An Exploration of Complex Matrix Factorization as a Tool for Single-Channel Musical Source Separation

James (Chuck) Bronson

Department of Music Research
Schulich School of Music
McGill University
Montreal, Canada

December 2013

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Arts.

2013/12/13

Abstract

Since the introduction of non-negative matrix factorization (NMF) as a tool for single-channel musical source separation (SC-MSS) in the early part of the 21^{st} century, it has steadily increased both in practicality and popularity, and continues to be a major area of focus within the signal processing community. It is acknowledged in the literature, however, that the fundamental assumption of the mixture model for the spectral representation of the acoustic sources, to which NMF is applied, violates the true nature of the superposition of acoustic signals by disregarding the phase relationship between overlapping sources. It is also recognized that NMF-based source separation procedures require additional post-processing, in order to reincorporate the phase information back into the spectral representation of the sources, once the factorization is complete.

This thesis explores Complex Matrix Factorization (CMF), a recently proposed variant of NMF, which incorporates the phase information of the acoustic sources into the matrix factorization framework, allowing for the development of source separation procedures founded on a mixture model rooted in the complex-spectrum domain (in which the superposition of overlapping sources is preserved). CMF has the additional benefit of integrating the estimation of the phase of the constituent sources directly into the factorization algorithm, eliminating the need for post-factorization phase estimation. Three experiments were conducted, investigating the behaviour of CMF, as it compares to NMF, when applied to simple test mixtures constructed from overlapping acoustic instruments. The main contribution of this thesis is the development of a physically motivated phase-based constraint, which restricts the relation between the phase parameter estimates over time. The CMF-based separation procedure, armed with this novel phase constraint, is demonstrated to offer promising results when employed as a tool for SC-MSS on the simple acoustic test case considered.

Résumé

Depuis l'introduction de la factorisation en matrices non-négatives (NMF) en tant qu'outil pour la séparation de sources musicales monophoniques (SC-MSS) au début du 21e siècle, son utilité et sa popularité ont augmenté de façon régulière et elle continue d'être un grand domaine d'intérêt au sein de la communauté de traitement du signal. Cependant, il est reconnu dans la littérature sur le sujet que l'hypothèse fondamentale du modèle de mélange pour la représentation spectrale de la source acoustique, à laquelle la NMF est appliquée, enfreint la véritable nature de la superposition des signaux acoustiques en ignorant la combinaison de phase entre sources superposées. Il est aussi reconnu que les procédures de type NMF pour la séparation de sources nécessitent des procédures de post-traitement additionnelles pour réintégrer l'information de phase à la représentation spectrale des sources une fois la factorisation complétée.

Cette thèse explore la factorisation de matrices complexes (CMF), une variante de la NMF récemment proposée qui intègre l'information de phase des sources acoustiques dans le cadre de la factorisation de matrices. Ceci permet le développement de procédures de séparation de sources fondées sur un modèle de mixage ancré dans le domaine des spectres complexes (dans lequel le mélange de sources qui se superposent est préservée). La CMF apporte aussi l'avantage d'intégrer l'estimation de la phase des sources constituantes directement à l'algorithme de factorisation, ce qui élimine le besoin d'estimation de phase a posteriori. Trois expériences ont été réalisées afin d'étudier le comportement de la CMF comparé à la NMF lorsqu'appliqué à des mélanges simples fait à partir d'instruments acoustiques qui se superposent. La principale contribution de cette thèse est la formalisation d'une contrainte de phase, fondée sur des propriétés physiques, qui impose des liens entre ses estimations au court du temps. Nous montrons que l'algorithme de la CMF, armée de cette nouvelle contrainte de phase, donne des résultats prometteurs lorsqu'il est utilisé sur des signaux acoustiques simples dans le cadre de la SC-MSS.

Acknowledgments

There are many people I would like to thank for all of the generous support they provided throughout my Master's research. First off, I would like to thank my supervisor, Philippe Depalle, for all of his invaluable guidance throughout that last two years. I would also like to give thanks to the administrative staff of the Faculty of Music; In particular, I would like to extend my gratitude to Helene Drouin for easing all procedural matters throughout my stay here at McGill. Similarly, I would like to thank Darryl Cameron for taking the time to help me with the various technical issues I had while conducting my research. I would also like to acknowledge the assistance offered by both Brian King and Jonathan Le Roux. Their valued correspondence greatly helped in my understanding of the CMF framework.

My studies were made all the more welcoming thanks to the support and friendship received from my colleagues. In particular I would like to thank, in no specific order, Hannah Robertson, Greg Burlet, Aaron Krajeski, Mike Winters, and everyone in the Signal Processing and Control Laboratory (SPCL). I would also like to extend a special thanks to Brodie Conley and Erin Kean, for all the support they provided and for being such awesome people in general.

Lastly, I would like to thank my loving family and wonderful partner Emma Wemyss. Their kindness and encouragement cannot be stressed enough.

Contents

List of Figures

List of Tables

List of Acronyms

BSS	Blind Source Separation
CASA	Computational Auditory Stream Analysis
CMF	Complex Matrix Factorization
CMFWISA	Complex Matrix Factorization with Intra-Source Additivity
DFT	Discrete Fourier Transform
DTFT	Discrete Time Fourier Transform
EUC-NMF	Euclidean Distance NMF
EP	Estimated Phase
FT	Fourier Transform
FHMM	Factorial Hidden Markov Model
GMM	Gaussian Mixture Model
GTR	Guitar Source
HMM	Hidden Markov Model
ICA	Independent Component Analysis
IDFT	Inverse Discrete Fourier Transform
IFT	Inverse Fourier Transform
IS	Itakura-Saito
IS-NMF	Itakura-Saito Divergence NMF
ISA	Independent Subspace Analysis
ISTFT	Inverse Short Time Fourier Transform
KKT	Karush-Kuhn-Tucker
KL	Kullback-Leibler
KL-NMF	Kullback-Leibler Divergence NMF
MM	Majorization-Minimization

MLE	Maximum Likelihood Estimation
MP	Modelled Phase
MSE	Mean Square Error
NBSS	Non-Blind Source Separation
NMF	Non-Negative Matrix Factorization
NMF2D	Non-Negative Matrix Factor 2-D Deconvolution
NMFD	Non-Negative Matrix Factor Deconvolution
OLA	Overlap-Add
OP	Oracle Phase
PCA	Principle Component Analysis
PE	Phase Evolution
PMF	Positive Matrix Factorization
PN	Piano Source
RMS	Root Mean Square
S/N	Signal to Noise Ratio
SAR	Source to Artifact Ratio
SBSS	Semi-Blind Source Separation
SC-MSS	Single-Channel Musical Source Separation
SDR	Source to Distortion Ratio
SIR	Source to Interference Ratio
SNR	Source to Noise Ratio
SSDR	Spectral Source to Distortion Ratio
STFT	Short Time Fourier Transform
TMR	Target-to-Masker Ratio
VQ	Vector Quatization

Chapter 1

Introduction

Single-channel musical source separation (SC-MSS) refers to the task of decomposing a musical mixture (e.g., a recording of an orchestral performance) into approximations of the constituent unmixed sources (e.g., separate recordings of each instrument) when only a monaural observation of the mixture is available for analysis. Many musical applications, such as automatic transcription, instrument classification, and data denoising/reduction, benefit from the processing of extracted musical sources in isolation.

Non-negative matrix factorization (NMF) (Lee and Seung 1999) has been demonstrated to be an effective tool for performing SC-MSS when applied to a time-frequency matrix representation (e.g., a magnitude/power spectrogram) of the musical mixture (Wang and Plumbley 2005). This approach, however, incorrectly models the spectrogram mixed sources as the sum of the spectrograms of the unmixed sources, disregarding any phase discrepancies between the overlapping spectra. In order for temporal domain source estimates to be constructed, NMF also requires that the phase of the constituent sources be estimated and reincorporated into their corresponding spectral representations. Complex Matrix Factorization (CMF) (Kameoka 2009) was proposed in 2009 to address these shortcomings by incorporating the previously excluded phase information directly into the factorization framework. To date, most of the research concerning CMF deals almost exclusively with acoustic mixtures of speech samples. There exist a substantial gap in the literature concerning the use of CMF as a tool for SC-MSS. This thesis specifically explores the use of CMF when applied to acoustic mixtures of musical sources. Note that the CMF was originally called "complex non-negative matrix factorization" in (Kameoka 2009). The

author of (King 2012), however, refers to the algorithm as "complex matrix factorization" for the sake of brevity and to disambiguate between the complex and non-negative methods. This name change shall also be adopted for the purposes of this thesis.

1.1 Overview of Experiments

Three experiments, outlined below, were conducted as a means of investigating the behaviour of CMF, as it compares to NMF, as a tool for SC-MSS. Each experiment analyzes sources extracted from the McGill University Master Samples Collection (Opolko and Wapnick 1987). Specifically, acoustic piano and classical guitar samples with 8 notes ranging from C4-C5, along a diatonic scale, for the piano and a fixed note of C4 for the guitar, were considered. The dataset was intentionally limited to only 9 notes and two instruments, favoring interpretability of the factorization over the ability to extrapolate beyond the current test cases considered, which could be obtained if sampling from a larger set of instruments with a wider range of notes. That being said, these test cases do represent (on a basic level) the general interaction of harmonic notes within a musical mixture and, as such, offer valuable insight into the nature of the temporal and spectral interaction of harmonic sources commonly occurring within musical mixtures.

Experiment 1: NMF vs. CMF Parameter Analysis:

The NMF-based and CMF-based source separation procedures depend on several problem-specific parameters associated with the Short-Time Fourier Transform ($STFT$) analysis, factorization algorithms, and Inverse Short-Time Fourier Transform ($ISTFT$) synthesis. The first experiment was conducted in order to gain insight into which analysis/factorization/synthesis parameter value combinations could potentially yield low/high NMF-based and CMF-based separation performances, evaluated using established source separation performance measures, outlined in (Vincent et al. 2006).

Experiment 2: NMF vs. CMF with Oracle Phase

A second experiment was conducted in order to investigate the potential benefits of the CMF-based separation procedure, over the NMF-based separation procedure, in conditions for which the NMF mixture model assumptions are most violated, given this particular

dataset/parameter value combinations. Namely, in conditions for which the sum of the magnitude spectrograms of the sources differs the most from the magnitude spectrogram of the mixture. This experiment also investigates the performance of the CMF-based separation procedure when the exact phase of the unmixed sources are known and incorporated into the factorization algorithm (oracle phase conditions). By analyzing the performance of the CMF-based separation procedure under oracle phase conditions, insight could be gained concerning a possible upper bound of achievable separation performance using the CMF-based separation procedure for this particular dataset/parameter value combinations, in conditions for which the NMF mixture model assumptions are substantially violated.

Experiment 3: NMF vs. CMF with Modelled Phase

Based on the results of the second experiment, a model for the phase behaviour of each source was developed and incorporated into the CMF framework in the form of a phase-based constraint, restricting the evolution of the phase parameter over time. A third experiment was conducted, as a proof of concept, in which the CMF-based separation procedure, together with this newly proposed phase constraint, was used to separate a mixture of harmonic sources possessing a strong violation of the NMF mixture model assumption. The performance of this newly proposed CMF-based separation procedure (with the phase constraint) was compared against the performance of the NMF-based separation procedure, the CMF-based separation procedure (without the phase constraint) and the CMF-based separation procedure (under oracle phase conditions).

1.2 Overview of Thesis

The remainder of this thesis is organized as follows. Chapter 2 reviews the relevant background literature concerning musical source separation, NMF, and CMF. Chapter 3 discusses pre-examination considerations and establishes the final version of the CMF algorithm used throughout the experimentation. Chapters 4 - 6 presents three experiments, introduced above, conducted as a means of investigating the behaviour of CMF as it compares to NMF, when applied as a tool for SC-MSS. Lastly, Chapter 7 recaps the main conclusions and contributions of this thesis and suggests possible future directions for further experimentation.

Chapter 2

Background

2.1 Musical Source Separation

The extraction of musical sources from a monaural mixture is a challenging task which has received a significant amount of attention from the signal processing community over the last decade. In order to fully appreciate the nature of the problem, several essential concepts, which provide the groundwork for many of the source separation techniques developed to date, will be addressed.

2.1.1 Problem Formulation

As described in (Jutten and Comon 2010), the basic relationship between an observed mixture and the underlying unmixed sources can be defined mathematically as follows:

$$\mathbf{x}(l) = \mathcal{M}(\mathbf{s}(l)) \tag{2.1}$$

where $\mathbf{x}(l) = (x_1(l), ...x_J(l))^T \in \mathbb{R}^J$ represents the J observed mixtures obtained from a sensor array (such as a microphone array), at sample index l, $\mathbf{s}(l) = (s_1(l), ...s_P(l))^T \in \mathbb{R}^P$ represents the P unknown constituent sources and $\mathcal{M}$ represents the unknown mixture mapping from $\mathbb{R}^P$ to $\mathbb{R}^J$. Given the above definition, source separation refers to the problem of extracting estimates, $\hat{\mathbf{s}}(l)$, of the unknown constituent sources, given the observed mixtures, $\mathbf{x}(l)$ (Jutten and Comon 2010).

2.1.2 Problem Classification

Three commonly discussed distinguishing characteristics of source separation problems are: 1) the number of sensors vs. the number of underlying sources, 2) the degree of a priori information incorporated into the separation technique, and 3) the mixture model adopted to describe the known/assumed physical nature of the mixing process. The various techniques developed for source separation problems differ based on these three distinguishing characteristics.

1) Number of Sensors vs. Number of Sources

Overdetermined/Determined ($J \geq P$): When the number of observed mixtures is greater than or equal to the number of sources, the separation problem is classified as overdetermined or determined, respectively. Note that this terminology parallels that used in linear algebra to describe a linear system of equations.

Underdetermined ($J < P$): When the number of sensors is less than the number of sources, the separation problem is said to be underdetermined. Taken to the extreme, when the mixture from only one sensor is observed (and the number of sources is non-trivial), the problem is referred to as a single-channel source separation. Underdetermined problems are ill-posed and additional a priori information is often needed to develop an effective separation.

Number of Observed Musical Mixtures: Most musical mixtures are either available as monophonic or stereophonic recordings. As such, musical-source separation tends to fall into the realm of single-channel source separation. This complicates the problem of extracting suitable source estimates due to the lack of immediately available information. Techniques developed for this type of situation differ substantially from those used in the overdetermined/determined case.

Number of Musical Sources: The exact definition of musical source is somewhat arbitrary. As discussed in (Virtanen 2006), a musical source may refer to every individual instrument making up a mixture (e.g., several violas playing in unison would all be considered as separate sources). Another possible definition would be to consider identical instruments playing in unison as one distinct musical source (e.g., several violas playing in unison

would be considered to be a single source). A distinction between the two interpretations of source is not considered further for the purposes of this thesis as the dataset used for the CMF-based source separation tasks, outlined in Chapters 4 - 6, consists of only two interacting instruments. In this case, the one-to-one correspondence between musical source and the original sounding instrument is clear. However, a more detailed discussion of the interpretation of source, rooted in auditory psychology and scene analysis (Bregman 1984), can be found in (Siamantas 2009).

2) A Priori Information

Blind Source Separation(BSS): When no (or minimal) prior information concerning the unmixed sources and the mixture model is incorporated into the separation task, the problem is commonly referred to as blind source separation (BSS). In extreme cases, even the number of sources is unknown (Virtanen 2006). In BSS, the sources are often assumed to be statistically stationary, independent and identically distributed with zero mean. Due to the limited amount of a prior information, BSS techniques typically require the problem to be overdetermined/determined. As such, BSS methods are rarely applicable to musical source separation tasks for which the problem is typically underdetermined.

Non-Blind Source Separation (NBSS): When a large amount of prior information concerning the unmixed sources and/or the manner in which they were mixed is incorporated into the separation algorithms, the problem is commonly referred to as non-blind. When dealing with musical source separation, the prior information in non-blind separation tasks is usually present in the form of higher-level (human informed) score-based information such as note onsets/ note durations, fundamental frequencies and/or timbral characteristics (Itoyama 2011).

Semi-Blind Source Separation (SBSS): A middle ground between blind and non-blind source separation exists, referred to as semi-blind source separation, in which the amount of incorporated prior information is not quite on the level of non-blind separation tasks but is substantially larger than the source-independent assumptions of BSS (Siamantas 2009). Semi-blind musical source separation approaches typically incorporate source-specific features, such as spectral/temporal smoothness, in the form of prior distributions governing the parameters of the mixture model (Erdogan and Grais 2010), (Virtanen 2004). It should

be noted that the amount of prior information incorporated into a source separation problem can be seen as lying along a continuum and the boundaries between semi-blind and the extreme cases of blind and non-blind are often blurred (Siamantas 2009).

Unsupervised/Supervised Source Separation: A concept that is closely related to the degree of blindness of a source separation task is the amount of incorporated human supervision. The exact definition of supervised and unsupervised separation may also vary somewhat depending on the literature, as discussed in (Siamantas 2009). For the purposes of this thesis, a source separation task is considered to be supervised if the algorithms used to infer the source estimates are trained beforehand on a database of isolated sources. Similarly, a source separation task is considered to be unsupervised if no training processes are used to infer source estimate parameters beforehand. Typically, BSS methods are unsupervised, NBSS are supervised, and SBSS methods can be either supervised or unsupervised

3) Mixture Model

A musical mixture is usually obtained by either: 1) recording two or more simultaneously sounding instruments, or 2) mixing together two individually sounding instruments using a mixing desk or a digital mixing environment (Jutten and Comon 2010). The way in which a musical mixture is obtained motivates the mixture model used to describe the physical system. As discussed in (Burred 2009), the mixture models developed to approximate the relationship between mixture observations and unmixed sources, described in Equation (2.1), typically account for the following observation/environment/source conditions: 1) delayed vs. instantaneous observations, 2) reverberant vs. anechoic environment, 3) noiseless vs. noisy environment 4) non-stationary vs. stationary sources. Here, "non-stationary" refers to the motion through space of the object generating the observed mixture (relative to the sensor). In order to simplify the musical source separation task, a commonly adopted mixture model is one which assumes stationary sources in a noiseless, anechoic space with instantaneous observations. This form is often referred to as the "instantaneous mixture model". It should be noted, however that musical sources are rarely perfectly stationary and the obtained mixtures are rarely perfectly anechoic. As such, a compromise must be made between accurately modelling the relationship between observed mixture and unmixed source and the computational complexity of the proposed model.

Mathematically, the instantaneous mixture model can be described as follows:

$$x_j(l) = \sum_{p=1}^{P} a_{jp} s_p(l) + e(l), \qquad j = 1, 2, ...J \tag{2.2}$$

where, $e(l)$ is the additive noise term and a_{jp} is a gain factor applied to source p from sensor j. As discussed above, most musical mixtures are either available as monophonic or stereophonic recordings. In the case of a single-channel mixture observation ($j = 1$) Equation (2.2) can be reduced to:

$$x(l) = \sum_{p=1}^{P} a_p s_p(l) + e(l) \tag{2.3}$$

where the subscript on x is dropped for convenience. This relationship is usually simplified further by assuming that the environment is approximately noiseless and that the mixing coefficients, a_p, are source independent and set to unity (Siamantas 2009).

$$x(l) = \sum_{p=1}^{P} s_p(l) \tag{2.4}$$

Using vector notation, this can be expressed as follows:

$$\mathbf{x} = \sum_{p=1}^{P} \mathbf{s_p} \tag{2.5}$$

where $\mathbf{x} = (x(0), x(1), ..., x(\tilde{L}-1))$ and $\mathbf{s_p} = (s_p(0), s_p(1), ..., s_p(\tilde{L}-1))$ ($\tilde{L}$ being the length of the observed mixture/sources).

2.1.3 Source Separation Methods

The goal of a source separation procedure is to approximate the underlying sources, $\mathbf{s}_p$, with source estimates $\widehat{\mathbf{s}}_p$. Techniques developed for extracting the desired source estimates vary depending on the aforementioned characteristics of the source separation problem to

which they are applied. Certain approaches are specifically suited to address the overdetermined/determined separation problem, whereas others are designed to approach the underdetermined case. The amount of available a priori information and the known/ assumed mixture model also influence the mathematical framework used to optimally extract the source estimates. A thorough review of sound source separation methods can be found in (Virtanen 2006), (Siamantas 2009), (Burred 2009), (Jutten and Comon 2010), and (Itoyama 2011).

Various informal typologies, which are all closely related in nature, have been proposed in attempt to categorize the separation techniques. For instance, (Siamantas 2009) suggests that musical source separation techniques can be divided into two classes: 1) those that are related to/inspired by the field of computational auditory stream analysis (CASA) and 2) those related to the field of BSS. On the other hand, the authors of (Duan et al. 2008), suggest that in the specific case of single-channel sound mixtures, the separation methods can be grouped into three, often overlapping, categories: 1) CASA-based methods, 2) Model-based methods 3) Spectral-decomposition-based methods. It should be noted that separation techniques do not always fall nicely into a given category and that these suggested typologies should serve merely as indications as to the motivation behind the development of a given source separation method. A brief review of single-channel musical source separation techniques, based on the typology presented in (Duan et al. 2008), follows.

CASA-Based Methods: The human auditory system is exceptionality tuned for selective listening of a single voice in a complex mixture of interfering voices (Brokhorst 2000). This is often referred to as the cocktail party phenomenon. Computational auditory scene analysis, (Brown and Cooke 1994), models the known processes of the human auditory system, using developed signal processing techniques, in attempt to imitate the source separation capabilities achievable by humans. Typically, CASA related approaches involve designing binary masks, (Weiss et al. 2006), which act on the mid-level time-frequency representation of an observed mixture and are used to separate the distinct perceptual auditory streams based on grouping principles outlined in (Bregman 1984). Streams are often identified based on temporal cues such as common onset/offset time and spectral cues such as harmonicity (Brown and Wang 2005).

Model-Based Methods: Statistical models such as factorial hidden Markov models (FHMMs) and Gaussian mixture models (GMMs) are very useful machine learning techniques which have demonstrated promising results when applied to musical source separation, as in (Mysore et al. 2010) and (Badeau 2011). Typically, model-based methods are computationally demanding, requiring several parameters to be learnt from training data. However, these methods also offer several benefits, such as accounting for temporal continuity of contiguous time frames using a HMM structure or allowing for phase-aware complex spectral information to be incorporated into the separation algorithm, using GMMs (Badeau 2011). More information regarding model-based methods can be found in (Duan et al. 2008).

Spectral-Decomposition-Based Methods: Mid-level time-frequency data representations are often incorporated into SC-MSS techniques to address the limited amount of directly exploitable information associated with low-level time-amplitude representation of the initial mixture observation. For instance, taking an $STFT$ of the time domain mixture, assuming the noiseless instantaneous model of Equation (2.5), yields:

$$\mathbb{X} = STFT\{\mathbf{x}\} = STFT\{\sum_{p=1}^{P} \mathbf{s_p}\} = \sum_{p=1}^{P} STFT\{\mathbf{s_p}\} = \sum_{p=1}^{P} \mathbb{S}_p \tag{2.6}$$

where the second last equality in Equation (2.6) holds due to the linearity of the $STFT$. Thus, in this case, the source separation problem can be framed as one of decomposing the complex $STFT$ of the mixture, $\mathbb{X}$, into the complex $STFT$s of the P sources. Most established decomposition methods, however, are developed for real-valued data, as opposed to the complex-valued data of an $STFT$. As such, spectral-decomposition-based methods are typically performed on a real-valued (magnitude/power) spectrogram of a single-channel mixture. Thus, the commonly adopted mixture model is as follows:

$$|\mathbb{X}| = X = \sum_{p=1}^{P} S_p \tag{2.7}$$

where X represents the spectrogram of the mixture and S_p represents the spectrogram of the p^{th} source. However, as previously stated in the introduction, and as will be discuss further in Subsection 2.3, it does not hold in general that the sum of the spectrogram of the sources, $\sum_{p=1}^{P} S_p$, is equal to the spectrogram of the mixture, X, (Parry and Essa 2007b). For

clarity, this problem is restated mathematically, using the established notation, as follows:

$$X = |\mathbb{X}| = |\sum_{p=1}^{P} \mathbb{S}_p| \neq \sum_{p=1}^{P} |\mathbb{S}_p| = \sum_{p=1}^{P} S_p \tag{2.8}$$

This idea will be revisited in Section 2.3. Note that, for the remainder of this thesis, the term spectrogram should be taken as referring to the magnitude spectrogram of a signal, unless specified otherwise.

Independent subspace analysis (ISA) and NMF are two popular examples of spectral-decomposition methods. ISA is motivated by independent component analysis (ICA) and, as such, a brief description of both ICA and ISA concludes this section on musical source separation. A thorough review of the main spectral-decomposition-based method considered in this thesis, NMF, is discussed in Section 2.2, followed by the complex variant, CMF, in Section 2.3.

ICA is a BSS technique applicable to source separation problems in which the number of observations is greater than or equal to the number of sources (overdetermined/determined problem). This approach is not directly applicable to SC-MSS due to the lack of mixture observations, as such, only a brief outline of ICA will be presented here, but the reader is referred to (Burred 2009), (Virtanen 2006), and (Jutten and Comon 2010) for a more in-depth review. ICA assumes the sources are independent, identically distributed and non-Gaussian. Source estimates are then calculated using optimization algorithms based on maximizing measures of independence or nongaussianity of the pre-whitened mixture observations. Here, the mixture observations, $\mathbf{x}(l)$, are treated as realizations of a multidimensional random vector and the pre-whitening process refers to centering the data about zero, decorrelating the data and scaling each element to normalize the variance.

Independent subspace analysis, as described in (Casey and Westner 2000), is motivated by the limitations of ICA, namely the restrictions of ICA to the overdetermined/ determined separation problems. For the underdetermined case of single-channel musical source separation, ISA can be considered an extension of ICA in which the single mixture observation is first mapped onto a higher dimensional space using a suitable transformation (e.g., the $STFT$). In doing so, the absolute value of the elements at each time frame of the newly transformed data can then be regarded as an observed mixture. In other words, each time frame can be considered as being made up of a mixture of spectral bases, which are assumed to be statistically independent and hence separable using the ICA framework.

2.2 Non-negative Matrix Factorization

The following section presents an overview of NMF, providing a brief summary of the history, problem formulation, motivations, and application of NMF as a spectral-decomposition-based method for SC-MSS. This section also reviews the theory concerning NMF cost functions, update algorithms, implementation, and convergence considerations, followed by a summary of various structural reformations and constraints imposed on NMF. This section concludes with a presentation of the final form of NMF, with the inclusion of a sparsity penalty term, which is used throughout the experiments of Chapters 4 - 6.

2.2.1 Brief History

One of the earliest developments of NMF is attributed to the work by D.D Lee and H.S Seung, which was published in the scientific journal, Nature, in 1999 (Lee and Seung 1999). The profundity of this paper lies in its presentation of NMF as a non-subtractive, part-based, data representation and for its detailing of efficient multiplicative update algorithms, which facilitated the generalization/applicability of NMF across various fields of research. However, it should be noted that NMF had been analyzed earlier under the name Positive Matrix Factorization (PMF) (Paatero and Tappert 1994) and is rooted even earlier still in the theory of curve resolution, which concerns the problem of separating a mixture of overlapping, non-negative, linearly independent functions, as discussed in (Lawton and Sylvestre 1971).

2.2.2 Problem Statement

As depicted in figure 2.1, NMF consists of approximately factorizing a given $N \times M$ non-negative data matrix, X, into an $N \times K$ non-negative basis matrix, W, and $K \times M$ non-negative activation matrix, H.

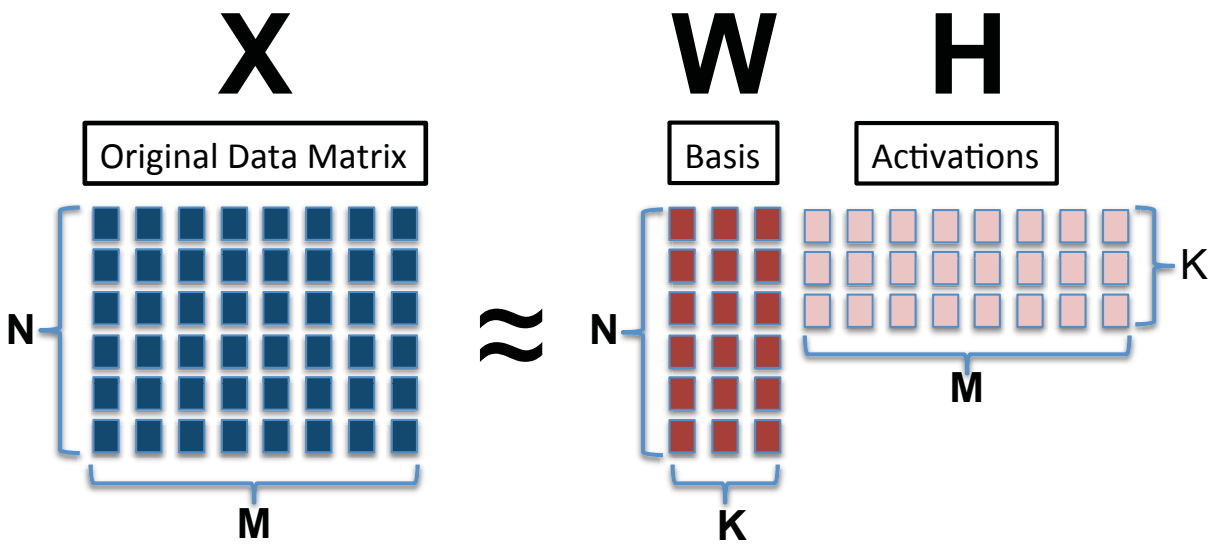

Fig. 2.1: NON-NEGATIVE MATRIX FACTORIZATION: $X \approx WH$

Basic NMF can be described mathematically as follows:

$$\begin{aligned} \text{Given} &: X \in \mathbb{R}_{\geq 0}^{N \times M} \text{ and } K \in \mathbb{R}_{>0} \\ \text{Factorize} &: X \approx \widehat{X} = WH \\ \text{Subject to} &: W \in \mathbb{R}_{\geq 0}^{N \times K} \text{ and } H \in \mathbb{R}_{\geq 0}^{K \times M} \end{aligned} \tag{2.9}$$

The inner dimension of the factorization, K, is typically chosen such that $K(N+M) \ll NM$ or $K \ll \min(N, M)$. In this sense, NMF is regarded as a data reduction technique (Lee and Seung 1999), (Berry et al. 2007). In the full-rank case of $K = N$, the approximation in the factorization statement above becomes a strict equality and the NMF becomes an exact factorization problem: $X = WH$ (Klingenberg et al. 2009). The smallest value of K for which the exact factorization exists is referred to as the non-negative rank (Wang and Zhang 2012), (Gillis and Glineur 2012), denoted $\text{rank}_+(X)$. Sufficient conditions for which the exact NMF problem is uniquely solvable (up to a scaling and permutation of the factors) is discussed in several papers, notably (Donoho and Stodden 2003) and (Laurberg et al. 2008). It should be noted that the exact NMF problem has been proven to be NP-hard, meaning it is at least as hard as any non-deterministic polynomial-time decision problem (Vavasis 2009), (Gillis 2012).

2.2.3 Motivations

Several matrix factorization techniques exist, such as Principle Component Analysis (PCA), Vector Quantization (VQ), and NMF, in which the matrix factors are structured according to imposed constraints/penalty functions required/desired for a given application. NMF, as stated above, refers to a basic matrix factorization under the constraints of element-wise non-negativity. Data extracted in many fields of study, including image processing and audio spectra analysis, is often constrained to the set of non-negative real numbers. By enforcing the constraints of non-negativity of the matrix factors W and H, the corresponding factorization adopts an purely additive, interpretable, part-based decomposition structure that is not present in more holistic factorization techniques, such as PCA and VQ (Lee and Seung 1999).

The part-based decomposition structure inherent in NMF is evident when one considers the column-by-column approximation: $\mathbf{x}_m = [X]_{:,m} \approx W[H]_{:,m} = W\mathbf{h}_m$, where $\mathbf{x}_m$ and $\mathbf{h}_m$ correspond to the m^{th} columns of X and H, respectively (Lee et al. 2001). In this form, it is clear that the m^{th} N-dimensional observed data vector, $\mathbf{x}_m$, is approximated as a linear combination of the N-dimensional basis vectors which make up the columns of W, scaled according to the non-negative weights specified by $\mathbf{h}_m$. The non-negativity constraints imposed on H guarantees that the linear combination is purely additive, motivating a part-based decomposition. The non-negativity constraints imposed on W guarantees that the column of X and W exist in the same vector space of $\mathbb{R}^N_{\geq 0}$, motivating the interpretability of the columns of W as physically meaningful underlying components of X (Wang and Zhang 2012).

2.2.4 NMF-Based Single-Channel Musical Source Separation

As was first examined in (Smaragdis and Brown 2003), when the data matrix in question X, corresponds to a non-negative time-frequency representation (e.g., a magnitude/power spectrogram) of a musical passage, the *ideal* non-negative factorization results in matrix factor, W, whose column vectors correspond to spectral templates of the generative musical structures (e.g., sustained tones, transients, and noise). These spectral profiles combine in a purely additive fashion, according to the non-negative temporal activations, specified by the rows of H, to form an approximation of the original time-frequency representation of the musical passage.

In 2005, NMF was explored as a tool for single-channel musical source separation, in (Wang and Plumbley 2005). The NMF-based musical source separation methods discussed in (Wang and Plumbley 2005) and later in (Wang and Plumbley 2006), closely parallel the factorization approach adopted in (Smaragdis and Brown 2003) for polyphonic music transcription. In order to best describe the steps typically followed in an NMF-based SC-MSS, a mathematical recapitulation of the process is presented below.

NMF-Based SC-MSS Procedure

As described above in Equation (2.7), the spectral-decomposition-based methods for single-channel source separation model the time-frequency representation, X, of the musical mixture as the sum of the time-frequency representations of the P sources, $S_1, ..., S_p$. For clarity, Equation (2.7) is restated as follows:

$$X = \sum_{p=1}^{P} S_p \tag{2.10}$$

The goal of spectral-based separation procedures is to determine suitable estimates, $\widehat{S}_1, ..., \widehat{S}_p$, for the underlying spectrograms of the sources, $S_1, ..., S_p$. NMF-based source separation procedures are developed based on the idea that the NMF algorithm is able to extract the spectral profiles and the temporal activations of the rank one components, $\widehat{C}_k$, which make up the estimates, $\widehat{S}_p$, for the underlying sources S_p. These rank one components are obtained by multiplying each spectral template, specified by the columns of W, by the corresponding temporal activations, specified by the rows of H, as follows:

$$\widehat{C}_k = [W]_{:,k}[H]_{k,:} \tag{2.11}$$

A musical source is typically composed of several notes, where each note may be described by a highly time-varying spectral shape (Hennequin et al. 2011). As such, several rank one components are often required to accurately describe the spectrogram of each source. Thus, a clustering strategy is needed to separate the K time-frequency representation of the components, $\widehat{C}_k$ into distinct groups, corresponding to the approximations of the spectrograms of the P sources. The concept of clustering will be elaborated upon in Subsection 3.3.6, however, assume that for the time being an appropriate strategy is chosen such that the components are adequately grouped into the underlying sources. This can be described

mathematically by considering the set of integers corresponding to the set of all component indices: $\mathcal{K} = [1, K]$. Letting $\mathcal{K}_p$ represent the set of all component indices corresponding to source p, it follows that $\mathcal{K}_p \subseteq \mathcal{K}$, $\mathcal{K}_p \bigcap \mathcal{K}_{\widetilde{p}} = \emptyset$ for $p \neq \widetilde{p}$ and $\sum_{p=1}^{P} |\mathcal{K}_p| = |\mathcal{K}| = K$, where $|\mathcal{K}|$ denotes the cardinality of the set. Using this notation, the synthesis of the source spectrogram estimates from the components $\widehat{C}_k$, can be expressed as follows:

$$\widehat{S}_p^{\text{SYNTH}} = \sum_{k \in \mathcal{K}_p} \widehat{C}_k = \sum_{k \in \mathcal{K}_p} [W]_{:,k}[H]_{k,:} \tag{2.12}$$

The entire NMF-based SC-MSS procedure can be summarized as follows:

$$\sum_{p=1}^{P} S_p = X \approx \widehat{X} = WH = \sum_{k=1}^{K} [W]_{:,k}[H]_{k,:} = \sum_{p=1}^{P} \sum_{k \in \mathcal{K}_p} \widehat{C}_k = \sum_{p=1}^{P} \widehat{S}_p^{\text{SYNTH}} \tag{2.13}$$

Note that an alternative source estimation procedure exists (Smaragdis 2007), in which the sources are extracted through a filtering process of the original spectrogram, X, as opposed to the direct synthesis of the sources via Equation (2.12). This filtering process, which is equivalent to Wiener filtering (Daudet 2012), can be described mathematically as follows:

$$\widehat{S}_p^{\text{FILT}} = \frac{\sum\limits_{k \in \mathcal{K}_p} \widehat{C}_k}{\sum\limits_{p=1}^{P} \sum\limits_{k \in \mathcal{K}_p} \widehat{C}_k} \bullet X = \frac{\sum\limits_{k \in \mathcal{K}_p} [W]_{:,k}[H]_{k,:}}{\sum\limits_{p=1}^{P} \sum\limits_{k \in \mathcal{K}_p} [W]_{:,k}[H]_{k,:}} \bullet X \tag{2.14}$$

where the division and multiplication of the rightmost expression is taken to be element-wise. For the remainder of this thesis, source estimation through direct addition of the grouped components, as described in Equation (2.12), will be referred to as *synthesis-based source estimation*, whereas source estimation using the filtering-based approach, as described in Equation (2.14), will be referred to as *filtering-based source estimation*.

Next, if the aim of the source separation is to obtain a time-domain waveform of the separated sources, the $ISTFT$ of the obtained time-frequency representations of the sources can be taken. To do this, however, an estimate of the phase information of the individual sources is needed. As was the case for choosing an appropriate clustering strategy, assume that for the time being the phase information of the individual sources is adequately deter-

mined. Approaches for recovering the discarded phase will be discussed in Subsection 3.3.7. Taking the $ISTFT$ of the extracted time-frequency representation of the synthesis-based source estimates, together with their appropriately determined phase information, yields:

$$ISTFT\{(\sum_{k\in\mathcal{K}_p}\widehat{C}_k)^{\Phi}\} = ISTFT\{\widehat{\mathbb{S}}_p^{\text{SYNTH}}\} = \widehat{\mathbf{s}}_p^{\text{SYNTH}} \tag{2.15}$$

where the Φ superscript is used here to denote the fact that the determined phase information has been incorporated into the time-frequency representation of the source estimates and $\widehat{\mathbb{S}}_p$ represents the $STFT$ estimate of the p^{th} source. Similarly, taking the $ISTFT$ of the extracted time-frequency representation of the filter-based source estimates, together with their appropriately determined phase information, yields:

$$ISTFT\{(\frac{\sum_{k\in\mathcal{K}_p}\widehat{C}_k}{\sum_{p=1}^{P}\sum_{k\in\mathcal{K}_p}\widehat{C}_k}\bullet X)^{\Phi}\} = ISTFT\{\widehat{\mathbb{S}}_p^{\text{FILT}}\} = \widehat{\mathbf{s}}_p^{\text{FILT}} \tag{2.16}$$

2.2.5 NMF Cost Functions

In order to solve for matrix factors W and H, an appropriate quantifiable measure of the quality of approximation must first be defined. This can be accomplished by establishing a separable measure of "distance" or "divergence" between the matrix X and WH, which is then minimized, subject to element-wise non-negativity. Note that a distinction is made between the mathematical concepts of "distance" and "divergence", the former being a symmetric measure and the latter asymmetric (Lee et al. 2001). The constrained optimization problem can be formulated as follows:

$$\begin{aligned}
&\text{Given} : X \in \mathbb{R}_{\geq 0}^{N\times M} \text{ and } K \in \mathbb{R}_{>0} \\
&\text{Optimize} : \min_{W,H} D(X||WH) = \sum_{n,m} d([X]_{n,m}||[WH]_{n,m}) \\
&\text{Subject to} : W \in \mathbb{R}_{\geq 0}^{N\times K} \text{ and } H \in \mathbb{R}_{\geq 0}^{K\times M}
\end{aligned} \tag{2.17}$$

Here, $d(a||b)$ is typically any suitable scalar cost function such that: 1) $d(a||b) \geq 0$ for

$a \in \mathbb{R}_{\geq 0}$ given $b \in \mathbb{R}_{\geq 0}$ and 2) $d(a||b) = 0 \iff a = b$, $\forall (a, b) \in \mathbb{R}^2_{\geq 0}$. Various cost functions have been examined in the literature, including the Euclidean distance and (generalized) Kullback-Leibler (KL) divergence, which are both detailed in (Lee et al. 2001), and more recently the Itakura-Saito (IS) divergence, which is discussed in (Févotte et al. 2009).

As examined in (Févotte and Idier 2011), these cost functions can be considered as special cases of the parameterized family of cost functions known as the β-divergence (Basu et al. 1998). The β-divergence can be expressed mathematically as follows:

β-divergence:

$$d_\beta(a||b) = \begin{cases} \frac{1}{\beta(\beta-1)}(a^\beta + (\beta-1)b^\beta - \beta a b^{\beta-1}), & \beta \in \mathbb{R}\backslash\{0,1\} \\ a \log \frac{a}{b} - a + b, & \beta = 1 \\ \frac{a}{b} - \log \frac{a}{b} - 1, & \beta = 0 \end{cases} \tag{2.18}$$

For $\beta = 2$ the β-divergence reduces to the square of the Euclidean distance and the constrained optimization problem becomes:

Euclidean Distance NMF (EUC-NMF):

$$\begin{aligned} &\text{Given} : X \in \mathbb{R}^{N \times M}_{\geq 0} \text{ and } K \in \mathbb{R}_{>0} \\ &\text{Optimize} : \min_{W,H} D_{EUC}(X||WH) = \frac{1}{2}\sum_{n,m}([X]_{n,m} - [WH]_{n,m})^2 \\ &\text{Subject to} : W \in \mathbb{R}^{N \times K}_{\geq 0} \text{ and } H \in \mathbb{R}^{K \times M}_{\geq 0} \end{aligned} \tag{2.19}$$

Note that, as mentioned in (Févotte and Idier 2011), the β-divergence for $\beta = 2$ differs from the square of the Euclidean distance in the sense that the square of the Euclidean distance is formally defined for vectors, whereas the β-divergence considers scalar inputs. For the remainder of the thesis, this consideration will be overlooked and no distinction will be made between the β-divergence for $\beta = 2$ and the Euclidean distance, as in (Févotte and Idier 2011).

The limiting cases of $\beta = 1$ and $\beta = 0$ correspond to the Kullback-Leibler and Itakura-Saito divergences, respectively:

Kullback-Leibler Divergence NMF (KL-NMF):

$$\begin{aligned} \text{Given} &: X \in \mathbb{R}_{\geq 0}^{N \times M} \text{ and } K \in \mathbb{R}_{>0} \\ \text{Optimize} &: \min_{W,H} D_{KL}(X||WH) = \sum_{n,m} ([X]_{n,m} \log \frac{[X]_{n,m}}{[WH]_{n,m}} - [X]_{n,m} + [WH]_{n,m}) \\ \text{Subject to} &: W \in \mathbb{R}_{\geq 0}^{N \times K} \text{ and } H \in \mathbb{R}_{\geq 0}^{K \times M} \end{aligned} \tag{2.20}$$

Itakura-Saito Divergence NMF (IS-NMF):

$$\begin{aligned} \text{Given} &: X \in \mathbb{R}_{\geq 0}^{N \times M} \text{ and } K \in \mathbb{R}_{>0} \\ \text{Optimize} &: \min_{W,H} D_{IS}(X||WH) = \sum_{n,m} (\frac{[X]_{n,m}}{[WH]_{n,m}} - \log \frac{[X]_{n,m}}{[WH]_{n,m}} - 1) \\ \text{Subject to} &: W \in \mathbb{R}_{\geq 0}^{N \times K} \text{ and } H \in \mathbb{R}_{\geq 0}^{K \times M} \end{aligned} \tag{2.21}$$

It should be noted that the β-divergence is a continuous function of β under the identities: $\lim_{\beta \to 0} d_\beta(a||b) = \frac{a}{b} - \log \frac{a}{b} - 1$ and $\lim_{\beta \to 1} d_\beta(a||b) = a(\log a - \log b) + (b - a)$, (Nakano et al. 2010). In fact, several other values of β have been considered as cost functions for NMF in the literature, (Fitzgerald et al. 2009). The optimal choice of β for a given application is still an open and challenging question. It should also be noted that various other families of cost functions have been considered as well, such as the α-divergence (Cichocki et al. 2008), α-β-divergence (Cichocki et al. 2011), Csiszrs divergence (Cichocki et al. 2006), and Bregman divergence (of which the β-divergence is a subclass (Hennequin et al. 2011)). The reader is referred to (Cichocki et al. 2009) and (Wang and Zhang 2012) and the references therein for more information regarding these divergences and others as they pertain to NMF.

2.2.6 NMF Update Algorithms

The aforementioned Euclidean distance, $d_{EUC}(a||b)$, and KL Divergence, $d_{KL}(a||b)$, are convex cost functions of the argument b (Févotte 2011). As such, the corresponding measure of fit, $D(X||WH)$, is convex as a function of W only or H only but not as a function of both W and H (Lee et al. 2001). The IS divergence, $d_{IS}(a||b)$, on the other hand is a convex function of b on $(0, 2a]$ and a concave function of b on $[2a, \infty)$ (Févotte 2011). Due to the absence of strict convexity of the problem formulation, these non-convex formulations are typically adopted and optimized via alternating minimization schemes (Wang and Zhang 2012), such as the multiplicative update algorithms originally presented in (Lee and Seung 1999) and (Lee et al. 2001). These multiplicative alternating update algorithms are based on updating one matrix factor at a time while keeping the other fixed. They were first developed for the EUC-NMF and KL-NMF in (Lee and Seung 1999) and later established for IS-NMF in (Févotte et al. 2009).

EUC-NMF Multiplicative Updates:

$$H \leftarrow H \bullet \frac{(W^T X)}{(W^T W H)} \qquad W \leftarrow W \bullet \frac{(X H^T)}{(W H H^T)} \tag{2.22}$$

KL-NMF Multiplicative Updates:

$$H \leftarrow H \bullet \frac{W^T((WH)^{\bullet -1} \bullet X)}{W^T \cdot \mathbf{1}} \qquad W \leftarrow W \bullet \frac{((WH)^{\bullet -1} \bullet X)H^T}{\mathbf{1} \cdot H^T} \tag{2.23}$$

IS-NMF Multiplicative Updates:

$$H \leftarrow H \bullet \frac{W^T((WH)^{\bullet -2} \bullet X)}{W^T(WH)^{\bullet -1}} \qquad W \leftarrow W \bullet \frac{((WH)^{\bullet -2} \bullet X)H^T}{(WH)^{\bullet -1}H^T} \tag{2.24}$$

where $A \bullet B$, $A^{\bullet n}$ and $\frac{A}{B}$ are being used here to denote, respectively, element-wise multiplication (Hadamard product), element-wise exponentiation, and element-wise division. Finding a global minimum of the NMF optimization problem formulated above is generally not feasible given that the NMF problem is NP-hard and non-convex (Wang and Zhang 2012), (Gillis 2011). As such, only a local minimum is typically sought after, reducing the problem to a small neighbourhood of the cost function in question (Ho 2008). The derivation of

the three update schemes presented above is based on an adaptive, diagonally-rescaled, gradient descent optimization approach, as presented in (Lee et al. 2001) and discussed in several other papers, including (Gonzalez and Zhang 2005), (Chu et al. 2004), and (Kirchhoff et al. 2012). Another derivation of the multiplicative updates is presented in (Schmidt 2008), (Févotte et al. 2009), (Févotte and Idier 2011), and discussed in (Hennequin et al. 2011) and (Bertin et al. 2009), which is based on the following heuristic:

$$\theta \leftarrow \theta \frac{P_\theta}{Q_\theta} \tag{2.25}$$

Where θ represents an element of W or H, and $\nabla_\theta D(\theta) = Q_\theta - P_\theta$. The above update heuristic guarantees that θ remains non-negative if initialized with non-negative values (Schmidt 2008) and evolves in the descent direction (opposite direction of the gradient). It is also structured such that a stationary point of the algorithm is one for which either $\theta = 0$ or $P_\theta - Q_\theta = 0 \implies \nabla_\theta D(\theta) = 0$. An interesting property of the β-divergence, $d_\beta(a||b)$, is that the first partial derivative with respect to the argument b is also continuous in β and can be expressed as follows:

$$\bigtriangledown_b \, d_\beta(a||b) = b^{\beta-2}(b-a) \tag{2.26}$$

Thus, as discussed in (Févotte et al. 2009), the gradients of the measure of the quality of approximation, $D_\beta(X||WH)$, with respect to W and H are given by

$$\begin{aligned} \bigtriangledown_H D_\beta(X||WH) &= W^T((WH)^{\bullet[\beta-2]} \bullet (WH - X)) \\ \bigtriangledown_W D_\beta(X||WH) &= ((WH)^{\bullet[\beta-2]} \bullet (WH - X))H^T \end{aligned} \tag{2.27}$$

Substituting these expressions into the above update heuristic formulation (2.25), yields:

$$H \leftarrow H \bullet \frac{W^T((WH)^{\bullet(\beta-2)} \bullet X)}{W^T(WH)^{\bullet(\beta-1)}} \qquad W \leftarrow W \bullet \frac{((WH)^{\bullet(\beta-2)} \bullet X)H^T}{(WH)^{\bullet(\beta-1)}H^T} \tag{2.28}$$

Ultimately, setting $\beta = 2, 1, 0$ in Equation (2.28) above, the multiplicative updates stated in Equations (2.22), (2.23), and (2.24) follow respectively.

2.2.7 β-NMF with Multiplicative Updates

The generalized β-NMF using a multiplicative update optimization scheme can summarized as follows:

Algorithm 1 β-NMF WITH MULTIPLICATIVE UPDATES

Input: $X \in \mathbb{R}_{\geq 0}^{N \times M}$ and $K \in \mathbb{R}_{>0}$

Output: $WH \approx X, W \in \mathbb{R}_{\geq 0}^{N \times K}$ and $H \in \mathbb{R}_{\geq 0}^{K \times M}$

Initialize W, H such that $W \in \mathbb{R}_{>0}^{N \times K}$ and $H \in \mathbb{R}_{>0}^{K \times M}$

while stopping criteria not met **do**

Compute $\widehat{X} = WH$

$H \leftarrow H \bullet \frac{W^T((WH)^{\bullet(\beta-2)} \bullet X)}{W^T(WH)^{\bullet(\beta-1)} + \Delta}$

Compute $\widehat{X} = WH$

$W \leftarrow W \bullet \frac{((WH)^{\bullet(\beta-2)} \bullet X)H^T}{(WH)^{\bullet(\beta-1)}H^T + \Delta}$

Normalize W **and scale** H **accordingly**

iter = iter + 1

end while

Note that the Δ term in the denominator of the matrix factor updates is included to avoid possible division by zero (Berry et al. 2007). Also note that the normalization of the matrix factor W after every update (specifically, normalization of the columns of W to unity in the L_1 or L_2 sense, assuming no columns of W are entirely zero) is needed to account for the indeterminacy of the arbitrary matrix factor scaling of W and H which could lead to numerical instabilities, ill-conditioning (Wang and Zhang 2012), and many trivial multiple solutions (Finesso and Spreij 2004). Any non-negative diagonal matrix G, the NMF $X \approx WH$ also yields the solution $X \approx WGG^{-1}H$ with equivalent cost: $D(X||WH) = D(X||WGG^{-1}H)$ (Laurberg et al. 2008). As discussed in (Lin 2007b) and (Wang and Zhang 2012) however, introducing this normalization factor after every update alters the original problem and, as such, may complicate the optimization procedure. It

is important to mention that enforcing a normalization on the columns of W does not alleviate the non-uniqueness of the NMF problem, which can easily be seen in the fact that an NMF solution is still non-unique up to a permutations of the columns of W and rows of H, expressed mathematically as $WH = W\Pi\Pi^{-1}H$ for a given permutation matrix Π (Finesso and Spreij 2004).

It should also be restated that the multiplicative update algorithm presented above is simply one of many optimization schemes which have been considered in the literature. As outlined in (Berry et al. 2007), NMF algorithms can be tentatively divided into three, possibly overlapping, update approaches: 1) multiplicative update algorithms, 2) gradient descent algorithms (Hoyer 2004), (Chu et al. 2004), and 3) alternating least squares algorithms (Paatero and Tappert 1994), (Langville et al. 2006), (Cichocki and Zdunek 2007), (Kim and Park 2008).

2.2.8 Convergence of NMF Multiplicative Update Algorithms

In (Lee et al. 2001), the Euclidean distance and KL-divergence based NMF cost functions, $D_{EUC}(X||WH)$ and $D_{KL}(X||WH)$, were proven, based on the theory of maximization-minimization (MM) algorithms, to be monotonically non-increasing, under the multiplicative updates presented above in Equations (2.22) and (2.23). MM algorithms are a class of algorithms rooted in the construction of a majorizing auxiliary functions, which can be optimized in place of the primary function for situations in which the primary function is difficult or impossible to optimize directly (Hunter and Lange 2003). Formally, a majorizing auxiliary function (here taken to have only non-negative real arguments as presented in (Févotte 2011)) can be defined as follows:

Majorizing Auxiliary Function: $F^+(\theta, \theta^{(\rho)}) : \mathbb{R}^N_{\geq 0} \times \mathbb{R}^N_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$, where $\theta^{(\rho)}$ represents a fixed value for the parameter θ, is a majorizing auxiliary function of $F(\theta)$ provided:

$$\begin{aligned} &\bullet\ F(\theta) \leq F^+(\theta, \theta^{(\rho)}), \forall(\theta, \theta^{(p)}) \in \mathbb{R}^N_{\geq 0} \times \mathbb{R}^N_{\geq 0} \\ &\bullet\ F(\theta^{(\rho)}) = F^+(\theta^{(\rho)}, \theta^{(\rho)}), \end{aligned} \tag{2.29}$$

Thus, the majorizing auxiliary function defines an upper bound on the associated primary function, with equality holding at $\theta = \theta^{(\rho)}$, as discussed in (Févotte 2011). The optimization problem is then framed as a minimization of the majorizing function, $F^+(\theta, \theta^{(\rho)})$, rather

than the primary function, $F(\theta)$. This is typically realized through iterative update schemes on θ, such as:

$$\theta^{(\rho+1)} = arg \min_{\theta \geq 0} F^{+}(\theta, \theta^{(\rho)}) \tag{2.30}$$

under which it can be proven that $F(\theta)$ is monotonically non-increasing, as follows:

$$F(\theta^{(\rho+1)}) \leq F^{+}(\theta^{(\rho+1)}, \theta^{(\rho)}) \leq F^{+}(\theta^{(\rho)}, \theta^{(\rho)}) = F(\theta^{(\rho)}) \tag{2.31}$$

The first inequality holds by definition of the majorizing auxiliary function and the second inequality holds from the definition of the update scheme. A graphical representation illustrating the basic concepts of MM algorithms is depicted below in figure 2.2.

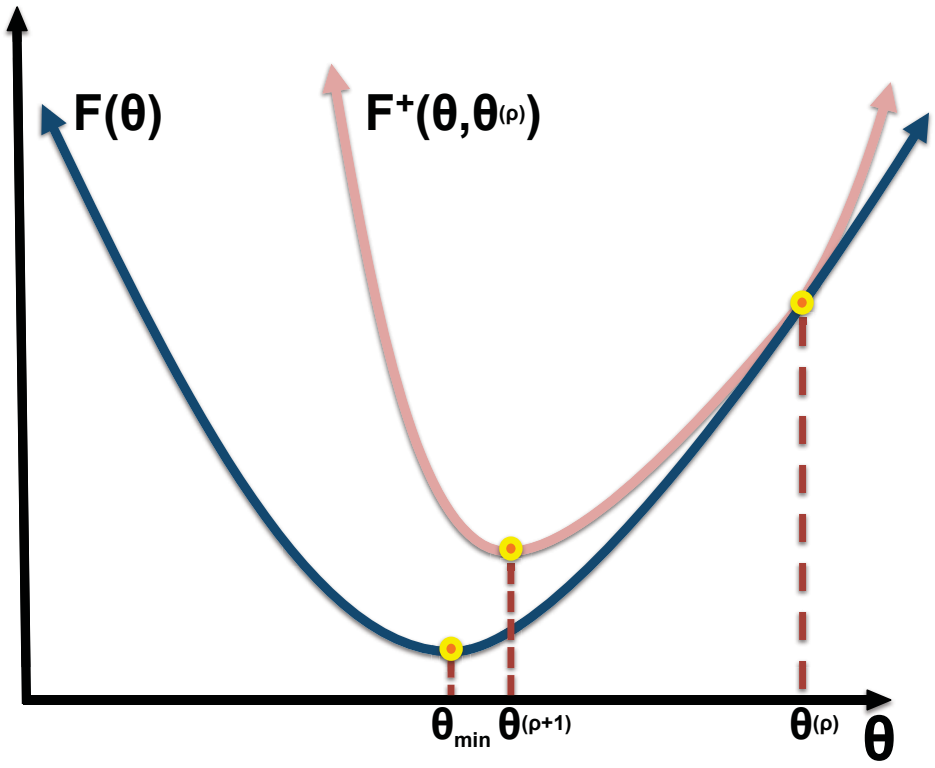

Fig. 2.2: EXAMPLE FUNCTION $F(\theta)$ AND ASSOCIATED AUXILIARY FUNCTION $F^{+}(\theta, \theta^{(\rho)})$

In (Lee et al. 2001), majorizing auxiliary functions were constructed for the *EUC-NMF* and *KL-NMF* cost functions, $D_{EUC}(X||WH)$ and $D_{KL}(X||WH)$, and the updates which

resulted from the majorization-minimization scheme presented in Equation (2.30) were shown to be equivalent to the NMF updates presented in Equation (2.22) and (2.23). As such, the primary functions, corresponding to the Euclidean distance and KL-divergence based NMF cost functions, $D_{EUC}(X||WH)$ and $D_{KL}(X||WH)$, were proven to be monotonically non-increasing based on the result of Equation (2.31).

This proof was originally framed as a one of convergence of the algorithm to a stationary point that is also a local minimum, which was later questioned in (Gonzalez and Zhang 2005) through the non-convergence of a numerical example, and more rigorously demonstrated to be incorrect in (Lin 2007a), (Lin 2007b), (Berry et al. 2007), (Finesso 2006) under the consideration of the Karush-Kuhn-Tucker (KKT) bounded optimality conditions for the NMF (Bertsekas 1999). Appropriate conclusions that can be drawn regarding the convergence of the algorithm under the multiplicative updates of Equations (2.22), corresponding to the Euclidean distance cost function, are summarized in (Berry et al. 2007) as follows: "When the algorithm has converged to a limit point in the interior of the feasible region $[[H]_{n,k} > 0, [W]_{k,m} > 0]$, this point is a stationary point. This stationary point may or may not be a local minimum. When the limit point lies on the boundary of the feasible region, its stationarity can not be determined". Similarly, in (Dessein et al. 2013), it is stated that the multiplicative updates of Equations (2.22) and (2.23) are "applied in turn until convergence, and ensure both non-negativity of the factors W and H as well as monotonic decrease of the cost, but not necessarily convergence of the factors nor local optimality". For more information regarding the convergence considerations of the NMF algorithms, the reader is referred to (Berry et al. 2007), (Dessein et al. 2013), and the references therein.

Since, (Lee et al. 2001), additional auxiliary functions have been constructed to prove the monotonic non-increase of the NMF multiplicative-update based algorithm for values of β other than $\beta = 1$ and $\beta = 2$. For instance, in (Cao et al. 1999) it is proven that $D_0(X||WH)$ is non-increasing under the corresponding multiplicative update scheme for $\beta = 0$. In (Kompass 2007), it is proven that $D_\beta(X||WH)$ is non-increasing using the multiplicative updates with costs corresponding to values of $\beta \in [1, 2]$. In (Févotte and Idier 2011), auxiliary functions are developed and used to prove the monotonicity of the NMF algorithm for cost functions corresponding to values of $\beta \in (0, 1)$. Thus, combining all of these results, it is known that the multiplicative NMF algorithm described above in Equation (2.28) is monotonic for $\beta \in [0, 2]$ (Févotte and Idier 2011). Although

the monotonic non-increase of the NMF algorithm using a cost function with values of β outside these ranges has not yet been proven, it has been observed in practice, as discussed in (Févotte and Idier 2011).

Structural Reformations and Constrained Factorizations

There exist several variations of NMF which offer improvements over the classical NMF when applied to SC-MSS. In (Smaragdis 2004) for instance, the matrix factor, W, is extended in dimension to include time-varying spectral templates. This structural reformation is called non-negative matrix factor deconvolution (NMFD). By incorporating a temporal dimension into the spectral templates, NMFD addresses some of the shortcomings of the classic approach to NMF, mainly the assumption that the spectral templates are static. Although the structure of NMFD does not inherently cluster notes of the same source, it does cluster templates of a source corresponding to a single note. In (Schmidt and Morup 2006), this idea is extended to include temporal and spectral information in both matrix factors, W and H, and is referred to as non-negative matrix factor 2-D deconvolution (NMF2D). This structural reformation operates on the log-frequency magnitude spectrogram and allows for individual sources to be represented as a time-varying spectral template which is then convolved in both time and frequency.

A filtering based approach has also been adopted into the NMF structure to address the issue of time-varying spectral shapes, as considered in papers such as (Hennequin et al. 2011),(Yoshii and Goto 2012), and the references therein. The methods proposed in these papers are often referred to as source/filter non-negative factorizations due to the fact that they are based on decomposing a non-negative spectrogram into a source template and a filter template along with time-varying gains.

Another popular variation on NMF is based on the inclusion of musically motivated constraints/penalty functions imposed on the matrix factors. These constraints/penalty functions can come in several forms and be applied to either W or H separately or both together. Typically, these musically motivated constraints/penalty functions take the form of enforced spectral template harmonicity and/or enforced temporal smoothness (Virtanen 2006). Various approaches to constrained NMF can be found in (Bertin et al. 2010).

2.2.9 Sparse NMF

NMF in its standard form, with no additional constraints other than element-wise non-negativity, is observed to produce a sparse representation of the data (Lee and Seung 1999). Here, an NMF is defined as being sparse if at least one matrix factor, W and/or H, is sparse, as defined in (Cichocki et al. 2009). This observed sparseness can be explained by the fact that the factors, W and H, which solve the NMF optimization problem will typically lie close to the boundary of the permissible solution space of $\mathbb{R}_{\geq 0}^{N\times K} \times \mathbb{R}_{\geq 0}^{K\times M}$ (Gillis 2011). The basic NMF, however, has since been extended to allow for the sparsity to be adjusted explicitly via an additional constraint imposed on the factorization as in (Hoyer 2004), or via an additional penalty function augmenting the objective function as in (Eggert and Korner 2004). Note that EUC-NMF with a sparseness penalty, as developed in (Eggert and Korner 2004), is closely related to the structural form of CMF and, as such, will be the version of NMF considered for the comparisons made in the experiments of Chapters 4 - 6. The sparse NMF optimization problem can be described as follows:

$$\begin{aligned} \text{Given} &: X \in \mathbb{R}_{+}^{NxM} \text{ and } K \in \mathbb{R}_{>0} \\ \text{Minimize} &: \frac{1}{2}\sum_{n,m} |[X]_{n,m} - [\widehat{X}]_{n,m}|^2 + \lambda \sum_{k,m} |[H]_{k,m}|^g \\ \text{Subject to} &: \sum_{n} [W]_{n,k}^2 = 1 \ (\forall k = 1, ..., K),\ W \in \mathbb{R}_{\geq 0}^{N\times K} \\ &\quad H \in \mathbb{R}_{\geq 0}^{K\times M} \end{aligned} \tag{2.32}$$

The first term in the minimization of Equation (2.32) corresponds to the quality of approximation of $\widehat{X} = WH$ to X, whereas the second term corresponds to the sparsity factor, penalizing larger values of the elements of H. Here, λ is a weighting parameter corresponding to the relative importance of the sparsity factor on the overall optimization and g is a parameter influencing the shape of the sparsity distribution. A solution to the optimization problem of Equation (2.32) can be found in (Eggert and Korner 2004) in which the updates for the W and H parameters are given as follows:

Sparse EUC-NMF Multiplicative Updates:

$$
\begin{aligned}
[H]_{k,m} &\leftarrow [H]_{k,m} \bullet \frac{[X]_{:,m}^{\top}\frac{[W]_{:,k}}{\|[W]_{:,k}\|}}{(\sum_{k'}[H]_{k',m}\frac{[W]_{:,k'}}{\|[W]_{:,k'}\|})^{\top}(\frac{[W]_{:,k}}{\|[W]_{:,k}\|})+\lambda} \\
[W]_{:,k} &\leftarrow [W]_{:,k} \bullet \frac{\sum_{m}[H]_{k,m}\Big[[X]_{:,m}+(\sum_{k'}[H]_{k',m}\frac{[W]_{:,k'}}{\|[W]_{:,k'}\|})^{\top}(\frac{[W]_{:,k}}{\|[W]_{:,k}\|})(\frac{[W]_{:,k}}{\|[W]_{:,k}\|})\Big]}{\sum_{m}[H]_{k,m}\Big[(\sum_{k'}[H]_{k',m}\frac{[W]_{:,k'}}{\|[W]_{:,k'}\|})+([X]_{:,m}^{\top}\frac{[W]_{:,k}}{\|[W]_{:,k}\|})(\frac{[W]_{:,k}}{\|[W]_{:,k}\|})\Big]}
\end{aligned}
\tag{2.33}
$$

Note that the form of the multiplicative updates in Equation (2.33) correspond specifically to the case for which the sparsity shaping parameter is set to $g = 1$ and the L_2 norm is used to normalize the columns of W, as established in (Eggert and Korner 2004). Finally, the sparse NMF multiplicative update optimization scheme used throughout the experiments of Chapters 4 - 6 can be summarized as follows:

Algorithm 2 SPARSE NMF WITH MULTIPLICATIVE UPDATES

Input: $X \in \mathbb{R}_{\geq 0}^{N\times M}$ and $K \in \mathbb{R}_{>0}$

Output: $WH \approx X, W \in \mathbb{R}_{\geq 0}^{N\times K}, H \in \mathbb{R}_{\geq 0}^{K\times M}$ and $\sum_{n}[W]_{n,k}^{2} = 1$ $(\forall k = 1, ..., K)$

Initialize W, H such that $W \in \mathbb{R}_{>0}^{N\times K}$ and $H \in \mathbb{R}_{>0}^{K\times M}$

while stopping criteria not met **do**

Update H

$$[H]_{k,m} \leftarrow [H]_{k,m} \bullet \frac{[X]_{:,m}^{\top}\frac{[W]_{:,k}}{\|[W]_{:,k}\|}}{(\sum_{k'}[H]_{k',m}\frac{[W]_{:,k'}}{\|[W]_{:,k'}\|})^{\top}(\frac{[W]_{:,k}}{\|[W]_{:,k}\|})+\lambda}$$

Update W

$$[W]_{:,k} \leftarrow [W]_{:,k} \bullet \frac{\sum_{m}[H]_{k,m}\Big[[X]_{:,m}+(\sum_{k'}[H]_{k',m}\frac{[W]_{:,k'}}{\|[W]_{:,k'}\|})^{\top}(\frac{[W]_{:,k}}{\|[W]_{:,k}\|})(\frac{[W]_{:,k}}{\|[W]_{:,k}\|})\Big]}{\sum_{m}[H]_{k,m}\Big[(\sum_{k'}[H]_{k',m}\frac{[W]_{:,k'}}{\|[W]_{:,k'}\|})+([X]_{:,m}^{\top}\frac{[W]_{:,k}}{\|[W]_{:,k}\|})(\frac{[W]_{:,k}}{\|[W]_{:,k}\|})\Big]}$$

iter = iter + 1

end while

It is important to note that even though KL-NMF and IS-NMF are known ((Virtanen 2006), (Févotte et al. 2009)) to offer superior separation performances compared to EUC-NMF, it is not the intention of this thesis to compare the best NMF algorithm against the newly proposed CMF algorithm. Rather, as explained in (King and Atlas 2011), the intention is to observe how introducing phase estimation into matrix factorization affects source separation performance. As such, the closest variant of NMF to CMF will be used. As CMF uses the Euclidean distance measure of factorization approximation and an L_1 - norm sparsity penalty term, the EUC-NMF with an L_1-norm sparsity penalty term will be used as grounds for comparison.

2.3 Complex Matrix Factorization

Throughout the description of NMF-based SC-MSS in Section 2.2.4, the adopted spectral mixing model set the sum of the spectrograms of the sources as being equal to the spectrogram of the mixture: $X = \sum_{p=1}^{P} S_p$. As discussed in Subsection 2.1.3, however, this model holds only approximately due to the fact that the additivity of the $STFT$ of the sources does not imply the additivity of the spectrogram of the sources. This previously established concept can be restated mathematically, as follows:

$$X = |\mathbb{X}| = |\sum_{p=1}^{P} \mathbb{S}_p| \neq \sum_{p=1}^{P} |\mathbb{S}_p| = \sum_{p=1}^{P} S_p \tag{2.34}$$

Another way to approach this issue is to understand that the sum of the spectrogram of the source does not equal the spectrogram of the mixture unless only one source is active at a time or overlapping sources are perfectly in phase, (Parry and Essa 2007b). This, however, is highly unlikely for real-world situations. As outlined in (Parry and Essa 2007b), consider the n^{th} frequency bin and m^{th} time frame of the $STFT$ of a mixture resulting from overlapping sources, as follows:

$$
\begin{aligned}
[\mathbb{X}]_{n,m} &= \sum_{p=1}^{P} [\mathbb{S}_p]_{n,m} \\
|[\mathbb{X}]_{n,m}]|^2 &= \langle [\mathbb{X}]_{n,m}, [\mathbb{X}]_{n,m} \rangle \\
|[\mathbb{X}]_{n,m}]|^2 &= \langle [\sum_{p=1}^{P} [\mathbb{S}_p]_{n,m}, \sum_{q=1}^{P} [\mathbb{S}_q]_{n,m} \rangle \\
|[\mathbb{X}]_{n,m}]|^2 &= \Re(\sum_{p=1}^{P} \sum_{q=1}^{P} \langle [\mathbb{S}_p]_{n,m}, [\mathbb{S}_q]_{n,m} \rangle) \\
|[\mathbb{X}]_{n,m}]|^2 &= \sum_{p=1}^{P} \sum_{q=1}^{P} \Re(\langle [\mathbb{S}_p]_{n,m}, [\mathbb{S}_q]_{n,m} \rangle) \\
|[\mathbb{X}]_{n,m}]| &= \sqrt{\sum_{p=1}^{P} \sum_{q=1}^{P} |[\mathbb{S}_p]_{n,m}||[\mathbb{S}_q]_{n,m}| \cos([\Phi_p]_{n,m} - [\Phi_q]_{n,m})}
\end{aligned}
\tag{2.35}
$$

where $z_1, z_2 \in \mathbb{C}$, $\Re(z_1)$ refers the real part of z_1, and $[\Phi_p]_{n,m} - [\Phi_q]_{n,m}$ represents the difference in phase between source p and source q at the n^{th} frequency bin and m^{th} time frame. As discussed in (Parry and Essa 2007a), if only one source, say S_p, is active at the n^{th} frequency bin and m^{th} time frame then the last line of Equation (2.35) yields: $|[\mathbb{X}]_{n,m}]|^2 = |[\mathbb{S}_p]_{n,m}]|^2$. Similarly, if $[\Phi_p]_{n,m} = [\Phi_q]_{n,m}$ (i.e., the sources share the same phase) then $|[\mathbb{X}]_{n,m}]|^2 = \sum_{p=1}^{P} \sum_{q=1}^{P} |[\mathbb{S}_p]_{n,m}||[\mathbb{S}_q]_{n,m}| = (\sum_{p=1}^{P} |[\mathbb{S}_p]_{n,m}|)^2$. Thus, for both these conditions, the NMF mixture model assumptions hold. However, for mixtures of two distinct sources, these conditions rarely hold in practice. For instance, as discussed in (Woodruff et al. 2008), Western music often consist of overlapping notes which lie at pitch intervals that are very close to simple integer ratios ($\approx \frac{5}{4}, \frac{4}{3}, \frac{3}{2}, \frac{5}{3}$). As such, many overlapping notes will share common harmonics, hence possess strong spectral overlap within a mixture. As will be examined in the experiments of Chapters 4 - 6, strong spectral overlap may also occur if the spectral resolution is too low and unable to resolve closely spaced harmonics of overlapping sources (e.g., two overlapping notes spaced a tone or semitone apart).

2.3.1 Principles of CMF

CMF was introduced in 2009 (Kameoka 2009) as a structural reformulation of NMF, which incorporates the phase of the observed mixture directly into the factorization model. Mathematically, CMF can be described as follows:

$$\begin{aligned} \text{Given} &: [\mathbb{X}]_{n,m} \in \mathbb{C} \text{ and } K \in \mathbb{R}_{>0} \\ \text{Factorize} &: [\mathbb{X}]_{n,m} \approx [\widehat{\mathbb{X}}]_{n,m} = \sum_{k=1}^{K} [\widehat{C}_k]_{n,m} = \sum_{k=1}^{K} [W]_{n,k}[H]_{k,m} \exp(i[\Phi]_{n,k,m}) \\ \text{Subject to} &: \sum_{n} [W]_{n,k} = 1 \ (\forall k = 1, ..., K),\ W \in \mathbb{R}_{\geq 0}^{N \times K} \\ &\quad H \in \mathbb{R}_{\geq 0}^{K \times M} \text{ and } \Phi \in \mathbb{R}^{N \times K \times M} \end{aligned} \tag{2.36}$$

Here, $|[\hat{C}_k]_{n,m}| = [W]_{n,k}[H]_{k,m}$, which parallels the NMF as typically presented. However, the factorization now incorporates the missing phase information, $\exp(i[\phi]_{n,k,m})$, at each time-frequency point, for each component. Note that the inclusion of the phase term, $\exp(i[\phi]_{n,k,m})$ prevents CMF from being expressed as a true factorization. However, it is referred to as such due to the similarities in structure shared with NMF. Let $\theta = \{W, H, \Phi\}$, where W and H are as defined in the context of NMF and Φ is the 3-dimensional $N \times K \times M$ matrix with entries $\phi_{n,k,m}$. In order to estimate the optimal values of $\theta = \{W, H, \Phi\}$, (Kameoka 2009) proposes the following optimization problem, which bares close resemblance to the sparse NMF formulation of Equation (2.32):

$$\begin{aligned} \text{Given} &: \mathbb{X} \in \mathbb{C}_{+}^{NxM} \text{ and } K \in \mathbb{R}_{>0} \\ \text{Optimize} &: \min_{W,H,\Phi} C(\theta) = \sum_{n,m} |[\mathbb{X}]_{n,m} - [\widehat{\mathbb{X}}]_{n,m}|^2 + 2\lambda \sum_{n,m} |[H]_{m,n}|^g \\ \text{Subject to} &: \sum_{n} [W]_{n,k} = 1 \ (\forall k = 1, ..., K),\ W \in \mathbb{R}_{\geq 0}^{N \times K} \\ &\quad H \in \mathbb{R}_{\geq 0}^{K \times M} \text{ and } \Phi \in \mathbb{R}^{N \times K \times M} \end{aligned} \tag{2.37}$$

where λ is again a weighting parameter corresponding to the influence of the sparsity factor on the optimization and g is a parameter influencing the shape of the sparsity

distribution. In a similar fashion to the development of NMF, as described in (Lee et al. 2001), an auxiliary functions was proposed from which multiplicative updates were derived for parameter estimation. The definition of an auxiliary function presented in (Kameoka 2009) is as follows:

Majorizing Auxiliary Function Revisited: $G^{+}(\theta, \bar{\theta})$, is a majorizing auxiliary function of $G(\theta)$ provided:

$$G(\theta) = \min_{\bar{\theta}} G^{+}(\theta, \bar{\theta}) \tag{2.38}$$

Note that this definition of an auxiliary function, which is discussed in (Leeuw 1994), is simply a more generalized way of viewing the auxiliary function approach presented in Subsection 2.2.8, in the context of NMF. As proven in (Kameoka 2009) and (Nakano et al. 2010), $G(\theta)$ is non-increasing under the updates:

$$\begin{aligned} &\bullet \quad \bar{\theta}^{(\rho+1)} \leftarrow arg \min_{\bar{\theta}} G^{+}(\theta^{(\rho)}, \bar{\theta}) \\ &\bullet \quad \theta^{(\rho+1)} \leftarrow arg \min_{\theta} G^{+}(\theta, \bar{\theta}^{(\rho+1)}) \end{aligned} \tag{2.39}$$

Update equations for the parameters $\theta = \{W, H, \Phi\}$ of the CMF are constructed using the following auxiliary functions with corresponding auxiliary variables $\bar{\theta} = \{\bar{\mathbb{X}}, \bar{H}\}$:

$$\begin{aligned} C(\theta, \bar{\theta}) = &\sum_{n,k,m} \frac{|[\bar{\mathbb{X}}]_{n,k,m} - [W]_{n,k}[H]_{k,m} \exp(i[\Phi]_{n,k,m})|^2}{[B]_{n,k,m}} \\ &+ \lambda \sum_{k,m} (g|[\bar{H}]_{k,m}|^{g-2}[H]_{k,m}^2 + 2|[\bar{H}]_{k,m}|^g - g|[\bar{H}]_{k,m}|^g) \end{aligned} \tag{2.40}$$

where $\sum_{k=1}^{K} [\bar{\mathbb{X}}]_{n,k,m} = [\mathbb{X}]_{n,m}$, $0 \leq g \leq 2$, and $\sum_{k=1}^{K} [B]_{n,k,m} = 1$, $B_{n,k,m} \in \mathbb{R}_{\geq 0}^{N,K,M}$.

Optimizing the auxiliary function over the the auxiliary variables yields the following auxiliary variable updates:

$$[\bar{\mathbb{X}}]_{n,k,m} = [W]_{n,k}[H]_{k,m} \exp(i[\Phi]_{n,k,m}) + [B]_{n,k,m}([\mathbb{X}]_{n,m} - [\widehat{\mathbb{X}}]_{n,m}) \tag{2.41}$$

$$[\bar{H}]_{k,m} = H_{k,m} \tag{2.42}$$

Next, optimizing the auxiliary function over the primary variables yields the following primary variables updates:

$$[W]_{n,k} = \frac{\sum_{m=1}^{M} \frac{[H]_{k,m}}{[B]_{n,k,m}} \Re[\bar{\mathbb{X}}_{n,k,m} \exp(-i[\Phi]_{n,k,m})]}{\sum_{m=1}^{M} \frac{[H]_{k,m}^2}{[B]_{n,k,m}}} \tag{2.43}$$

$$[H]_{k,m} = \frac{\sum_{n=1}^{N} \frac{[W]_{k,m}}{[B]_{n,k,m}} \Re[\bar{\mathbb{X}}_{n,k,m} \exp(-i[\Phi]_{n,k,m})]}{\sum_{n=1}^{N} \frac{[W]_{k,m}^2}{[B]_{n,k,m}} + \lambda g |[\bar{H}]_{k,m}|^{g-2}} \tag{2.44}$$

$$\exp(i[\Phi]_{n,k,m}) = \frac{[\bar{\mathbb{X}}]_{n,k,m}}{|[\bar{\mathbb{X}}]_{n,k,m}|} \tag{2.45}$$

Finally, as discussed in (Kameoka 2009), the parameter B is updated as follows:

$$[B]_{n,k,m} = \frac{[W]_{n,k}[H]_{k,m}}{\sum_{k=1}^{K} [W]_{n,k}[H]_{k,m}} \tag{2.46}$$

The entire CMF algorithm is presented below in Algorithm 3.

Algorithm 3 COMPLEX MATRIX FACTORIZATION

Input: $\mathbb{X} \in \mathbb{C}_{+}^{NxM}$ and $K \in \mathbb{R}_{>0}$

Output: W, H, Φ s.t $[\mathbb{X}]_{n,m} \approx \sum_{k=1}^{K} [W]_{n,k}[H]_{k,m} \exp(i[\Phi]_{n,k,m})$

$W \in \mathbb{R}_{\geq 0}^{NxK}$, $H \in \mathbb{R}_{\geq 0}^{KxM}$, $\Phi \in \mathbb{R}^{N \times K \times M}$ and $\sum_{n}[W]_{n,k} = 1$ $(\forall k = 1, ..., K)$

Initialize W, H, Φ, such that $W \in \mathbb{R}_{>0}^{NxK}$, $H \in \mathbb{R}_{>0}^{KxM}$ and $\Phi = \frac{\mathbb{X}}{|\mathbb{X}|}$

while stopping criteria not met **do**

Compute B

$$[B]_{n,k,m} = \frac{[W]_{n,k}[H]_{k,m}}{\sum_{k}[W]_{n,k}[H]_{k,m}}$$

Compute $\bar{\mathbb{X}}$

$$[\bar{\mathbb{X}}]_{n,k,m} = [W]_{n,k}[H]_{k,m} \exp(i[\Phi]_{n,k,m}) + [B]_{n,k,m}([\mathbb{X}]_{n,m} - [\widehat{\mathbb{X}}]_{n,m})$$

Compute $\bar{H}$

$$[\bar{H}]_{k,m} = H_{k,m}$$

Compute Φ

$$\exp(i[\Phi]_{n,k,m}) = \frac{[\bar{\mathbb{X}}]_{n,k,m}}{|[\bar{\mathbb{X}}]_{n,k,m}|}$$

Compute W

$$[W]_{n,k} = \frac{\sum_{m} \frac{[H]_{k,m}}{[B]_{n,k,m}} \Re[\bar{\mathbb{X}}_{n,k,m} \exp(-i[\Phi]_{n,k,m})]}{\sum_{m} \frac{[H]_{k,m}^{2}}{[B]_{n,k,m}}}$$

Compute H

$$[H]_{k,m} = \frac{\sum_{n} \frac{[W]_{n,k}}{[B]_{n,k,m}} \Re[\bar{\mathbb{X}}_{n,k,m} \exp(-i[\Phi]_{n,k,m})]}{\sum_{n} \frac{[W]_{n,k}^{2}}{[B]_{n,k,m}} + \lambda g |[\bar{H}]_{k,m}|^{g-2}}$$

Normalize W and scale H accordingly

iter = iter + 1

end while

2.3.2 CMF-Based Single-Channel Musical Source Separation

Following the approach of NMF-based SC-MSS procedure, described in Subsection 2.2.4, CMF synthesis/filter-based source estimates can be derived, as follows:

CMF Synthesis-Based Source Estimate:

$$[\widehat{\mathbb{S}}_p]_{n,m}^{\text{SYNTH}} = \sum_{k \in \mathcal{K}_p} [\widehat{C}_k]_{n,m} = \sum_{k \in \mathcal{K}_p} [W]_{n,k}[H]_{k,m} \exp(i[\Phi]_{n,k,m}) \tag{2.47}$$

$$\widehat{\mathbf{s}}_p^{\text{SYNTH}} = ISTFT\{\widehat{\mathbb{S}}_p^{\text{SYNTH}}\} \tag{2.48}$$

CMF Filter-Based Source Estimate:

$$[\widehat{\mathbb{S}}_p]_{n,m}^{\text{FILT}} = \frac{|\sum_{k \in \mathcal{K}_p} [\widehat{C}_k]_{n,m}|}{\sum_p |\sum_{k \in \mathcal{K}_p} [\widehat{C}_k]_{n,m}|} \bullet [\mathbb{X}]_{n,m} = \frac{|\sum_{k \in \mathcal{K}_p} [W]_{n,k}[H]_{k,m} \exp(i[\Phi]_{n,k,m})|}{\sum_p |\sum_{k \in \mathcal{K}_p} [W]_{n,k}[H]_{k,m} \exp(i[\Phi]_{n,k,m})|} \bullet [\mathbb{X}]_{n,m} \tag{2.49}$$

$$\widehat{\mathbf{s}}_p^{\text{FILT}} = ISTFT\{\widehat{\mathbb{S}}_p^{\text{FILT}}\} \tag{2.50}$$

As defined in the context of NMF, $\mathcal{K}_p$ represents the set of all component vector indices corresponding to source p. Note that, by using the synthesis-based source estimation method, the estimate phase parameter is incorporated directly into the source estimates.

2.3.3 Previous CMF-Based Experiments

The following subsection details previous CMF-based experiments, all of which focus primarily on applying CMF to mixtures of speech samples. There exists a substantial gap in the literature regarding the application of CMF to mixtures of musical samples. As such, although they do not focus on simple musical tones, as is considered in this thesis, the

previous CMF-based speech studies offer valuable insight into the behaviour of the CMF algorithm when applied to spectrograms of audio data.

In (Kameoka 2009), the author presents two experiments in which CMF is used to decompose a spectrogram of speech excerpts from the ATR B-set speech database. The first experiment involved decomposing the spectrogram of a single female voice signal. The number of extracted components was set to $K = 10$. The authors observed that the spectra contained in the columns of W clearly represented distinct harmonic patterns in a similar fashion to what is typically observed using NMF. The second experiment focused on applying CMF to a mixture of two speech speech signals, both of which were female. The number of extracted components was set to $K = 30$ for the second experiment. The true unmixed signals of both speakers were used to cluster the columns of W into the respective sources (adjusting the rows of H and the corresponding phase accordingly), based on some measure of distance[1] between the columns of W and the columns of the unmixed sources. The authors observed an improvement in the signal to noise ratio[2] (S/N) of both extracted sources over the original mixture, suggesting that the extracted sources correspond to a single voice spectrum. Although not explicitly stated, the S/N in (Kameoka 2009) is taken to mean ten times the log of the power ratio of the signal to the unwanted noise/interference. For both experiments, the sparsity weight and shaping parameter were set to $\lambda = \sum_{n,m} |[X]_{n,m}|^2 / K^{1-\frac{g}{2}} \times 10^{-5}$ and $g = 1.2$.

As discussed in (Le Roux et al. 2009), an $STFT$ is a redundant signal representation which has a particular structure. Thus, not all sets of complex numbers correspond to the $STFT$ of a real-world signal. Mathematically this can be expressed as follows:

$$\mathbb{Q} \neq STFT(ISTFT(\mathbb{Q})) \tag{2.51}$$

where $\mathbb{Q} \in (C)^{N \times M}$ represents some set of complex numbers in the complex time-frequency domain. STFT's for which equality does hold in Equation (2.51) are referred to as being "consistent". Consistency constraints are developed in (Le Roux et al. 2009) and incorporated into the CMF optimization problem as a penalty function in attempt to further improve the model by favouring a solution in which the extracted $STFT$ estimates of ev-

[1] The measure of distance remained unspecified in the paper.

[2] Here the signal to noise ratio is abbreviated as S/N to distinguish it from the source to noise ratio (SNR) discussed in Subsection 3.3.7.

ery component are consistent. An experiment was conducted using a mixture constructed from chimes sounds and male speech taken from the TIMIT database. A supervised source separation task was performed in which K = 20 spectral bases were first extracted via factorization of the spectrograms of the unmixed chime and male speech training samples. A test mixture was then constructed from an unused portion of the chime and male speech samples. The spectral bases extracted in the training phase were concatenated to form W, which remained fixed throughout the factorization. The sparsity parameters were set as in (Kameoka 2009). The author demonstrates that CMF with consistency constraints offered an increase in performance, with respect to the S/N performance measure, compared to the results obtained using an NMF-based and CMF-based separation procedure without the use of consistency constraints. Consistency constraints will be discussed further in Section 3.1.

CMF-based speech separation and NMF-based speech separation were also compared in (King and Atlas 2010), in which a new training approach is suggested. This training approach is referred to as "copy-to-train" initialization method. Unlike the "no-train" initialization method—which initializes the basis matrix, W, with positive random numbers—or the "factorize-to-train" initialization method—which initializes the basis matrix, W, with the results from a matrix decomposition on the spectrogram of the training sources—the "copy-to-train" initialization method initializes the basis matrix, W, with an exact copy of the spectrograms of the training sources concatenated into a single matrix. Three tests were carried out in order to compare CMF with NMF using the "copy-to-train" and "factorize-to-train" initialization methods. Training samples and test mixtures were created using hand-transcribed speech from one male and one female speaker, which were extracted from a television broadcast news show. Five target to masker ratios (TMR) were considered for the testing mixture (-6, -3, 0, 3, 6 dB) and three levels of the number of extracted bases in the "factorize-to-train" initialization method were considered ($K = 50, 100, 200$). When the NMF-based and CMF-based separation procedures were applied as a preprocessing step to an automatic speech recognition system, it was found that the "copy-to-train" method outperformed the "factorize-to-train" initialization method for CMF for all levels of K, but only outperform the "factorize-to-train" method for NMF when $K = 50$. It was also demonstrated that the CMF-based separation procedure outperformed the NMF-based separation procedure using the "copy-to-train" initialization method when applied as a preprocessing step to an automatic speech recognition system for all levels of TMR

considered.

Single-channel source separation using CMF is also discussed in (King and Atlas 2011). A larger dataset than the one considered in (King and Atlas 2010) is used, consisting of male and female speech samples taken from the Boston University Radio News corpus. Six levels of the number of extracted bases in the "factorize-to-train" initialization method were considered ($K = 50, 100, 200, 400, 800, 1600$) and the TMR of the test mixture was set to 0 dB. The sparsity weight was set to $\lambda = 0.01$, which was observed to work well with the test data considered. It was observed by the authors of (King and Atlas 2011) that the results of the CMF-based separation procedure produced separated signals that sounded very natural and did not suffer from any "musical noise", as if the unwanted interfering signal were manually attenuated from a mixing board. It was also observed that the CMF-based separation procedure was able to better approximate the spectrum of the the original signals in higher frequency ranges compared to the NMF-based separation procedure. It was discovered that, when applied as a preprocessing step to automatic speech recognition systems, CMF outperformed NMF on average. Unlike the initial experiment presented in (King and Atlas 2010), however, it was observed that the "factorize-to-train" initialization method, using K = 1600 bases, outperformed the "copy-to-train" initialization method. In fact, it was discovered that the "copy-to-train" initialization method only performed the best in 6% of the clips on average. Note that the results of the Experiments presented in (King and Atlas 2010) and (King and Atlas 2011) are also presented in (King 2012)

A variant of CMF, referred to as Complex Matrix Factorization with Intra-Source Additivity (CMFWISA), is presented in (King 2012) and implemented in (King and Atlas 2012). CMFWISA limits the previously formulated CMF to include only one unique phase matrix estimate per source. In doing so, CMFWISA significantly reduces the parameterization of the factorization, offers a more realistic representation of the unmixed sources, and lowers both the memory and computation time required for the implementation of the factorization. Three experiments are carried out in (King 2012) which investigate the performance of CMFWISA as it compares to NMF and CMF when applied to speech separation using the same database described in (King and Atlas 2011). Five target-to-masker ratios (TMR) were considered for the testing mixture (-10, -5, 0, 5, 10 dB) and only the "copy-to-train" method was used for initialization of the spectral templates. The source to distortion (SDR), source to interference (SIR), and source to artifact (SAR) performance measures were used to quantitatively assess the quality of separation obtained for each

factorization-based separation procedures, (Vincent et al. 2006). These three performance measures describe a global measure of distortion, a measure of how much interfering sources are rejected, and a measure of how much unwanted distortions and artifacts are rejected, respectively. The SDR, SIR, and SAR performance measures will also be used throughout this thesis and will be discussed further in Subsection 3.3.7. The first experiment examined the CMFWISA-based separation procedure when the phases of the individual sources are known to be estimated correctly. It was demonstrated that under these oracle phase condition, the CMFWISA-based separation procedure outperformed the NMF-based separation procedure, yielding higher SDR, SIR, and SAR performance measures for both source estimation methods (synthesis-based and filtering-based). CMFWISA was adapted in (King 2012) to include consistency constraints as a penalty function based on the work of (Le Roux et al. 2009). Unlike (Le Roux et al. 2009), however, the approach taken in (King 2012) enforces the consistency of the extracted $STFT$ estimates of each source, as opposed to the consistency of the extracted $STFT$ estimates of each component. An experiment was conducted to see how the performance of the CMFWISA-based separation procedure was affected when the consistency penalty was applied to the updates for the temporal activations matrix, H, and the updates for the phase parameter, Φ. Three levels of the consistency weight (0, 10, 100) were considered. It was observed that applying a consistency penalty on the updates of the phase parameter improved the CMWISA performance, offering higher SIR and SDR performance measures using the filter-based source estimation method and higher SIR performance measures using the synthesis-based source estimation method. It was also observed, however, that the SAR performance measure decreased under the presences of a consistency penalty for both source estimation methods. Additionally, it was concluded that adding a consistency penalty to the updates of the temporal activations matrix, H, degraded the speech-separation performance, lowering all three performance measures. A third experiment was conducted to examine how the performance of the CMFWISA-based separation procedure fared against the NMF-based and CMF-Based separation procedures. A notable conclusions drawn from this experiment is that the CMFWISA-based separation procedure seemed to offer a middle ground between the higher SIR, but lower SAR, performance measures obtained using the NMF-based separation procedure and the lower SIR, but higher SAR, performance measures obtained using the CMF-based separation procedure. As suggested by the author of (King 2012), the biggest room for improvement lies in the estimation of the phase parameter.

Chapter 3

Pre-Experimentation Considerations:

Before proceeding to the experiments found in Chapters 4, 5, 6, a final modification of the CMF algorithm is discussed. This modification is a refinement of the consistency constraint auxiliary function, as first described in (Le Roux et al. 2009), which simplifies the calculation of the iterative CMF parameter updates under the previously established consistency constraint formulation and is intuitively grounded in the theory of signal estimation from a modified $STFT$, as presented in (Griffin and Lim 1984). This chapter also presents the definition of the $STFT/ISTFT$ used throughout the experiments, a qualitative/quantitative description of the audio dataset upon which the NMF/CMF algorithms were compared, preliminary comments on the choice of $STFT$ analysis window length, and other pre-experiment considerations, such as initialization strategies/ stopping criteria, spectrogram phase recovery, component-to-source clustering strategies, the BSS performance measures/evaluation regions, and the handling of undetected sources.

3.1 Consistency Constraint Reformulation

As discussed in Subsection 2.3.3, the consistency constraint/penalty function proposed in (Le Roux et al. 2009) is motivated by the fact that not all sets of complex numbers correspond to the $STFT$ of a real-world signal. As formulated in (Le Roux et al. 2010), consistent spectrograms can be described as the kernel of the operator $\mathcal{F}$ from $\mathbb{C}^{N\times M}$ to $\mathbb{C}^{N\times M}$ defined below in Equation 3.1 as follows:

$$[F(\mathbb{Q})]_{n,m} = [\mathcal{G}(\mathbb{Q})]_{n,m} - [\mathbb{Q}]_{n.m} \tag{3.1}$$

where $\mathcal{G}(\mathbb{Q}) = STFT(ISTFT(\mathbb{Q}))$, for any $\mathbb{Q} \in \mathbb{C}^{N\times M}$. Consistency is presented in the form of a soft constraint imposed on the CMF optimization problem in (Le Roux et al. 2009) and (Le Roux et al. 2011). The CMF optimization problem with the addition of the consistency constraint as a penalty function can be described mathematically as follows:

$$\begin{aligned}
&\text{Given} : \mathbb{X} \in \mathbb{C}^{NxM} \text{ and } K \in \mathbb{R}_{>0} \\
&\text{Optimize} : \min_{W,H,\Phi} C(\theta) = \sum_{n,m} |[\mathbb{X}]_{n,m} - [\widehat{\mathbb{X}}]_{n,m}|^2 + 2\lambda \sum_{n,m} |[H]_{n,m}|^g \\
&\qquad + \gamma \sum_{n,k,m} |\mathcal{F}(\widehat{C}_k)_{n,m}|^2 \\
&\text{Subject to} : \sum_{n} [W]_{n,k} = 1 \ (\forall k = 1, ..., K),\ W \in \mathbb{R}_{\geq 0}^{N\times K} \\
&\qquad H \in \mathbb{R}_{\geq 0}^{K\times M} \text{ and } \Phi \in \mathbb{R}^{N\times K\times M}
\end{aligned} \tag{3.2}$$

The first two terms in the optimization problem described above, in Equation (3.2), correspond to the cost of the factorization approximation and the sparsity penalty function, respectively. The third term corresponds to the consistency constraint penalty function, where the objective now is to minimize the sum of the square of the modulus of the difference between the k^{th} estimated component, $[\hat{C}_k]_{n,m} = [W]_{n,k}[H]_{k,m}\exp(i[\Phi]_{n,k,m})$ and $\mathcal{G}(\hat{C}_k)_{n,m}$ at each time-frequency point. Here, γ is a scaling factor, weighting the relative importance of the penalty function in an analogous fashion to the λ scaling factor, which weights the relative importance of the sparsity penalty function.

As proven in (Griffin and Lim 1984) and discussed in (Le Roux et al. 2010), under certain $STFT/ISTFT$ conditions, namely that the windowed overlap-add (OLA) procedure is used with additional conditions imposed on the windowing function, discussed below in Subsection 3.1.1, the closest consistent spectrogram to $\mathbb{Q}$ in the least-squared sense is $G(\mathbb{Q})$. Mathematically, this fact can be expressed as follows:

$$\sum_{n,m} |[\mathcal{G}(\mathbb{Q})]_{n,m} - [\mathbb{Q}]_{n,m}|^2 = \min_{\bar{\mathbb{L}} \in Ker(\mathcal{F})} \sum_{n,m} |[\bar{\mathbb{L}}]_{n,m} - [\mathbb{Q}]_{n,m}|^2 \tag{3.3}$$

Based on this result, a refined auxiliary function for the consistency constraint penalty function described in Equation (3.2) is developed in (Le Roux et al. 2010) and (Le Roux et al. 2011) in the context of CMF, as follows:

$$C(\theta, \bar{\theta})_{\text{CON}} = \sum_{n,k,m} |[\bar{\mathbb{L}}]_{n,k,m} - [W]_{n,k}[H]_{k,m} \exp(i[\Phi]_{n,k,m})|^2 \tag{3.4}$$

where $\bar{\mathbb{L}} \in Ker(\mathcal{F})$ serves as the auxiliary variable and the "CON" subscript is used to indicate that $C(\theta, \bar{\theta})_{\text{CON}}$ represents only the consistency penalty portion of the entire CMF auxiliary function and must be later combined with the auxiliary functions developed for the factorization approximation and the sparsity penalty function, defined in Equation (2.40).

Following the MM-algorithm, as discussed in Subsection 2.3.1, the optimization of the primary variables, W, H, Φ, yields the final update scheme, which will be used throughout the experiments of Chapters 4 - 5. Note that due to the incorporation of the newly proposed phase-constraint, the update scheme is slightly modified for the third experiment, conduced in Chapter 6.

3.1.1 Window Specifications

As mentioned above and discussed in depth in (Griffin and Lim 1984) and (Le Roux et al. 2010), in order for the result of Equation (3.3), which underlies the theory behind the newly refined consistency constraint auxiliary function, to hold/be simplified in calculation, the windowing function, w, used for the $STFT/ISTFT$ analysis/synthesis should satisfy the following conditions:

1) The same windowing function should be used for both analysis and (OLA) synthesis

2) The windowing function should be normalized such that: $\sum_{m=-\infty}^{\infty} w^2(m\tilde{H} - l) = 1$

As suggested in (Griffin and Lim 1984), the following windowing function, which is formally regarded as a modified Hann window, satisfies the properties outlined above provided the window length, L, is a multiple of four times the hop size, $\tilde{H}$. As such, the modified Hann window with a window length of $L = 4\tilde{H}$ will be used as the STFT analysis/synthesis window for the experiments covered in this thesis.

Modified Hann Window:

$$w(l) = \begin{cases} \frac{2\sqrt{\frac{L}{\tilde{H}}}}{\sqrt{4(0.5)^2+2(-0.5)^2}}[(0.5) + (-0.5)\cos(\frac{2\pi l}{L} + \frac{\pi}{L})] & \text{if } 0 \leq l \leq L \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

More information regarding the theory behind the construction of the modified Hann window and its roll in the newly refined consistency constraint can be found in (Griffin and Lim 1984), (Le Roux et al. 2010), and (Le Roux et al. 2011).

3.2 Current CMF Update Scheme

Algorithm: 4, found below, details the CMF with refined consistency constraints[1], which shall be used for the experiments discussed in Chapters 4 - 5.

[1]Note that the newly refined consistency constraint auxiliary function can be viewed as a specific instant of the old consistency constraints auxiliary function developed in (Le Roux et al. 2009) and (King 2012), for which the parameter "a", where $a = (\sum_{n',m'} A_{n,m,n',m'})^2$, as defined in equation (7) of (Le Roux et al. 2009) and equation (4.42) of (King 2012), is set to unity. Here, $[A]_{m,n,m',n'}$ is the matrix representation of $[F(\mathbb{Q})]_{n,m}$, expressed mathematically as $[F(\mathbb{Q})]_{n,m} = \sum_{n',m'} [A]_{n,m,n',m'}[\mathbb{Q}]_{n',m'}$.

Algorithm 4 CMF with Consistency Constraint

Input: $\mathbb{X} \in \mathbb{C}_{+}^{NxM}$ and $K \in \mathbb{R}_{>0}$

Output: W, H, Φ s.t $[\mathbb{X}]_{n,m} \approx \sum_{k=1}^{K} [W]_{n,k}[H]_{k,m} \exp(i[\Phi]_{n,k,m})$

$W \in \mathbb{R}_{\geq 0}^{NxK}$, $H \in \mathbb{R}_{\geq 0}^{KxM}$, $\Phi \in \mathbb{R}^{N \times K \times M}$ and $\sum_{n}[W]_{n,k} = 1$ $(\forall k = 1, ..., K)$

Initialize W, H, Φ, such that $W \in \mathbb{R}_{>0}^{NxK}$, $H \in \mathbb{R}_{>0}^{KxM}$ and $\Phi = \frac{\mathbb{X}}{|\mathbb{X}|}$

while Stopping criteria not met **do**

Compute B

$[B]_{n,k,m} = \frac{[W]_{n,k}[H]_{k,m}}{\sum_{k}[W]_{n,k}[H]_{k,m}}$

Compute $\bar{\mathbb{X}}$

$[\bar{\mathbb{X}}]_{n,k,m} = [W]_{n,k}[H]_{k,m} \exp(i[\Phi]_{n,k,m}) + [B]_{n,k,m}([\mathbb{X}]_{n,m} - [\widehat{\mathbb{X}}]_{n,m})$

Compute $\bar{H}$

$[\bar{H}]_{k,m} = H_{k,m}$

Compute $[\bar{\mathbb{L}}]$

$[\bar{\mathbb{L}}]_{n,k,m} = \mathcal{G}(\widehat{C}_k)_{n,m}$

Compute Φ

$[\Phi]_{n,k,m} = Arg(\frac{[\bar{\mathbb{X}}]_{n,k,m}}{[B]_{n,k,m}} + \gamma[\bar{\mathbb{L}}]_{n,k,m})$

Compute W

$[W]_{n,k} = \frac{\sum_{m}[H]_{k,m}\Re[(\frac{\bar{\mathbb{X}}_{n,k,m}}{[B]_{n,k,m}} + \gamma[\bar{\mathbb{L}}]_{n,k,m}) \exp(-i[\Phi]_{n,k,m})]}{\sum_{m}[H]_{k,m}^{2}(\frac{1}{[B]_{n,k,m}} + \gamma)}$

Compute H

$[H]_{k,m} = \frac{\sum_{n}[W]_{n,k}\Re[(\frac{\bar{\mathbb{X}}_{n,k,m}}{[B]_{n,k,m}} + \gamma[\bar{\mathbb{L}}]_{n,k,m}) \exp(-i[\Phi]_{n,k,m})]}{\sum_{n}[W]_{n,k}^{2}(\frac{1}{[B]_{n,k,m}} + \gamma) + \lambda g([\bar{H}]_{k,m})^{g-2}}$

Normalize W and scale H accordingly

iter = iter + 1

end while

3.3 Analysis/Factorization/Synthesis Considerations

3.3.1 STFT/ISTFT Definition

The STFT used throughout the experiments conducted in Chapters 4 - 6 is defined as follows:

$$[STFT\{\mathbf{x}\}]_{n,m} = [\mathbb{X}]_{n,m} = \sum_{l=0}^{\tilde{N}-1} x(l + m\tilde{H})w(l)\exp(\frac{-i2\pi ln}{\tilde{N}}) \tag{3.6}$$

The ISTFT used throughout the experiments conducted in Chapters 4 - 6 is defined as follows:

$$[ISTFT\{\mathbb{X}\}](l) = x(l) = \sum_{m=-\infty}^{\infty}(\sum_{n=0}^{\tilde{N}-1}[\mathbb{X}]_{n,m}\exp(\frac{i2\pi(l - m\tilde{H})n}{\tilde{N}}))w(l + m\tilde{H}) \tag{3.7}$$

where n is the frequency bin index, m is the time frame index, $\tilde{N}$ is the length of the Discrete Fourier Transform (DFT) (in samples), $\tilde{H}$ is the window hop size (in samples). Note that the magnitude spectrogram, X, is created by taking the absolute value of the $STFT$, $\mathbb{X}$. As such, the spectrum components at each analysis frame of the spectrogram are the magnitudes of the spectral bins arising from the DFT taken for each window segment of the signal. No warping is applied to the frequency nor time axes. Only frequency bins in the set $n \in \{0, ..., N = \tilde{N}/2\}$ are considered in the factorization due to the symmetry of the DFT of real signals, namely, $|DFT\{\mathbf{x}\}(n)| = |DFT\{\mathbf{x}\}(\tilde{N} - n)|$.

3.3.2 Dataset Description

The experiments outlined in the next three chapters analyze sources extracted from the McGill University Master Samples Collection (Opolko and Wapnick 1987). Specifically, acoustic piano and classical guitar samples with 8 notes ranging from C4-C5, along a diatonic scale, for the piano and a fixed note of C4 for the guitar were considered. Both recordings are originally sampled at $Fs = 44100$ Hz and later down sampled by a factor of 4 to a rate of $Fs = 11025$ Hz using Matlab's® built in decimate function, which uses an eighth-order lowpass Chebyshev Type I filter with a cutoff frequency of $0.8*(Fs/8)$ Hz. The audio samples were also truncated in time so that each sample was exactly one second long.

A logarithmic decrease in amplitude was also enforced after 0.75 seconds which smoothly drove the magnitude of each sample to zero, eliminating any hard clipping. Both sources were normalized by their root mean square (RMS) signal levels, so that the decibel ratio of the RMS signal levels between the two sources was 0 dB, and rescaled so that maximum absolute amplitude of the mixture was 0.9 normalized units. As previously stated in Section 1.1, the dataset was intentionally limited to only 9 notes and two instruments, favouring interpretability of the factorization over the ability to extrapolate beyond the current test cases considered, which could be obtained if sampling from a larger set of instruments with a wider range of notes. The guitar source and piano source were specifically chosen because they exhibit a nice balance between possessing strong spectral similarities while still retaining clear timbral distinction, which could potentially be exploited to yield a quality separation. Again it is important to note that CMF is still in its infancy as a tool for source separation, and as such, a highly controlled test samples, consisting a limited range of sources, is favoured. That being said, it is important to restate that these test cases do represent the general interaction of harmonic notes within a musical mixture and, as such, offer valuable insight into the nature of the temporal and spectral interaction of harmonic sources commonly occurring within musical mixtures. To give a better idea of the structure of each source, the magnitude spectrogram of each piano source and the guitar source are plotted below, in figure 3.1.

The spectrograms in figure 3.1 were calculated using a 512 (46 ms @ 11025 Hz) length symmetric Hann window, a hop size of 128 (12 ms @ 11025 Hz) samples and no zero-padding. From the spectrograms of the unmixed sources, it is clear that the first source, the piano, is active from $t = 0$ s to $t = 1$ s and from $t = 2$ s to $t = 3$ s, whereas the guitar source is active from $t = 1$ s to $t = 2$ s and from $t = 2$ s to $t = 3$ s. In other words, the two sources are first played in isolation and then with 100% overlap. The piano note played during the first second was an exact duplicate of the piano note played during the final second of the mixture. Similarly, the guitar note played after the first second was an exact duplicate of the guitar note played during the final second of the mixture.

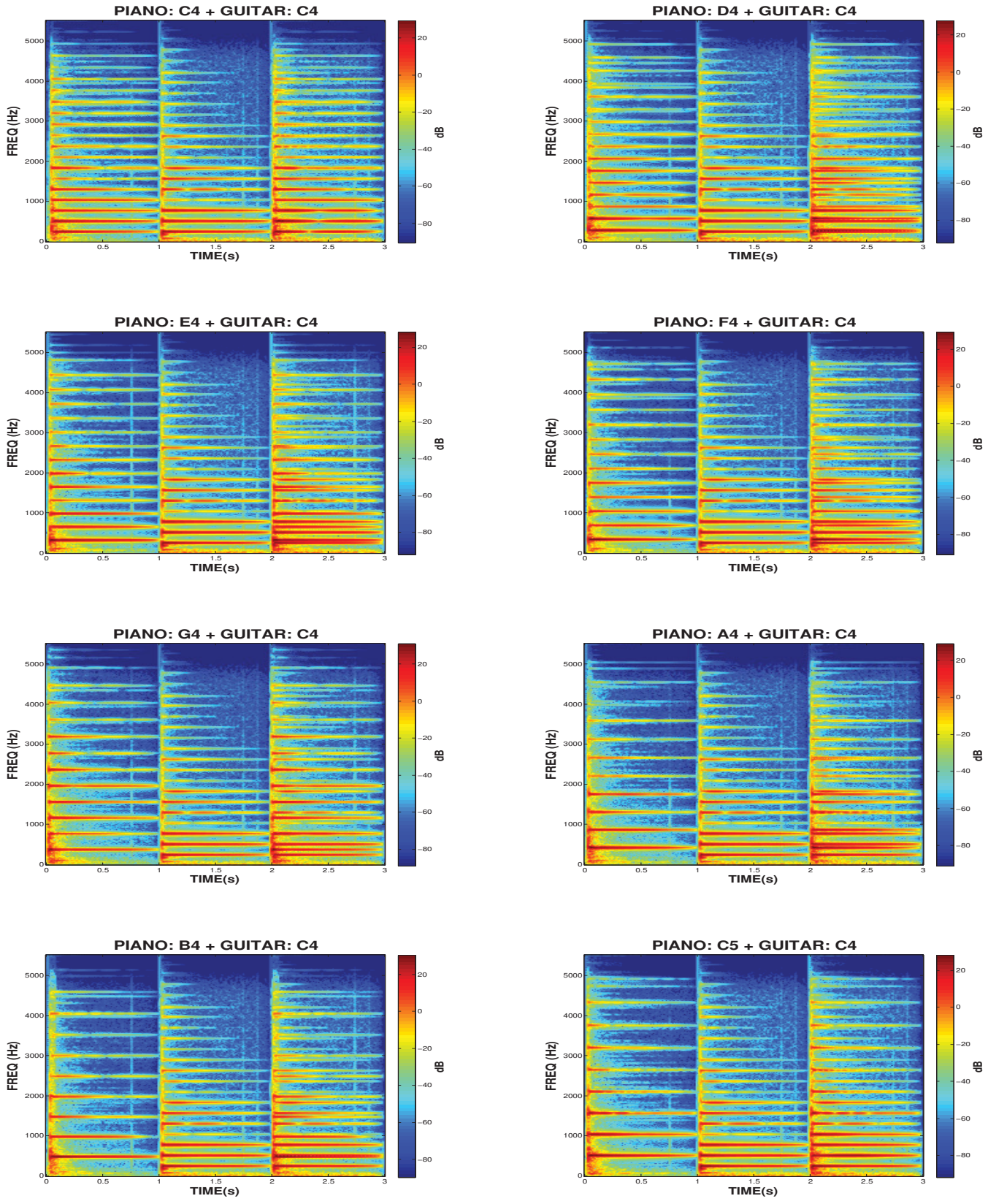

Fig. 3.1: DATASET OF PIANO/GUITAR SOURCES

3.3.3 Preliminary Window Length Considerations

The window length, L, used for the $STFT$ analysis of the musical mixture is an important factor within the NMF-based and CMF-based separation procedures as it dictates the tradeoff between temporal and spectral resolution obtained within the mixture spectrogram to be factorized. Larger window lengths offer higher spectral resolution at the cost of lower temporal resolution, whereas smaller window lengths offer higher temporal resolution at the cost of lower spectral resolution. Within the literature of NMF-based source separation applied to musical mixtures, window lengths are typically chosen within the set $L \in \{46, 93, 186, 372\}$ ms (Chen et al. 2012). Using a sampling rate of $Fs = 11025$ Hz, this corresponds to window lengths in the range of $L \in \{512, 1024, 2048, 4096\}$ samples. Poor frequency resolution could result in a high violation of the NMF mixture model assumption, especially when two high energy partials are close in frequency, which could lead to poor NMF-based separation performance. However, poor temporal resolution could also imply a high violation of the NMF mixture model assumption. For instance, when a note onset coincides with another note offset, a lack of sufficient temporal resolution may carry energy from one note onto the other across time, potentially leading to phase interactions which could have otherwise been avoided. Poor frequency and temporal resolution could also lead to lower NMF-based and CMF-based separation performance simply by allowing errors in approximation to spread over a greater frequency or temporal region, respectively. This may result in a smearing of the frequencies or smearing of the note onsets/offsets in the final source estimate reconstruction. The choice of window length and its effect on the performance of the NMF-based and CMF-based separation procedures remains an open problem and shall be investigated in the experiments of Chapter 4 - 6.

3.3.4 Initialization Strategy

As originally presented in (Lee and Seung 1999), initializations of the elements of the matrix factors could simply be taken to be any random positive numbers. Due to the form of the multiplicative updates, a zero valued entry will always remain zero. As such, the random initializations should be void of any zero elements, which will never update to anything other than zero and may contribute to a poor approximation (Wild 2003). The NMF multiplicative update scheme is highly sensitive to the choice of initialization and both the rate of convergence and the quality of approximation could suffer if W or H are optimized

to a local minimum far removed from any global solution (Albright et al. 2006),(Razaei et al. 2011). Several papers have since considered effective factor initializations, some of which are rooted in the theory of SVD (Boutsidis and Gallopoulos 2008), (Langville et al. 2006), (Albright et al. 2006), PCA (Zhao 2008), k-means clustering (Frederic 2008), (Wild 2003), and Gabor wavelets (Zheng et al. 2007), as discussed in (Ho 2008), (Razaei et al. 2011), (Zheng et al. 2007), and the references therein.

The initialization strategies for CMF is similar to NMF, however, differ with the addition of the phase parameter, which must now be initialized. In (Kameoka 2009), the W and H parameters are initialized to random positive real numbers, whereas the phase parameter is initialized to the phase of the mixture. Training is introduced in (Le Roux et al. 2009) where the spectral templates matrix, W, was set to the results of an NMF based factorization on unmixed samples of the sources distinct from those used to form the mixture to be analyzed. The temporal activations matrix, H, was initialized to either random positive real numbers or the results of an NMF on the training data, whereas the phase parameter is initialized to the phase of the mixture. As mentioned in Subsection 2.3.3, several initialization strategies are considered in (King and Atlas 2011) and (King 2012) in which the W matrix is initialized using one of the three following approaches: 1) "no-train" using random positive real numbers, 2) "factorize-to-train" using the result of an NMF on the training data and 3) "copy-to-train" using the entire magnitude spectrogram of the training sources. The H matrix is also initialized using either 1) "no-train" using random positive real numbers or 2)"factorize-to-train" using the result of an NMF on the training data. In (King and Atlas 2011) and (King 2012), the phase parameters are set to the phase of the mixture, as in (Le Roux et al. 2009).

Due to the simplicity of the dataset considered for the purposes of this thesis, and the desire to keep any possible confounding factors to a minimum (such as the influence of a trained initialization on the factorization), positive random numbers, drawn from a uniform distribution, were considered for the initialization of W and H for both the NMF and CMF algorithms and the phase was initially set to be the phase of the mixture for the CMF algorithm for the experiments conducted throughout Chapters 4 - 6.

3.3.5 Stopping Criteria

Another important aspect of the NMF/CMF algorithms that must be addressed is the definition of appropriate stopping criteria for the alternating update scheme. Typically, the NMF updates will continue until a suitable approximation has been obtained, as specified by some established threshold on the the cost function (i.e., $D(X||WH) \leq \epsilon$). Alternatively, the NMF updates can be stopped once the difference in the value of the cost function between subsequent approximations is below a given threshold, a maximum number of iterations have been performed, or a maximum allotted time for the updates to occur has expired (Cichocki et al. 2009). It is also possible to have stopping criteria which is based on a variation of any or all of these conditions, as examined in (King 2012) in which both a maximum number of iterations and a pre-specified approximation threshold are used, which is actualized in the code provided in (King and Atlas 2012). Throughout the experiments conducted in Chapters 4 - 6, the factorization was terminated after either 100 iterations had been performed or the value the cost function, $C(\theta)$, at the current iteration was less than 120 dB relative to the cost function at the preceding iteration.

3.3.6 Component-to-Source Clustering Strategies

As discussed in Subsection 2.2.4, the NMF algorithm extracts the spectral profiles of the generative components (e.g., sustained tones, transients, and noise) and several components are often required to accurately describe each source. In other words, the number of extracted components, K, is usually much larger than the number of sources, P. NMF in its classic form, however, does not inherently structure the columns of W in any particular order and may in fact be subject to several permutations of the columns/rows of the matrix factors, given a random initialization. Thus, clustering of the extracted NMF components into sources is an important processing step which has a significant impact on the quality of the separation. Component grouping strategies can range from manual clustering based on direct observation of the components by the user, as in (Wang and Plumbley 2005), to more elaborate approaches, such as those based on harmonic/temporal cues as in (Nakano, Kitano, Ono, and Sagayama 2010) or the inclusion of additional, statistically motivated, group-sparsity cost functions applied to the activations matrix, H, which is minimized when the components are appropriately grouped into sources, as in (Lefevre et al. 2011). Various other grouping strategies have been addressed in (Virtanen 2004), (Virtanen 2006),

(Jaiswal et al. 2012), (Casey and Westner 2000), and the references therein.

Two simplifications were made regarding the number of extracted components/component-to-source clustering scheme employed in the experiments of Chapters 4 - 6. The first simplification made was to only extracted $K = P = 2$ components, one component for each source. In doing so, CMF is reduced to a form more akin to CMFWISA, for which only one phase matrix is associated with a given source. This simplification eases the interpretability of both the extracted components and the learnt phase matrix but comes at the cost of losing the ability to fully describe the time varying characteristics of each source. It was decided that the gain in interpretability of the estimated components and phase matrix outweighed the increase in expressivity, which would be gained if using more than one component per source. The second simplification made was to adopt a non-blind component-to-source clustering procedure, motivated by the work of (Virtanen 2006), which uses the unmixed sources as reference. Even though K was set to be equal to P, it is still necessary to determine which source is associated with which components in order to calculate the BSS performance measures, defined below in Subsection 3.3.7, and to evaluate whether any source remained undetected, as discussed in Subsection 3.3.9. The spectral source to distortion ratio (SSDR) was calculated between the magnitude spectrogram of the p^{th} unmixed source, S_p, and the k^{th} extracted component, $\widehat{C}_k$, as follows:

Spectral Source to Distortion Ratio:

$$\text{SSDR} := 10\log_{10}\left(\frac{\sum_{n,m}|[S_p]_{n,m}|^2}{\sum_{n,m}|[S_p]_{n,m} - |[\widehat{C}_k]_{n,m}||^2}\right) \tag{3.8}$$

As in (Virtanen 2006), the k^{th} component was associated with the p^{th} source that leads to the highest SSDR.

3.3.7 Spectrogram Phase Recovery

As discussed in Subsection 2.2.4, in order to convert the extracted source estimates back into a time-domain representation using the NMF-based separation procedure, the phase information that was initially discarded when we considered the spectrogram of the mixed signal must be reincorporated into the extracted time-frequency representations of the

sources. In the case of the filtering-based source estimation method, the phase of each estimated source $STFT$ is typically taken to be the phase of the original mixture $STFT$. In the case of the synthesis-based source estimation method, the phase of each estimated source is also either taken to be the phase of the original mixture $STFT$, or taken to be the result of phase estimates based on methods such as those proposed in (Griffin and Lim 1984), (Slaney et al. 1994), and (Le Roux 2009). For the experiments conducted in Chapters 4 - 6, however, the phase of each estimated source $STFT$ using the NMF-based separation procedure is simply taken to be the phase of the original mixture $STFT$.

Using the CMF-based separation procedure, the phase parameter is now estimated. As such, in the case of the synthesis-based source estimation method, the phase of each estimated source is taken to be the phase estimated through the CMF update algorithm. Note however that in the case of the filter-based source estimation method, the phase of each estimate is once again simply taken to be the phase of the original mixture STFT.

Source Separation Performance Measures

In order to quantitatively assess the quality of the separation procedure, appropriate measures of distortion between the extracted source and the true source must be established. In (Vincent et al. 2006), performance measures for BSS are presented, which make use of the known pre-mixed sources but are independent of the mixing and demixing system. First, the source estimate, $\widehat{\mathbf{s}}_p$ is expressed as a sum of the target source, s_{target}, plus three different flavours of error, as follows:

$$\widehat{\mathbf{s}}_p = \mathbf{s}_{\text{target}} + \mathbf{e}_{\text{interf}} + \mathbf{e}_{\text{noise}} + \mathbf{e}_{\text{artif}} \tag{3.9}$$

As defined in (Vincent et al. 2006), $\mathbf{s}_{\text{target}} = \mathcal{D}(\mathbf{s}_p)$ is a version of $\mathbf{s}_p$ modified by an allowed distortion, $\mathcal{D}$. The interference error term, $\mathbf{e}_{\text{interf}}$, represents the part of estimated source perceived as coming from the unwanted sources. The noise error term, $\mathbf{e}_{\text{noise}}$, represents the part of the estimated source perceived as coming from sensor noise. Finally, the artifact error term, $\mathbf{e}_{\text{artif}}$, represents the part of the estimated source perceived as coming from other causes, such as forbidden distortions and/or "burbling" artifacts, as described in (Vincent et al. 2006). Next, in order to determine the relative value of each of these error terms, energy ratios are computed over the signal as follows:

Source to Distortion Ratio:

$$\text{SDR} := 10\log_{10}\frac{\|\mathbf{s}_{\text{target}}\|^2}{\|\mathbf{e}_{\text{interf}}+\mathbf{e}_{\text{noise}}+\mathbf{e}_{\text{artif}}\|^2} \tag{3.10}$$

Source to Interference Ratio:

$$\text{SIR} := 10\log_{10}\frac{\|\mathbf{s}_{\text{target}}\|^2}{\|\mathbf{e}_{\text{interf}}\|^2} \tag{3.11}$$

Source to Noise Ratio:

$$\text{SNR} := 10\log_{10}\frac{\|\mathbf{s}_{\text{target}}+\mathbf{e}_{\text{interf}}\|^2}{\|\mathbf{e}_{\text{noise}}\|^2} \tag{3.12}$$

Source to Artifact Ratio:

$$\text{SAR} := 10\log_{10}\frac{\|\mathbf{s}_{\text{target}}+\mathbf{e}_{\text{interf}}+\mathbf{e}_{\text{noise}}\|^2}{\|\mathbf{e}_{\text{artif}}\|^2} \tag{3.13}$$

For the purposes of this thesis, the mixtures considered are assumed to be noiseless. As such, only the SDR, SIR, and SAR performance measures are used throughout the experimentation of Chapters 4 - 6. As in (King 2012), the baseline of the SDR and SIR performance measures for the unmixed source estimates shall be taken to be the SDR and SIR performance measures of the unprocessed mixture. On the other hand, the baseline of the SAR performance measure shall be taken to be 0 dB. The reason for defining the baselines in such a way rests in the fact that the mixture in question will typically have a well-defined finite SDR and SIR before any source separation procedures have been applied, whereas the SAR of the mixture is infinite due to the absence of artifacts in the unprocessed mixture. As such, for the experiments described in Chapters 4 - 6, the SDR and SIR performance measure results of the unmixed source estimates will actually refer to the relative difference in SDR and SIR between the unmixed source estimates and the original mixture. The SAR performance measure results of the unmixed source estimates will simply refer to the SAR performance measure obtained for the unmixed source estimates (i.e., not relative to the original mixture). The calculation of the SDR, SIR and SAR performance measures used in the experiments of chapters: 4 - 6 were all performed using the MATLAB® BSS_EVAL toolbox (Févotte et al. 2005).

3.3.8 BSS Evaluation Region

As described above in Subsection 3.3.7, the BSS performance measures are determined based on the projection of the estimated sources on the original unmixed sources. If the BSS metrics are calculated over the entire signal, however, any non-zero projection of a given source onto an interfering source implies that the interfering source vector is present over the entire three seconds of the estimated source. This is not physically reasonable given that the notes only interfere during the last second of the sample, as depicted in figure 3.1. As such, a local BSS measure is employed, where only the last second of the mixture, the region of interest, is compared against the last second of the unmixed sources.

3.3.9 Undetected Sources

Using the component-to-source clustering scheme described above in Subsection 3.3.6, it is possible that the extracted components become grouped with a proper subset of the original unmixed sources, leaving the other source(s) undetected. When this situation occurs, the projections of the undetected source on the original unmixed sources will all be zero. Thus, the SDR, SIR, and SAR BSS performance measures will adopt an undefined numerical result. This is indicated by the NaN placeholder in Matlab®. Instead of removing the test cases in which undetected notes occur, the template and activation vector for the undetected note is set to random positive numbers. The BSS performance measures were calculated according to these new random templates and activations.

Chapter 4

Experiment 1: NMF vs. CMF Parameter Analysis

4.1 Motivation

The NMF-based and CMF-based source separation procedures considered in this thesis can both be roughly divided into three distinct processing steps: 1) $STFT$-based analysis of the musical mixture, 2) NMF-based/CMF-based spectrogram/$STFT$ factorization, and 3) $ISTFT$-based synthesis of the unmixed source estimates. All three of these processing steps have several problem-specific parameters which must be specified by the user. The $STFT$ analysis and $ISTFT$ synthesis of the audio mixture depend on the length of the Discrete Fourier Transform (DFT) analysis and Inverse Discrete Fourier Transform ($IDFT$) synthesis at each time frame, as well as windowing considerations, such as window length, shape, and hop size. Similarly, the NMF and CMF algorithms depend on several problem-specific parameters associated with the factorization cost functions and imposed sparseness/consistency penalty criterion. As such, the first experiment was conducted in order to gain insight into which analysis/factorization/synthesis parameter value combinations could potentially yield high/low NMF-based and CMF-based separation performances, as quantified using the SDR, SIR, and SAR performance measures.

4.2 Experimental Design

Determining appropriate factorization and $STFT/ISTFT$ parameter values, given a particular dataset, is a primary focus of the experiments presented in (Smaragdis 2007), (Smaragdis et al. 2009), (O'Grady and Pearlmutter 2007), (King et al. 2012), (Shashanka 2003), and (Schmidt and Olsson 2006). In all of these papers, appropriate factorization and/or $STFT/ISTFT$ parameters are determined following a trial-and-error approach: a range of values for each parameter is tested and the values which tend to yield the highest separation performance, given some suitable measure, is selected. Following the designs of these experiments, the parameter values in question were varied according to the levels outlined in table 4.1.

Table 4.1: STFT/FACTORIZATION/ISTFT PARAMETERS

	Notation	Value(s)
Factorization Parameters		
Number of Sources	P	{2}
Number of Components	K	{2}
Sparsity Weight	λ	{0 0.001 0.01 0.1 1}
Sparsity Power	g	{1}
Consistency Weight*	γ	{0 0.001 0.01 0.1 1}
STFT/ISTFT Parameters		
Window Length (samples)	L	{512 1024 2048 4096}
DFT Size	$\tilde{N}$	$\{\tilde{N} = L\}$
Hop Size	$\tilde{H}$	$\{L/4\}$
Window Type	w	{Modified Hann}

***note:** only applies to CMF

Using Matlab® R2012a, 8 audio mixtures were created using the piano (PN) and guitar (GTR) samples from the dataset described in Subsection 3.3.2, one for each piano note, ranging in pitch from C4 - C5. NMF-based and CMF-based separation procedures were performed using all possible combination of the parameters set according to the levels specified in Table 4.1. This resulted in (8 note pairings $\times$ 5 sparsity levels $\times$ 4 window lengths) = 160 tests performed using the NMF algorithm and (8 note pairings $\times$ 5 sparsity

levels × 5 consistency levels × 4 window lengths) = 800 tests performed using the CMF algorithm. Each test was conducted 10 times, randomizing the initializations. In total, this amounted to (160 NMF tests + 800 CMF tests) × 10 = 9600 tests. The NMF algorithm was able to perform roughly 32 of these separations per minute on average, whereas the CMF algorithm was only able to perform roughly 2 of these separations per minute on average. In total, the first experiment took nearly 3 full days of running[1] to complete.

4.3 Results/Analysis

Table 4.2: NMF MEDIAN AND INTER-QUARTILE RANGE RESULTS FOR PARAMETER VALUE COMBINATION YIELDING MAX/MIN PERFORMANCE

	NMF		
	Maximum Median [Q1, Q3]		
NOTE RANGE	[SDR]	[SIR]	[SAR]
	19.27[8.56, 21.48] dB	**29.61[15.60, 33.08] dB**	**19.89[12.33, 21.99] dB**
PN: C4-C5	K: 2	K: 2	K: 2
	λ: 0.001	λ : 0.001	λ :0.001
GTR: C4	γ -:	γ: -	γ: -
	N: 4096	N: 4096	N: 4096
	SM: FILT	SM: SYNTH	SM: FILT
	Minimum Median [Q1, Q3]		
NOTE RANGE	[SDR]	[SIR]	[SAR]
	-4.89[-6.66, -3.75] dB	**0.71[-0.18, 1.56] dB**	**-1.77[-5.92, 0.14] dB**
PN: C4-C5	K: 2	K: 2	K: 2
	λ: 1	λ : 1	λ :1
GTR: C4	γ: -	γ: -	γ: -
	N: 512	N: 512	N: 512
	SM: SYNTH	SM: FILT	SM: SYNTH

[1]Using a desktop computer with 2 × 2.66 GHz Dual-Core Intel Xeon processors.

Table 4.3: CMF MEDIAN AND INTER-QUARTILE RANGE RESULTS FOR PARAMETER VALUE COMBINATION YIELDING MAX/MIN PERFORMANCE

	CMF		
	MAXIMUM MEDIAN [Q1, Q3]		
NOTE RANGE	[SDR]	[SIR]	[SAR]
	16.25[7.23, 17.52] dB	**25.98[13.40, 29.17] dB**	**17.35[11.94, 18.36] dB**
PN: C4-C5	K: 2	K: 2	K: 2
	λ: 0.01	λ : 0.01	λ :0.01
GTR: C4	γ: 0	γ: 0	γ: 0
	N: 1024	N: 1024	N: 1024
	SM: FILT	SM: SYNTH	SM: FILT
	MINIMUM MEDIAN [Q1, Q3]		
NOTE RANGE	[SDR]	[SIR]	[SAR]
	-8.37[-17.59, -6.92] dB	**0.02[-20.22, 0.22] dB**	**-5.15[-17.82, -3.69] dB**
PN: C4-C5	K: 2	K: 2	K: 2
	λ: 1	λ :1	λ :1
GTR: C4	γ: 1	γ: 1	γ: 1
	N: 512	N: 512	N: 512
	SM: SYNTH	SM: FILT	SM: SYNTH

The suppression of the unwanted source and the reduction of unwanted artifacts were both deemed equally important for the purpose of this first experiment. As such, the SDR was chosen as a global performance measure, as suggested in (Vincent et al. 2006). For each parameter value combination, data vectors of length $2 \times 8 \times 10 = 160$ were created by pooling the results of the SDR performance measures obtained for both sources, for all 8 note pairings, and for all 10 trials. For any given parameter value combination, two such data vectors were created, one for the results obtained using the synthesis-based estimation method and one for the results obtained using the filter-based estimation method. The parameter value combination that resulted in the pooled performance measures with the maximum/minimum median were considered as yielding the maximal/minimal separation performance. A parameter value combinations which resulted in an undetected source was not considered as possible candidates for providing the best separation performance.

The median SDR, SIR, and SAR performance measures, given these maximal/minimal parameter value combinations, are presented in Table 4.2 and Table 4.3. To provide some insight into the spread of the data, the first and third quartiles for the results are presented in square brackets along with the median.

NMF vs. CMF Performance

From the data presented in Table 4.2 and Table 4.3, it can be seen that median SDR, SIR, and SAR results obtained using the NMF-based separation procedure under the parameter value combination resulting in the best NMF-based separation performance were larger than the median SDR, SIR, and SAR results obtained using the CMF-based separation procedure under the parameter value combination resulting in the best CMF-based separation performance. In fact, an improvement in maximal median SDR, SIR, and SAR, of 3.02 dB, 3.63 dB, 2.54 dB, respectively, were obtained using the NMF-based separation procedure over the CMF-based separation procedure. The spread of the data, as indicated by the first and third quartiles, is also suggestive of a better performance by the NMF-based separation procedure over the CMF-based separation procedure.

At first glance it seems that these results contradict the more promising results obtained using the CMF-based separation procedure in (Le Roux et al. 2009) and (King 2012). However, several key distinctions must be made which may shed light on this apparent discrepancy. Perhaps the largest distinction is the fact that in the work of (Le Roux et al. 2009) and (King 2012) the spectral template matrix, W, was initialized with $K >> P$ templates obtained from the separation performed on unmixed training data. In the work presented here, however, W was initialized using positive random numbers and only $K = P$ templates were being extracted. It is hypothesized that since only $K = P$ templates were being extracted, the factorizations obtained using NMF/CMF best approximated the original mixture spectrogram/$STFT$ if the extracted components reduce the error in describing the highest energy, which is found within the first few harmonics of each source. As such, the higher harmonics which possess lower energy relative to the lower harmonics may not be accurately described over time. As discussed in Subsection 2.3.3, it was observed (King and Atlas 2011) that the higher source frequencies are better modelled using CMF compared to NMF due to the inclusion of the phase parameter, resulting in a greater reduction of high frequency artifacts (higher SAR performance measure). This

result however was obtained using a value of $K >> P$, for which higher frequencies could be better described over time. Therefore, perhaps the fact that the experiment conducted here only uses $K = P = 2$ extracted templates limits the true potential of CMF as it was observed to perform in (Le Roux et al. 2009) and (King 2012). The results obtained in this first experiment and the results obtained in (King 2012) could suggest that in order to gain the full benefits of the CMF-based separation procedure over the NMF-based separation procedure in terms of better describing the higher frequencies resulting in a higher SAR performance measure, the number of extracted components should be $K >> P$.

Another possible source for the discrepancy found between the results obtained in this first experiment and those obtained in (Le Roux et al. 2009) and (King 2012) could be in the mixture signals considered. Unlike in the experiments conducted in (Le Roux et al. 2009) and (King 2012), which involved mixtures of speech signals, the sources considered for testing in the above experiment are highly structured harmonic musical signals. As such, it is hypothesized that high energy spectral overlap within the musical test signals is not as prevalent as it was in the mixtures involving less structured speech signals. Perhaps the incorrect NMF-mixture model assumption (which will always be present) is not affecting the high energy, low frequency, spectral regions enough to noticeably hinder the performance of the NMF-based separation procedure. This idea is the basis of the experiments conducted in Chapters 5 - 6.

Audible Differences in Performance

In general, the reconstructed source signal estimates using both the NMF-based and CMF-based source separation procedure, under the parameter value combinations which resulted in the maximal performance, sounded quite similar. Audio examples of the estimated sources can be found online at `www.music.mcgill.ca/~jbrons1`. Some noticeable high frequency artifacts are present in the results obtained from both separation procedures using the synthesis-based source estimation method. A clear timbral degradation is also observed in the source estimates obtained using the synthesis-based source estimation method due to the fact that only $K = P$ sources were being extracted. Overall, the filter-based source estimation method produced superior sounding separation results with little degradation to the timber of each source. An audible difference in note onset is present between the NMF-based and CMF-based separation results. The NMF-based separation procedure, in

general, produced noticeably delayed note onsets, which is almost certainly a result of the higher window length employed resulting in poorer temporal resolution, as compared to the results obtained using the CMF-based separation procedure.

Sparsity Weight Considerations

The minimum median SDR, SIR, and SAR performance measures obtained, for both the NMF-based and CMF-based separation procedures, occurred using a sparsity weight of $\lambda = 1$. This result is supported in the work of (O'Grady and Pearlmutter 2007), in which poorer performance measures were also observed for higher sparsity weights. As discussed in (King 2012), higher values of the sparsity weight could result in the values of the elements in the activation weight matrix, H, being driven to zero, compromising the quality of the factorization and hence separation. The maximum median SDR, SIR and SAR performance results for the NMF-based separation procedure were achieved with a lower sparsity weight of $\lambda = 0.001$. It was also concluded in (Virtanen 2006) that a lower sparsity weight of $\lambda = 0$ yielded the highest SDR performance for their dataset, albeit a slightly different formulation of the sparsity constraint from the one considered in this thesis was used in the work of (Virtanen 2006). The maximum median SDR, SIR, and SAR results for the CMF-based separation procedure were achieved with a sparsity weight of $\lambda = 0.01$. In (King 2012), a value of $\lambda = 0.01$ was also observed to work well.

Window Length Considerations

As discussed in Subsection 3.3.3, the window lengths of $L \in \{512, 1024, 2048, 4096\}$ samples correspond to window lengths of $L \in \{46, 93, 186, 372\}$ ms, using a sampling rate of $Fs = 11025$ Hz. As such, the NMF-based separation procedure obtained maximum median SDR, SIR, and SAR performance measures using a window length of $L = 372$ ms. The minimum median performance measures were obtained using a window length of $L = 46$ ms for both the NMF-based and the CMF-based separation procedures. The maximum median SDR, SIR, and SAR performance measures were obtained using a window length of $L = 93$ ms for the CMF-based separation procedure. As discussed in Subsection 3.3.3, the size of the analysis window, L, dictates the tradeoff between spectral resolution and temporal resolution. Larger window lengths offer higher spectral resolution but lower temporal resolution, whereas smaller window lengths offer higher temporal resolution but lower

spectral resolution. As such, the fact that the lowest performance measures for both the NMF-based and CMF-based separation procedures were obtained using a window length of $L = 46$ ms suggests that lower spectral resolution/higher temporal resolution yielded poorer performances for both the NMF-based and CMF-based separation procedures for this particular dataset. The best NMF results were obtained using a window length of $L = 372$ ms, which suggests that higher spectral resolution/lower temporal resolution yielded a superior NMF-based separation performance. It is hypothesized that, despite the decrease in temporal resolution, larger window lengths promote a reduction in the noticeable effects of the NMF mixture model violation by increasing the frequency resolution, thus minimizing strong energy spectral overlap between overlapping partials. This hypothesis is supported in the calculations performed in Subsection 5.2.1.

It is speculated in Subsection 3.3.3 that poor temporal resolution may contribute to poor separation results by allowing any error in approximation to be spread over greater temporal regions. In fact, as hypothesized in Subsection 5.4 the reduction in temporal resolution is contributing to a noticeable delay in note onset in the final source estimate reconstructions. However, based on the results of this first experiment it seems that this lack of temporal resolution did not outweigh the increase in performance obtained from the higher spectral resolution resulting from the use of a larger analysis window. This being said, it is hypothesized that the full effect of the poor temporal resolution associated with larger windows is not being observed in the SDR, SIR, and SAR performance measures due to the fact that the performance measures were only evaluated on the last second of the mixture, as explained in Subsection 3.3.8. As such, the effect of the poor temporal resolution during the high energy note onset occurs before the start of the performance evaluation region and does not contribute to the calculation of the performance measures. However, due to the relatively short duration of the errors obtained as a result of having poor temporal resolution, as compared to the duration of the errors obtained as a result of having poor spectral resolution, which may last the entire last second of the mixture, it is suspected that the lack of temporal resolution truly did not outweigh the increase in performance obtained from the higher spectral resolution resulting from the use of a larger analysis window.

Lastly, it is interesting to note that the window length resulting in the maximal CMF-based separation performance was $L = 96$ ms. This window length has a temporal resolution 4 times greater than that obtained using a window length of $L = 384$ ms. This result

is promising in the sense that it could suggest that the CMF-based separation procedure performs well for lower length windows, compared to the NMF-based separation procedure, allowing for the preservation of the temporal resolution. This idea is explored further in the experiments conducted in Chapters 5 - 6.

Consistency Weight Considerations

The maximum median SDR, SIR, and SAR results, using the CMF-base separation procedure, were obtained when the consistency weight was set to $\gamma = 0$. The minimum median SDR, SIR, and SAR results were obtained when using a consistancy weight of $\gamma = 1$. It was similarly observed in the work of (King 2012) that the inclusion of the consistency weight degraded performance when incorporated into both the phase parameter updates and the activation weight parameter updates. It was discussed in (King 2012) that higher values of the consistency weight resulted in the elements in the activation weight matrix being driven down to zero, resulting in a zero matrix, in a similar fashion to the effect of using a high sparsity weight. Promising results were obtained in (King 2012) when the consistency penalty was only included in the update of the phase matrix. It was demonstrated in (Le Roux et al. 2009) that the inclusion of the consistency constraint seemed to enable improvements in the results obtained using NMF-based and CMF-based separation procedures without the consistency constraint. However, it is important to note that the spectral templates used in (Le Roux et al. 2009) were trained prior to the factorization of the spectrogram/$STFT$ of the mixture signal, whereas the spectral templates in the study presented here were initialized with positive random numbers and estimated using update algorithm, described in Subsection 3.2, which contain parameters reflecting the inclusion of the consistency constraint. Future experiments should investigate the performance of the CMF-based separation procedure applied to a broader range of musical sources when this newly reformulated consistency constraint is only applied to the phase update equations.

Source Estimation Method Considerations

From Table 4.2 and Table 4.3, the highest resulting median SDR and SAR values, for both the NMF-based and the CMF-based separation procedures, occurred using the filter-based source estimation method and the lowest resulting median SDR and SAR values, for both the NMF-based and the CMF-based separation procedures, occurred using the

synthesis-based source estimation method. On the other hand, the highest resulting median SIR values were a result of the synthesis-based source estimation method and the lowest resulting median SIR values were a result of the filter-based source estimation method, for both the NMF-based and the CMF-based separation procedures. These results are supported by the fact that, as discussed in (King 2012), the filter-based source estimation method is typically better at suppressing the unwanted sources (higher SIR) at the cost of poor suppression of undesired artifacts (lower SAR). Similarly, the filter-based source estimation method typically has a higher suppression of unwanted artifacts (higher SAR) but a lower suppression of unwanted sources (lower SIR).

4.4 Conclusion

The experiment presented in this chapter investigated the parameter value combinations resulting in the best/worst separation performances obtained using the NMF-based and CMF-based separation procedures. Both the suppression of the unwanted source and the reduction of unwanted artifacts were deemed equally important for the purposes of this first experiment. As such, the quality of the separation performance was determined by taking the median of the results of the SDR performance measures for both sources, over all 8 note pairings, and over all 10 trials. The results of this experiment indicate that, under the maximal parameter value combinations, the median SDR, SIR, and SAR results obtained using the NMF-based separation procedure were all higher than the median SDR, SIR, and SAR results obtained using the CMF-based separation procedure. The parameter value combinations that yielded the highest separation performance are as follows: ($L_{NMF} = 4096$ and $L_{CMF} = 1024$), ($\lambda_{NMF} = 0.001$ and $\lambda_{CMF} = 0.01$), and ($\gamma_{CMF} = 0$).

Investigations into the behaviour of the reformulated constancy constraint led to the conclusion that the best performance results for the CMF algorithm were obtained when the consistency constraint weight was set to $\gamma = 0$. It should be stressed, however, that the experiments conducted here used very basic and highly controlled musical samples in order to ease the interpretation of the factorization results. As suggested in Subsection 4.3, future work should investigate the performance of the CMF-based separation procedure under the consistency constraint when applied to a larger database of musical sources. Future experimentation should also consider the performance behaviour of the CMF-based separation procedure when this newly reformulated consistency constraint is only applied

to the phase update equations.

As discussed in Subsection 3.3.6, a simplification was made to only extract $K = P$ components, which reduces CMF to a form resembling CMFWISA. This simplification was made because the inclusion of the phase parameter makes it very difficult to interpret the behaviour of the CMF algorithm, especially if more than one component is extracted per source, which would involve several magnitude and phase matrices interacting with one another to produce the final source estimate. In these early stages of testing is was deemed more beneficial to examine the behaviour of CMF when only one phase component was estimated per source. Based on the results of this first experiment, however, it was concluded that future experiments should investigate the performance of the CMF-based separation procedure for values of $K >> P$ in order to observe the possible benefits of CMF over NMF in describing the higher frequency content of the estimated sources, as observed in (King and Atlas 2011).

It was also observed in the results of this first experiment that when using only $K = P$ extracted templates, the benefits of the CMF-based separation procedure over the NMF-based separation procedure may be observed when using smaller analysis window lengths. It is hypothesized that smaller window lengths promote more noticeable effects of the NMF mixture model assumption violation due to the decrease in spectral resolution, which may promote high energy spectral overlap in the lower frequency regions. These low frequency regions are of particular importance because they posses most of the spectral energy, which must be described in order to have a strong approximation to the mixture spectrogram/$STFT$. The effects of "increasing" the NMF mixture model assumption violation on the NMF-based and CMF-based separation procedures was examined in the experiments of Chapters 5 - 6.

Chapter 5

Experiment 2: NMF vs. CMF with Estimate/Oracle Phase

5.1 Motivation

Based on the results of the first experiment, it was questioned whether the CMF-based separation procedure could offer potential benefits, over the NMF-based separation procedure, in conditions for which the effects of the incorrect NMF mixture model assumption are greatest, given the STFT analysis parameters considered. The experiment detailed in this chapter examines which of the analysis window lengths, in the range of $L \in \{46, 93, 186, 372\}$ ms, offers the highest violation of the NMF mixture model assumptions. This experiment also investigates the performance of the CMF-based separation procedure in conditions for which the exact phase of the unmixed sources are known and incorporated into the factorization algorithm (oracle phase conditions). By analyzing the performance of the CMF-based separation procedure under oracle phase conditions, insight could be gained concerning a possible upper bound of achievable separation performance using the CMF-based separation procedure for this particular dataset/parameter value combinations, in conditions for which the NMF mixture model assumption is most violated given the STFT analysis parameters considered. Henceforth, CMF with the inclusion of the oracle phase conditions shall be denoted CMF(OP), where OP stands for "Oracle Phase". Similarly, standard CMF, for which the phase matrices are estimated, will be referred to as CMF(EP), where the EP stands for "Estimated Phase".

5.2 Pre-Experimental Design Considerations

5.2.1 Violation of the NMF Mixture Model Assumptions

It is important to note that the incorrect NMF mixture model assumption will always be present for any spectrogram factorization, as defined in the inequality of Equation (2.35). However, the amount of error present in this incorrect mixture model assumption can vary depending on the amount of high energy spectral overlap present within the mixture spectrogram, which is ultimately dictated by the length/shape of the $STFT$ analysis widow employed. The strength of the NMF mixture model assumption violation, (i.e., the degree to which the sum of the spectrograms of the sources differed from the spectrogram of the mixture) was quantified for each window length in the set $L \in \{46, 93, 186, 372\}$ ms, by calculating the SSDR between the spectrogram of the mixture and the sum of the spectrogram of the sources, as follows:

$$\text{SSDR} = 10\log_{10}\left(\frac{\sum\limits_{n,m}|[X]_{n,m}|^2}{\sum\limits_{n,m}|[X]_{n,m} - [\sum\limits_{p=1}^{P} S_p]_{n,m}|^2}\right) \tag{5.1}$$

As the denominator of the ratio approaches zero, the error in the NMF mixture model assumption becomes smaller and the SSDR increases (approaching infinity in the limiting case for which the NMF assumptions hold true). On the other hand, as the denominator of the ratio increases, the error in the NMF mixture model assumption becomes larger and the SSDR decreases. The median SSDR over all 8 note pairings is calculated for each window length and the results are presented below in Table 5.1.

Table 5.1: SSDR QUANTIFYING NMF MIXTURE MODEL VIOLATION FOR EACH WINDOW LENGTH AVERAGED OVER ALL NOTE PAIRINGS

WINDOW LENGTH (L)	SSDR
46 ms	17.64 dB
93 ms	22.44 dB
186 ms	23.75 dB
372 ms	23.81 dB

The SSDR calculations suggest that, given this data set, the largest violation of the NMF mixture model assumption, indicated by the lowest median SSDR, occurs using a window length of $L = 46$ ms. The results presented in Table 5.1 support the hypothesis presented in Subsection 4.3 that larger window lengths promote a decrease in the violation of the NMF mixture model assumption by increasing the frequency resolution, hence decreasing strong energy spectral interference between overlapping sources. Though this idea may seem obvious, it is also important to consider that the decrease in temporal resolution resulting from the use of a larger analysis window could have also contributed to a strong violation of the NMF mixture model assumption, especially during the high energy, inharmonic, note onsets.

The difference between the spectrogram of the mixture and the sum of the spectrogram of the sources is displayed in Figure 5.1, for each note pairing, given a window length of $L = 46$ ms. The magnitude of the difference is represented by the colour of the image (red corresponding to a large magnitude difference and blue corresponding to a small magnitude difference). Other than the transition between the two sources, which occurs at $t = 1$ s, the NMF mixture model assumptions are not violated for the first two seconds of the mixture. However, it can be seen that, as one would expect, the greatest violation of the NMF mixture model occurs during the last second of the mixture, in which the piano source and guitar source are mixed together.

5.2.2 Window Length Considerations Revisited

As mentioned in Section 2.3, the pitch intervals commonly found in Western music are often very close to simple integer ratios ($\approx \frac{5}{4}, \frac{4}{3}, \frac{3}{2}, \frac{5}{3}$) (Woodruff et al. 2008). Thus, it is not uncommon for overlapping notes to share common harmonics and hence a spectral overlap potentially yielding a strong violation of the NMF mixture model assumption. For instance, in the dataset considered, the GTR:C4 note has harmonics in common with the PN:C4 (unison: $\frac{1}{1}$), PN:E4 (major third: $\frac{5}{4}$), PN:F4 (fourth: $\frac{4}{3}$), PN:G4 (fifth: $\frac{3}{2}$), PN:A4 (sixth: $\frac{5}{3}$), and PN:C5 (octave: $\frac{2}{1}$) notes. However, because these shared harmonics are very close to being exactly equal, no length of analysis window will suffice in resolving these shared harmonics. That being said, when two harmonics are closely spaced but not equal, it is possible to resolve each harmonic by adjusting the length of the analysis window employed. A conservative measure of spectral resolution, when trying to resolve two closely

spaced harmonics of overlapping sources, requires that no overlap occurs between the main lobe of the analysis window centred about the frequency of each harmonic. As will be further discussed in Subsection 6.3.4, the width of the main lobe of the modified Hann window spans 4 frequency bins[1]. Thus, the minimum difference in frequency between two harmonics, under this conservative spectral resolution requirement, is equal to $4\frac{Fs}{L}$. For a window length of $L = 512$ samples, this corresponds to a minimum difference in frequency of $4\frac{11025}{512} \approx 86$ Hz. The difference in frequency between the first harmonic of the PN:D4 source (≈ 294 Hz) and the first harmonic of the GTR:C4 source (≈ 261 Hz) fails to meet this requirement. As such, the (PN:D4 GTR:C4) mixture possesses a strong spectral overlap resulting in a large violation of the NMF mixture model assumption. Similarly, the difference in frequency between the first harmonic of the PN:B4 source (≈ 494 Hz) and the second harmonic of the GTR:C4 source (≈ 523 Hz) also fails to meet the conservative spectral resolution requirement. As such, the (PN:B4 GTR:C4) mixture also possesses a strong spectral overlap resulting in a large violation of the NMF mixture model assumption. Instances of these spectral overlaps producing strong violation of the NMF mixture model assumption can be observed for every note pairing in Figure 5.1. This suggests that using a window length of $L = 512$ samples presents good test cases for analyzing the performance of the NMF-based and CMF-based separation procedures in conditions for which the spectrogram of the mixture highly deviates from the sum of the spectrograms of the sources. Note that it would be possible to simply increase the length of the analysis window allowing for sufficient spectral resolution such that the strong spectral overlap no longer exists between the GTR:C4 and PN:D4/PN:B4 mixtures. However, one could argue that when dealing with sources lower in pitch the same problem would persist for larger window lengths. For instance, even if a window length of $L = 372$ ms is used, under the conservative spectral resolution requirements, frequency differences smaller than $4\frac{11025}{4096} \approx 11$ Hz in the harmonics of overlapping sources will also suffer from the same high energy overlap, potentially yielding a strong violation of the NMF mixture model assumption, with the additional cost of lower temporal resolution. These low frequency cases are not considered in the dataset chosen for this initial exploration, however, are well within the range of many musical instruments, including the acoustic guitar and piano.

[1]No zero padding is being considered for any of the experiments conducted throughout this thesis, as stated in Table 4.1

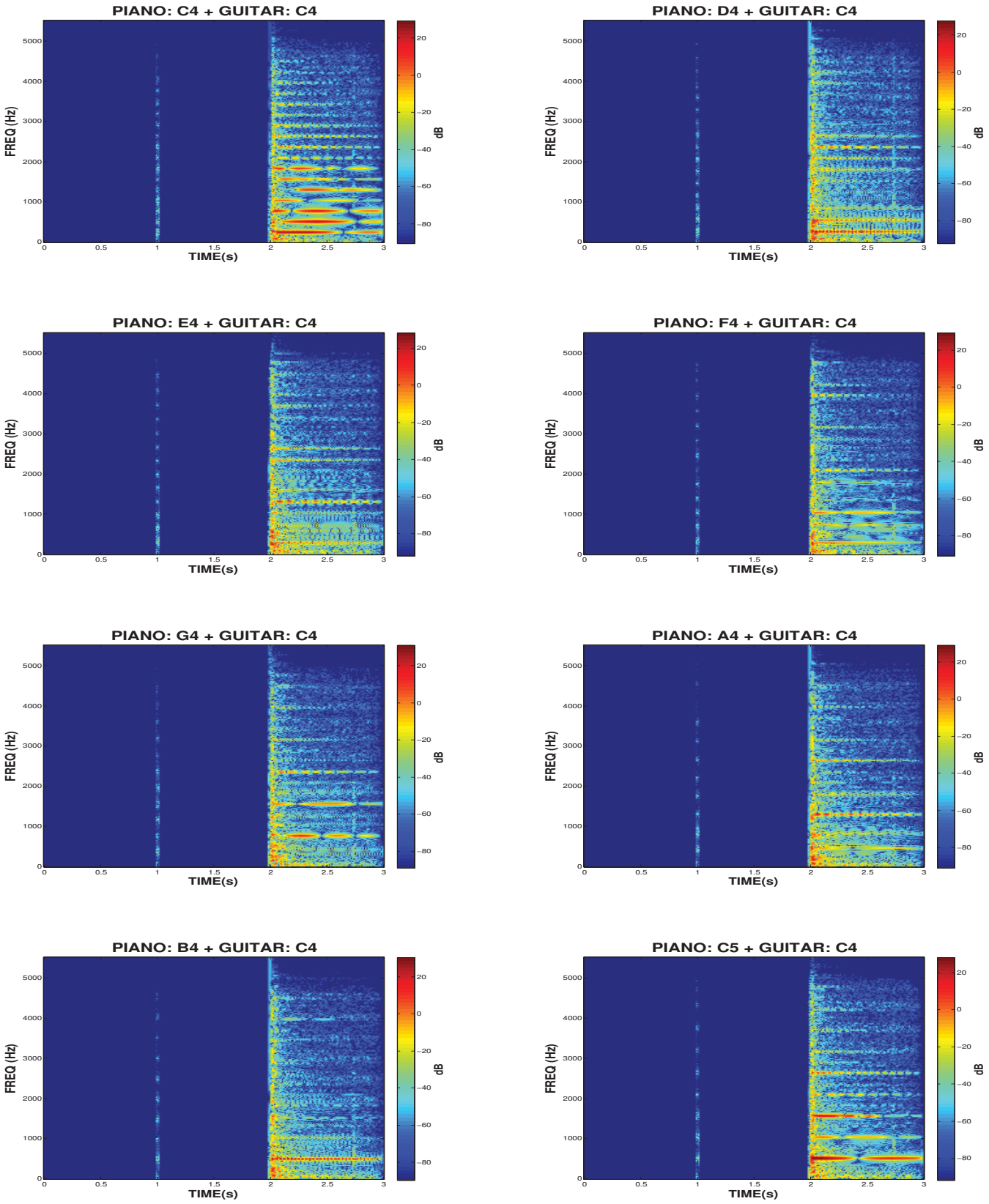

Fig. 5.1: REGION OF NMF MIXTURE MODEL ASSUMPTION VIOLATION FOR A WINDOW LENGTH OF L = 512 SAMPLES

5.2.3 Parameter Value Combination Selection

Along with investigating the behaviour of the NMF-based and CMF-based separation procedures in conditions for which a strong violation of the NMF mixture model assumption exists, the experiment described below also investigates the behaviour of the CMF-based separation procedure subject to refinements in the phase parameter update (estimated phase vs. oracle phase). As such, the synthesis-based source estimation method is of primary interest. Even though the refinements in the phase parameter could benefit the estimation of the templates and activation weights, which would in turn improve the source estimate using the filter-based source estimation method, only the synthesis-based source estimation method fully incorporates the refined phase parameter into the final source estimates. It was also desired to focus the attention of this experiment primarily on the SIR performance measure. Though the SAR performance measure is important and will still be considered throughout the following analysis, the SIR performance measure offers a better indication of how well the source separation procedure is able to isolated the individual sources from one another. Based on the data collected from the first experiment, the parameter value combinations which resulted in the maximum median SIR performance measures, given the synthesis-based source estimation method, $K = 2$ components, and a window length of $L = 512$ samples for both the NMF-based and CMF(EP)-based separation procedures, were determined. The median SDR, and SAR were also calculated using these parameter value combinations. The results are presented below in Table 5.2 and Table 5.3.

5.3 Experimental Design

Source separation tests were conducted, using the CMF(OP)-based separation procedure, for each of the 8 note pairings. Unlike in the case of the CMF(EP)-based separation procedure, the phase parameter was not updated, rather set to be equal to the true phase of the unmixed sources at every iteration. Based on the results of Table 5.3, the parameter value combination of $L = 512$, $K = 2$, $\lambda = 0.01$, and $\gamma = 0.001$ was used. In a similar fashion to the first experiment, each separation was conducted 10 times, randomizing the initializations. This resulted in 80 separations which took roughly 30 minutes to complete.

5.4 Results/Analysis

Performance Measure Analysis

Table 5.2: NMF MEDIAN RESULTS WITH MAX MEDIAN SIR PARAMETER VALUE COMBINATION

	NMF		
	MEDIAN [Q1, Q3]		
NOTE RANGE	[SDR]	[SIR]	[SAR]
PN: C4-C5 **GTR: C4**	**10.42[6.56, 13.10] dB** K: 2 λ: 0.001 γ -: N: 512 SM: SYNTH	**18.11[12.57, 28.21] dB** K: 2 λ : 0.001 γ: - N: 512 SM: SYNTH	**11.83[9.24, 13.22] dB** K: 2 λ :0.001 γ: - N: 512 SM: SYNTH

Table 5.3: CMF(EP) MEDIAN RESULTS WITH MAX MEDIAN SIR PARAMETER VALUE COMBINATION

	CMF(EP)		
	MEDIAN [Q1, Q3]		
NOTE RANGE	[SDR]	[SIR]	[SAR]
PN: C4-C5 **GTR: C4**	**9.93[3.22, 12.49] dB** K: 2 λ: 0.01 γ: 0.001 N: 512 SM: SYNTH	**18.29[8.38, 22.95] dB** K: 2 λ : 0.01 γ: 0.001 N: 512 SM: SYNTH	**11.15[8.30, 12.98] dB** K: 2 λ :0.01 γ: 0.001 N: 512 SM: SYNTH

From the results described in Table 5.2 and Table 5.3 it can be seen that a slightly higher median SIR measure was obtained using the CMF-based separation procedure over the NMF-based separation procedure, however the interquartile range of the SIR measure suggests that the spread of the NMF-based results is more favourable. Furthermore, no improvement in SDR and no improvement in SAR was obtained using the CMF(EP)-based separation procedure, compared to the NMF-based separation procedure. Despite

Table 5.4: CMF(OP) MEDIAN RESULTS WITH MAX MEDIAN SIR PARAMETER VALUE COMBINATION

	CMF(OP)		
	MEDIAN [Q1, Q3]		
NOTE RANGE	[SDR]	[SIR]	[SAR]
PN: C4-C5 **GTR: C4**	**13.67[12.60, 14.75] dB** K: 2 λ: 0.01 γ: 0.001 N: 512 SM: SYNTH	**45.35[36.85, 48.25] dB** K: 2 λ : 0.01 γ: 0.001 N: 512 SM: SYNTH	**13.66[13.47, 13.98] dB** K: 2 λ :0.01 γ: 0.001 N: 512 SM: SYNTH

the strong violation of the NMF-based mixture model assumption present throughout the test mixtures, the results described in Table 5.2 and Table 5.3 suggest that the CMF(EP)-based separation procedure did not offer any substantial improvements over the NMF-based separation procedure. It is interesting to note, however, that the highest CMF(EP)-based separation procedure performance was obtained using a consistency weight value $\gamma = 0.001$. This suggests that the presence consistency constraint, albeit a relatively small weighting, seemed to help the overall CMF(EP)-based separation performance under conditions for which a strong violation of the NMF mixture model exists. It is important to restate, however, that further testing is required to better understand the nature of this constraint.

Though these initial CMF(EP)-based results do not suggest a substantial improvement in performance over the NMF-based separation procedure, the results of the second experiment, which are presented in Table 5.4, are much more promising. These results suggest that the CMF(OP)-based separation procedure obtained a substantially higher median SIR performance measure, while still being able to retain higher median SDR and SAR performance measures, compared to both the NMF-based and CMF(EP)-based separation procedures. An improvement in median SDR, SIR, and SAR, of 3.25 dB, 27.24 dB, and 1.83 dB, respectively, was observed using the CMF(OP)-based separation procedure over the NMF-based separation procedure.

A graphical summary of the BSS performance measures obtained using the synthesis-based source estimation method for the NMF-based, CMF(EP)-based, and the CMF(OP)-based separation procedures, is presented in the form of box plots on a note-by-note basis,

in Figures 5.2 - 5.4. Each box plot is made up of a central line indicating the median of the data, upper and lower box edges indicating the 25^{th} and 75^{th} percentiles, vertical whiskers extending to the minimum and maximum extrema, and blue stars indicating the outliers. A data point is considered to be an outlier if it lies beyond 1.5 times the interquartile range, either below or above the first and third quartiles, respectively. Note that the data in each box-plot corresponds to the results obtain for both sources. As such, each box-plot is a summary of a sample of 20 data points (one data point for each of the 10 trials $\times$ 2 sources). The axis labelled "PIANO NOTE" indicates which piano note was used for the mixture (PN:C4 - PN:C5), the guitar note being held fixed at C4 (GTR:C4).

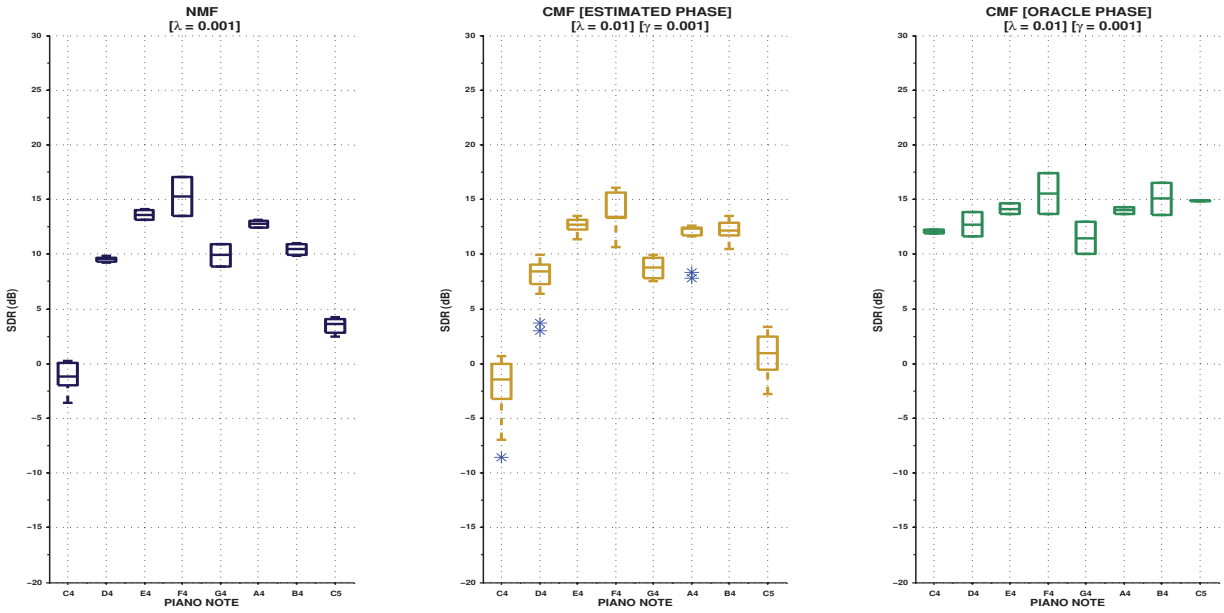

Fig. 5.2: COMBINED SDR RESULTS FOR BOTH SOURCES ACROSS ALL NOTE PAIRINGS USING THE SYNTHESIS-BASED SOURCE ESTIMATION METHOD FOR NMF-BASED (BLUE), CMF(EP)-BASED (YELLOW), AND CMF(OP)-BASED (GREEN) SEPARATION PROCEDURES. HIGHER VALUES ARE AN INDICATION OF BETTER PERFORMANCE

The SDR, SIR, and SAR results obtained for both the NMF-based and CMF(EP)-based separation procedures appear to be distributed fairly equally on a note-by-note basis. The box-plots indicate the lower BSS measures were obtained using the (PN:C4 GTR:C4) and (PN:C5 GTR:C4) note pairings, for which strong spectral overlap between the shared harmonics occurs. Better performances in all three BSS measures resulted from the other 6 note pairings (PN:D4 - PN:B4 GTR:C4), for which the amount of spectral overlap is not as high. The box-plots also indicate that the SDR, SIR, and SAR results obtained for the CMF(OP)-based separation procedure are not distributed in the same manner as the perfor-

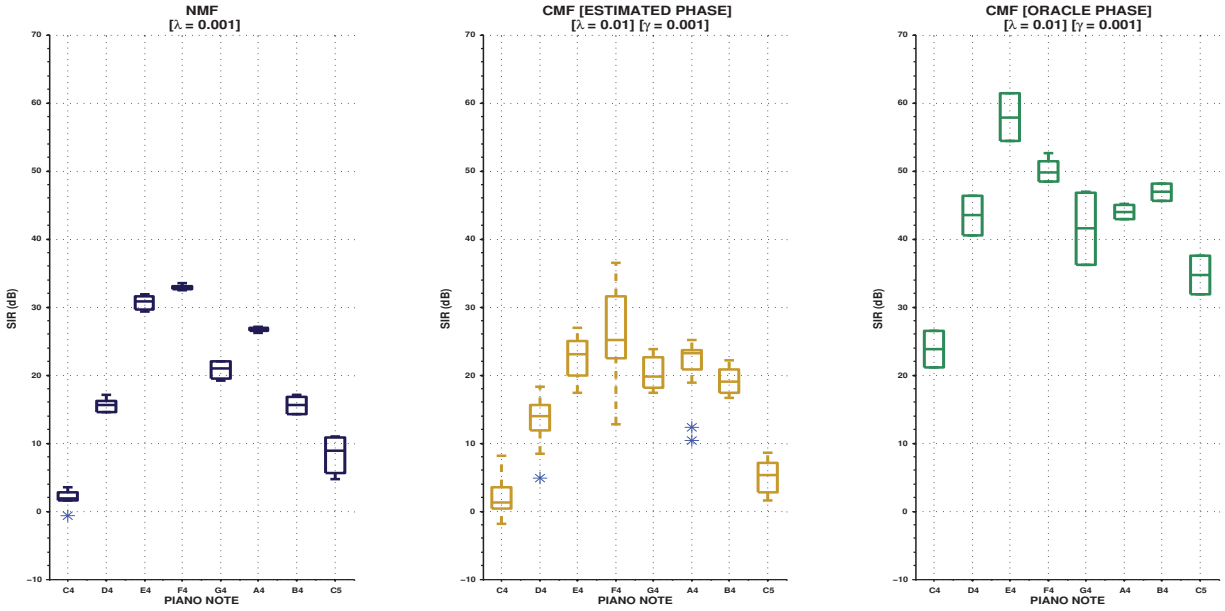

Fig. 5.3: COMBINED SIR RESULTS FOR BOTH SOURCES ACROSS ALL NOTE PAIRINGS USING THE SYNTHESIS-BASED SOURCE ESTIMATION METHOD FOR NMF-BASED (BLUE), CMF(EP)-BASED (YELLOW), AND CMF(OP)-BASED (GREEN) SEPARATION PROCEDURES. HIGHER VALUES ARE AN INDICATION OF BETTER PERFORMANCE

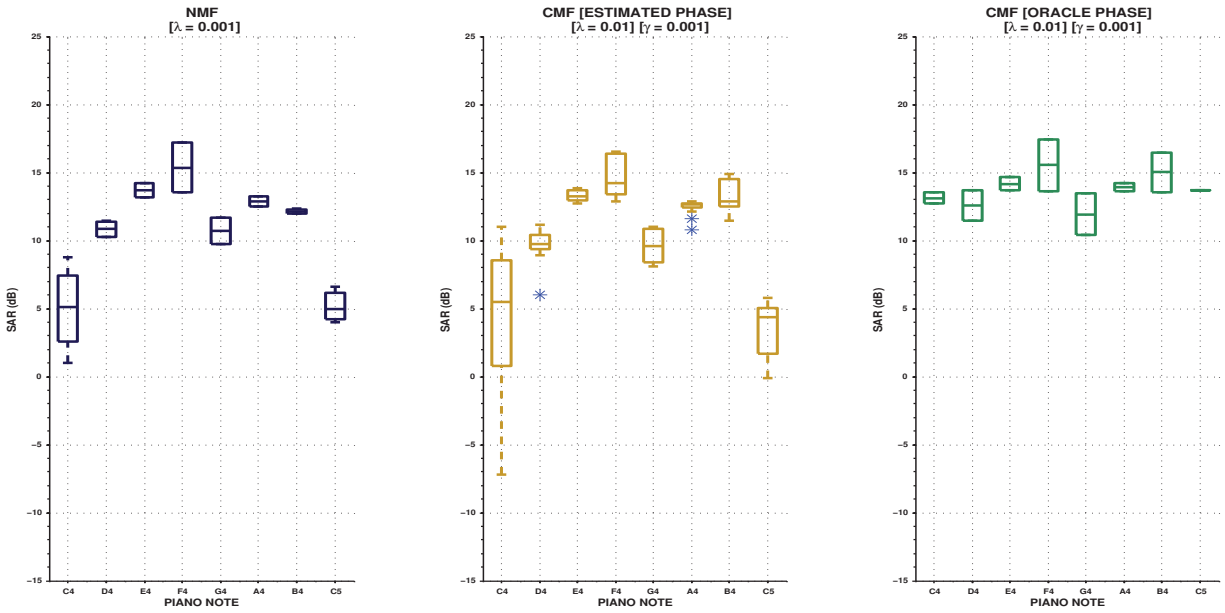

Fig. 5.4: COMBINED SAR RESULTS FOR BOTH SOURCES ACROSS ALL NOTE PAIRINGS USING THE SYNTHESIS-BASED SOURCE ESTIMATION METHOD FOR NMF-BASED (BLUE), CMF(EP)-BASED (YELLOW), AND CMF(OP)-BASED (GREEN) SEPARATION PROCEDURES. HIGHER VALUES ARE AN INDICATION OF BETTER PERFORMANCE

mance measures obtained for the NMF-based and CMF(EP)-based separation procedures. The median SIR performance measures obtained using the CMF(OP)-based separation procedure are substantially higher than the median SIR performance measures obtained for the other factorization-based separation procedures, on a note-by-note basis. From the box-plots, it can also be observed that the median SDR, and SAR performance measures for the CMF(OP)-based separation procedure are relatively flat and do not share the same downward arching shape, decreasing towards the (PN:C4 GTR:C4) and (PN:C5 GTR:C4) note pairings, as is observed in the SDR and SAR results of the NMF-based and CMF(EP)-based separation procedures. In fact, the CMF(OP)-based separation procedure obtained substantially higher SDR and SAR results compared to the other two factorization-based separation procedures when applied to the (PN:C4 GTR:C4) or (PN:C5 GTR:C4) note pairings, despite the strong spectral overlap of the shared harmonics.

It is interesting to note that the largest improvements in the median SIR performance measure, using the CMF(OP)-based separation procedure, seemed to occur with the mixtures involving the PN:C4, PN:D4, PN:E4, PN:B4, and PN:C5 sources. The mixtures resulting from the GTR:C4 source combined with any of these five piano sources, possess a strong violation of the NMF mixture model assumption within the first two partials of either source. As discussed in Section 4.3, it is hypothesized that since only $K = P$ templates were being extracted, a good NMF/CMF(EP) factorization was produced if the highest energy, which typically corresponds to the lower harmonics, is accurately described. Therefore, it is further hypothesized that since a strong violation of the NMF mixture model assumption was present in the first two partials of the (PN:C4 GTR:C4), (PN:D4 GTR:C4), (PN:E4 GTR:C4), (PN:B4 GTR:C4), (PN:C5 GTR:C4) mixtures, the CMF(OP)-based separation procedure was able to offer larger improvements in performance over the NMF-based separation procedure. This idea is explored further in the third experiment, described in Chapter 6.

Audible Differences in Performance

An noticeable improvement can be heard in the reconstructed sources estimated using the CMF(OP)-based separation procedure compared to the reconstructed sources estimated using the NMF-based and CMF(EP)-based separation procedures. Note that, the extracted PN:D4 and PN:B4 sources both contained an audible fluctuation in amplitude during the

last second of the mixture using both the NMF-based and CMF(EP)-based separation procedures. These fluctuations in amplitude are a result of the violation of the NMF mixture model assumption caused by the phase interactions between the closely spaced harmonics of the (PN:D4 GTR:C4) and (PN:B4 GTR:C4) mixtures, as depicted in Figure 5.1. These fluctuations, however, were audibly suppressed using the CMF(OP)-based separation procedure. Its also worth noting that, unlike the NMF-based or CMF(EP)-based separation procedures, the CMF(OP)-based separation procedure was able to produce a clear separation of the (PN:C4 GTR:C4) mixture, which can be heard in the audio playback of the reconstructed source estimates.

5.5 Conclusion

The experiment presented in this chapter explored the performance of the CMF-based separation procedure using a window length of $L = 46$ ms, which was demonstrated to offer the largest violation of the NMF mixture model assumptions compared to all other window lengths considered $L \in \{46, 93, 186, 372\}$ ms. The CMF-based separation performance was examined under oracle phase conditions for which the true phase of the unmixed sources was incorporated into the factorization algorithm. It was concluded that the CMF(EP)-based separation procedure did not offer any substantial improvements over the NMF-based separation procedure in terms of the SDR, SAR, and SIR performance measures. It was observed, however, that the CMF(OP)-based separation procedure was able to achieve a substantially higher median SIR performance result and more favourable median SDR and SAR results compared to both the NMF-based and CMF(EP)-based separation procedures. Although the phase of the unmixed sources are typically not available in practice, the results of the CMF(OP)-based separation procedure helps provide insight into a possible upper bound of achievable separation performance, for this particular dataset, using the CMF-based separation procedure with the parameter value combination of $L = 512$, $K = 2$, $\lambda = 0.01$, and $\gamma = 0.001$.

Overall, the results of this experiment suggest that a strong separation performance is obtainable if the phase estimates for each source can be accurately determined. The fact that the CMF(OP)-based separation procedure offered substantial improvements over the NMF-based separation procedure, while the CMF(EP)-based source separation procedure failed to offer substantial improvements over the NMF-based separation procedure, suggests

that the phase matrix estimated for each source using CMF(EP) may not be updating to a good approximation of the true phase for each source. In order to help improved the CMF(EP)-based separation procedure, it was questioned whether source specific information could be incorporated into the phase update algorithm to facilitate the estimation of the phase matrix for each source. This idea ultimately motivated the development of the phase constraint proposed in the Chapter 6.

Another notable conclusion drawn from the results of this experiment is that the CMF(OP)-based separation procedure was able to noticeably suppress the fluctuations in amplitude resulting from the violation of the NMF mixture model assumption caused by the phase interactions between the closely spaced harmonics of the (GTR:C4 - PN:D4) and (GTR:C4 - PN:B4) mixtures. This result could suggest that when a strong violation of the NMF mixture model assumption is present in the lower harmonics of the mixture, the CMF(OP)-based separation procedure may be more accurately describing the amplitude of the true source and not the amplitude of the resulting mixture, which is influenced by the phase of the overlapping sources. Further investigation on this matter is presented in the experiment of Chapter 6.

Chapter 6

Experiment 3: NMF vs. CMF with Modelled Phase

6.1 Motivation

Based on the results of the second experiment, it was hypothesized that, in conditions for which the oracle phase of the unmixed sources is unknown, a superior CMF(EP)-based separation procedure performance could be obtained if source specific knowledge, in the form of a physically motivated phase-based constraint, were incorporated into the CMF(EP) algorithm. This chapter details the development of a newly proposed phase constraint, henceforth referred to as the "phase evolution" constraint. The chapter also summarizes an experiment, conducted as a proof of concept, in which the newly proposed phase evolution constraint was integrated into the CMF(EP)-based separation procedure and applied to separate the mixture constructed from the (PN:D4 GTR:C4) note pairing. This note pairing was chosen, in particular, due to the high energy overlap between the first harmonic of each source. A strong oscillatory change in amplitude, suggesting a strong violation of the NMF mixture model assumption due to the phase difference between the sources, is observed during the last second of the mixture using this note pairing. This effect can be observed in the overlapping first harmonics of the (PN:D4 GTR:C4) note pairing, displayed in Figure 3.1 and in Figure 5.1. It was hypothesized that the CMF(EP)-based separation procedure, together with this newly proposed phase evolution constraint, could offer improvements over the performance of both the NMF-based and CMF(EP)-

based separation procedures without the phase evolution constraint, when applied to this note pairing. Henceforth, CMF with the inclusion of the phase evolution constraint shall be denoted CMF(MP), where MP stands for "Modelled Phase".

6.2 Phase Model/Assumptions

The following model/assumption were used for the development of the phase evolution constraint. Note that similar assumptions are also adopted in the work of (Li et al. 2009).

A-1 Assume that the number of sources is known and that only one component is set to be extracted per source (i.e., $K = P$).

A-2 Assume that the fundamental frequency of each source[1] is known.

A-3 Assume that each source is well modelled as a sum of sinusoids[2]:

$$\tilde{s}_p(t) = \sum_{r=1}^{R_p} A_r \exp(i(2\pi f_{0_p} r t + \phi_{0_{p,r}})), \tag{6.1}$$

where f_{0_p} corresponds to the fundamental frequency (in Hz) of the p^{th} source, r corresponds to the harmonic number starting from r = 1 (the fundamental frequency) to R_p (the total number of harmonics considered for source p) where $f_{0_p} R_p < \frac{Fs}{2}$, t corresponds to the continuous time variable, and $\phi_{0_{p,r}}$ corresponds to the initial[3] phase of the r^{th} harmonic of the p^{th} source.

A-4 Assume the energy of a given harmonic does not extend beyond the frequency bins which fall under the main lobe of the Fourier transform of the analysis window centered about the frequency of that harmonic.

[1] For the purposes of this constraint development, each source is considered to be a single musical note with a well defined fundamental frequency.

[2] For convenience of the following analysis, the analytic form of the signal is taken. In actuality, the signal should be modelled as either the real or complex component of the analytic form. However, as only positive frequencies are considered, the analysis of the phase evolution is identical using the analytic form the signal.

[3] The proposed model describes a particular instance of the activated source and, as such, does not account for the fact that the initial phase may change between different activations of the source. In other words, multiple onsets of the same source are not described using this model.

6.3 Phase Evolution Constraint Development

6.3.1 STFT Considerations

Recall from Chapter 3 that the definition of the $STFT$ used for the experiments conducted throughout this thesis is as follows:

$$[STFT\{\mathbf{x}\}]_{n,m} = [\mathbb{X}]_{n,m} = \sum_{l=0}^{\tilde{N}-1} x(l+mH)w(l)\exp(\frac{-i2\pi ln}{\tilde{N}}) \tag{6.2}$$

No zero padding is used, the hop size is set to $\tilde{H} = \tilde{N}/4$ and the window is a modified Hann window as defined in Equation (3.5).

6.3.2 Harmonic Analysis

After sampling in time at a rate of $1/T = Fs$ and using l as a sampling index (i.e., taking discrete time samples of $t = lT$ for $l = 0, ..., \tilde{L}$, where $\tilde{L}$ is the total number of samples of the signal being analyzed), the STFT of the sinusoidal model described in assumption A-2 above, can be expressed as:

$$\begin{aligned}
[STFT\{\tilde{\mathbf{s}}_\mathbf{p}\}]_{n,m} = [\tilde{\mathbb{S}}]_{n,m} &= \sum_{l=0}^{\tilde{N}-1}\sum_{r=1}^{R_p} A_r \exp(i(2\pi f_{0_p} r(l+mH)T + \phi_{0_{p,r}}))w(l)\exp(\frac{-i2\pi ln}{\tilde{N}}) \\
&= \sum_{l=0}^{\tilde{N}-1}\sum_{r=1}^{R_p} A_r \exp(i2\pi f_{0_p} rmHT + \phi_{0_{p,r}})w(l)\exp(-i2\pi l(\frac{n}{\tilde{N}} - f_{0_p}rT)) \\
&= \sum_{r=1}^{R_p} A_r \exp(i(2\pi f_{0_p} rmHT + \phi_{0_{p,r}}))\sum_{l=0}^{\tilde{N}-1} w(l)\exp(-i2\pi l(\frac{n}{\tilde{N}} - f_{0_p}rT)) \\
&= \sum_{r=1}^{R_p} A_r \exp(i(2\pi f_{0_p} rmHT + \phi_{0_{p,r}}))\mathbb{W}(2\pi(\frac{n}{\tilde{N}} - f_{0_p}rT))
\end{aligned} \tag{6.3}$$

where $\mathbb{W}$ is the Discrete Time Fourier Transform ($DTFT$) of $\mathbf{w}$:

$$\mathbb{W}(\Omega) = DTFT\{\mathbf{w}\}(\Omega) = \sum_{l=0}^{\tilde{N}-1} w(l)\exp(-il\Omega) \tag{6.4}$$

with $\Omega \in \mathbb{R}$ being the continuous variable representing frequency. Now, the centre frequency of the k^{th} bin can be expressed as $F_c(n) = \frac{nFs}{\tilde{N}}$ (Hz). Let $\hat{n}_{p,r}$ be the index of the frequency bin whose centre frequency is closest to the frequency of the r^{th} harmonic of the p^{th} source. Formally, $\hat{n}_{p,r} = \min_n |F_c(n) - f_{0_p} r|$. Under the assumption of A-4, the energy corresponding to all the other harmonics at bin $n = \hat{n}_{p,r}$ is taken to be negligible. As such, the STFT of the p^{th} source at the $\hat{n}_{p,r}$ frequency bin and m^{th} time frame can be expressed as:

$$[\tilde{\mathbb{S}}_p]_{\hat{n}_{p,r},m} = A_r \exp(i(2\pi f_{0_p} r m H T + \phi_{0_{p,r}}))\mathbb{W}(2\pi(\frac{\hat{n}_{p,r}}{\tilde{N}} - f_{0_p} r T)) \tag{6.5}$$

6.3.3 Phase Analysis

The observed phase of the $(n,m)^{th}$ element of $\tilde{\mathbb{S}} \in \mathbb{C}^{N \times M}$ can be expressed mathematically as:

$$[\Phi_p]_{n,m} = \arctan\left(\frac{\Im[\tilde{\mathbb{S}}]_{n,m}}{\Re[\tilde{\mathbb{S}}]_{n,m}}\right) \tag{6.6}$$

By keeping track of the signs of the numerator and denominator, the phase can be specified within the bounds of $[-\pi, \pi]$. Consider the "true phase" as being the phase of the observed complex sinusoid at a given time frame and frequency bin, keeping track of the exact number of rotations over the time frames starting from an initial phase $\phi_{0_{p,r}}$, whereas the "observed phase" is the phase calculated at each time frame and frequency bin using Equation (6.6). The true phase of source p at time frame m and frequency bin n shall be denoted as $[\Phi_p^{\text{TRUE}}]_{n,m}$, whereas the observed phase of source p at time frame m and frequency bin n shall be simply denoted as $[\Phi_p]_{n,m}$. The relationship between $[\Phi_p^{\text{TRUE}}]_{n,m}$ and $[\Phi_p]_{n,m}$ can be expressed as follows:

$$[\Phi_p^{\text{TRUE}}]_{n,m} = [\Phi_p]_{n,m} + 2\pi q_m \tag{6.7}$$

where $q_m \in \mathbb{Z}$ represents the unknown number of integer rotations of 2π after m frames, which are not accounted for using Equation (6.6). Now, the difference in "true phase" of the p^{th} source between frame m and $m-1$ (for $m \geq 2$) at bin $n_{p,r}$ can be determined from Equation (6.5) as follows:

$$\begin{aligned}
\Delta[\Phi_p^{\text{TRUE}}]_{n_{p,r},m} &= [\Phi_p^{\text{TRUE}}]_{n_{p,r},m} - [\Phi_p^{\text{TRUE}}]_{n_{p,r},m-1} \\
&= 2\pi f_{0_p} r m\tilde{H}T + \angle\mathbb{W}(2\pi(\frac{\hat{n}_{p,r}}{\tilde{N}} - f_{0_p}rT)) + \cancel{\phi_{0_{p,r}}} \\
&- 2\pi f_{0_p} r(m-1)HT - \angle\mathbb{W}(2\pi(\frac{\hat{n}_{p,r}}{\tilde{N}} - f_{0_p}rT)) - \cancel{\phi_{0_{p,r}}} \\
&= 2\pi f_{0_p} r m\tilde{H}T - 2\pi f_{0_p} r(m-1)\tilde{H}T \\
&= 2\pi f_{0_p} r\tilde{H}T
\end{aligned} \tag{6.8}$$

Using the relationship between the "true phase" and the observed phase expressed in Equation (6.7), the phase difference in observed phase of the p^{th} source between frame m and $m-1$ at bin $n_{p,r}$ can be determined as follows:

$$\begin{aligned}
\Delta[\Phi_p]_{n_{p,r},m} &= [\Phi_p]_{n_{p,r},m} - [\Phi_p]_{n_{p,r},m-1} \\
&= [\Phi_p^{\text{TRUE}}]_{n_{p,r},m} + 2\pi q_m - [\Phi_p^{\text{TRUE}}]_{n_{p,r},m-1} - 2\pi q_{m-1} \\
&= 2\pi f_{0_p} r\tilde{H}T + (2\pi q_m - 2\pi q_{m-1})
\end{aligned} \tag{6.9}$$

Taking the complex exponential of both sides of Equation (6.9):

$$\begin{aligned}
\exp(i([\Phi_p]_{n_{p,r},m} - [\Phi_p]_{n_{p,r},m-1})) &= \exp(i(2\pi f_{0_p} r\tilde{H}T + (2\pi q_m - 2\pi q_{m-1}))) \\
\exp(i[\Phi_p]_{n_{p,r},m})\exp(-i[\Phi_p]_{n_{p,r},m-1}) &= \exp(i2\pi f_{0_p} r\tilde{H}T)\cancel{\exp(i2\pi(q_m - q_{m-1}))}^{\,1} \\
\exp(i[\Phi_p]_{n_{p,r},m}) &= \exp(i[\Phi_p]_{n_{p,r},m-1})\exp(i2\pi f_{0_p} r\tilde{H}T) \\
\implies 0 &= \exp(i[\Phi_p]_{n_{p,r},m}) \\
&\quad - \exp(i[\Phi_p]_{n_{p,r},m-1})\exp(i2\pi f_{0_p} r\tilde{H}T)
\end{aligned} \tag{6.10}$$

6.3.4 Proposed Phase Evolution Cost function

Before introducing the phase evolution cost function, some notation is presented to make the definition of the cost function more concise. First, let $\mathcal{N}_{p,r}$ be the set of all frequency bins, n, which fall under the main lobe of the Fourier transform of the analysis window centered about the frequency $f_{0_p}r$ (Hz). In the case of the modified Hann window used throughout this thesis, the main lobe of the Fourier transform of the analysis window spans 4 frequency bins, as depicted below in Figure 6.1. Thus, for each of the p sources, and each

of the r harmonics, there will be a corresponding set, $\mathcal{N}_{p,r}$, consisting of 4 elements (one for each frequency bin which falls under the main lobe of the Fourier transform of the analysis window centered about the frequency $f_{0_p}r$ (Hz)). For example, in the case of the window depicted in Figure 6.1, the window is centered closely to (but not exactly at) frequency bin n = 10. For the purpose of this example, suppose that the fundamental frequency of the first source ($p = 1$) was slightly greater than the centre frequency of bin $n = 10$. Thus, using the notation introduced earlier, $n_{1,1} = \min_n |F_c(n) - f_{0_1}| = 10$. Since the main lobe of the window is 4 bins wide, when centered about the fundamental of the first source, the energy of the main lobe will extend from bin $n = 9$ to bin $n = 12$. Thus, the set $\mathcal{N}_{1,1}$ would be equal to $\mathcal{N}_{1,1} = \{9, 10, 11, 12\}$. Note that if the fundamental frequency of the first source ($p = 1$) were slightly less than the centre frequency of bin n = 10, then the main lobe of the window would span over the bins $n = 8$ to $n = 11$. In this case, the set $\mathcal{N}_{1,1}$ would be equal to $\mathcal{N}_{1,1} = \{8, 9, 10, 11\}$. Under the assumptions of the phase constraint development, the second harmonic (r = 2) would have a frequency of $2f_{0_1}$ which would fall (depending on the difference in frequency between the fundamental and the centre frequency of the 10^{th} bin) near bin $n = 20$. For the purpose of this example, suppose that the frequency of the second harmonic falls just above the centre frequency of the 20^{th} bin, (i.e., $n_{1,2} = \min_n |F_c(n) - 2f_{0_1}| = 20$). Thus the set $\mathcal{N}_{1,2}$ would be equal to $\mathcal{N}_{1,2} = \{19, 20, 21, 22\}$. To summarize, for a given source, p, and a given harmonic, r, the sets $\mathcal{N}_{p,r}$ are constructed, which contain the 4 frequency bins that fall under the main lobe of the Fourier transform of the analysis window centered about the frequency, $f_{0_p}r$, corresponding to the r^{th} harmonic of the p^{th} source.

Now, let $\mathbb{1}_{\mathcal{N}_{p,r}}$ be used as a indicator function specifying the membership of the n^{th} frequency bin to the set $\mathcal{N}_{p,r}$:

$$\mathbb{1}_{\mathcal{N}_{p,r}} = \begin{cases} 1 & : n \in \mathcal{N}_{p,r} \\ 0 & : n \notin \mathcal{N}_{p,r} \end{cases} \tag{6.11}$$

Using this notation and the results of Equation (6.10) the following cost function is introduced and incorporated into the CMF framework as follows:

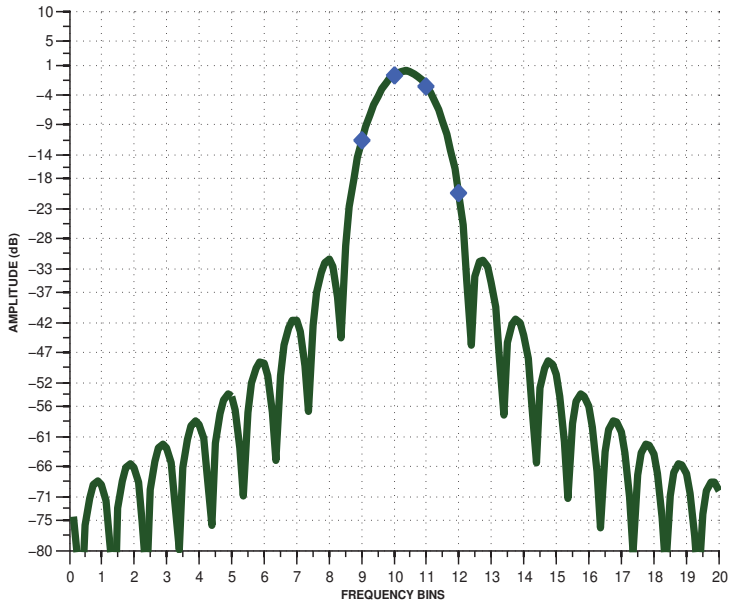

Fig. 6.1: FOURIER TRANSFORM OF MODIFIED HANN WINDOW (IN GREEN) CENTERED BETWEEN BINS $n = 10$ AND $n = 11$ AND THE VALUES OF THE WINDOW AT BINS $n = 9$ TO $n = 12$ (BLUE DIAMONDS) CORRESPONDING TO THE BINS WHICH FALL UNDER THE MAIN LOBE OF THE SPECTRUM OF THE WINDOW

Phase Evolution Cost Function:

$$C(\theta)_{\text{PE}} = \sum_{n,p,r,m} \mathbb{1}_{\mathcal{N}_{p,r}} |\exp(i[\Phi]_{n,p,m}) - \exp(i[\Phi]_{n,p,m-1}) \exp(i2\pi f_{0_p} r \tilde{H} T)|^2 \tag{6.12}$$

The cost function, $C(\theta)_{\text{PE}}$ is minimized when the complex exponential of the observed phase of the signal, at time frame m, and frequency bin $n \in \mathcal{N}_{p,r}$ (for some source, p, and harmonic, r) equals the complex exponential of the observed phase of the signal at time frame $m-1$ with an additional rotation of $2\pi f_{0_p} r \tilde{H} T$. In other words, the cost function is minimized if the harmonics of each source evolve according to the assumptions A-1 to A-4.

The validity of the phase evolution model/assumptions was analyzed by calculating the magnitude of the summand of the cost function described in Equation (6.12), for each source, p, using $R_p = 10$ harmonics. The results are presented below in Figure 6.3 for the piano sources and in Figure 6.2 for the guitar source. The highest possible value achieved for this constraint occurs when $\exp(i[\Phi]_{n,p,m}) = -\exp(i[\Phi]_{n,p,m-1}) \exp(i2\pi f_{0_p} r \tilde{H} T)$, which yields a value of $2^2 = 4$ in the summand of Equation (6.12). As such, the scale was set

to range from $20\log_{10}(4)$ dB to $20\log_{10}(4) - 20$ dB. Thus, the lower values (indicated in blue) correspond to magnitudes less than or equal to 0.1 (20 dB below a reference of unity). The figures indicate that the assumed phase evolution model holds fairly well (indicated in blue) for the lower harmonics whereas larger deviations from the model (indicated in red) are observed in the data for the upper harmonics[4]. The inaccuracies in the model, specifically for the upper harmonics, could possibly be attributed to the fact that neither source is perfectly harmonic (the piano source, in particular, is known to be inharmonic (Fletcher et al. 1962)) or to the fact that the an error in the fundamental frequency estimate will propagate proportional to the harmonic number. For instance, if the fundamental frequency estimate differs from the true fundamental frequency of the source by 5Hz, then the frequency estimate for the r^{th} harmonic will differ from the true frequency of the r^{th} harmonic by $5r$ Hz, assuming a perfectly harmonic structure for the sources. That being said, the model does seem to offer a fairly accurate description of the phase evolution of the first 5 harmonics of PN:D4 and the first 8 harmonics of GTR:C4, on which this experiment specifically focuses.

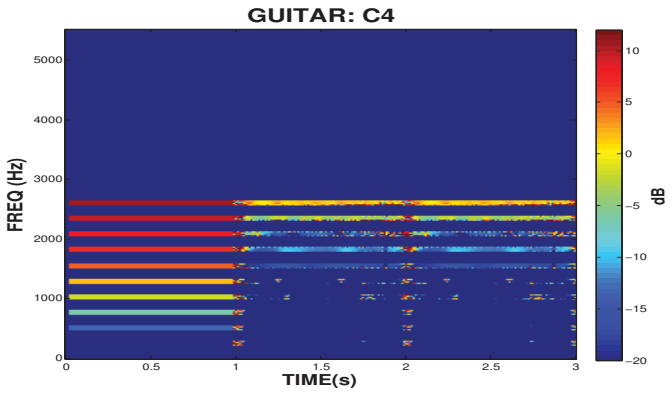

Fig. 6.2: PHASE EVOLUTION MODEL/ASSUMPTIONS VALIDITY FOR GUITAR SOURCE (HIGH DISCREPANCY BETWEEN MODEL AND OBSERVED DATA INDICATED IN RED - LOW DISCREPANCY BETWEEN MODEL AND OBSERVED DATA INDICATED IN BLUE)

[4]Note that only the values of the bins $n \in \mathcal{N}_{p,r}$ are of interest. As such, the smaller values (indicated in blue) in between the harmonics and after the 10th harmonic are not indicative of a good representation of the observed data using the model, rather they are indicative of the fact that the model does is not being applied to these regions.

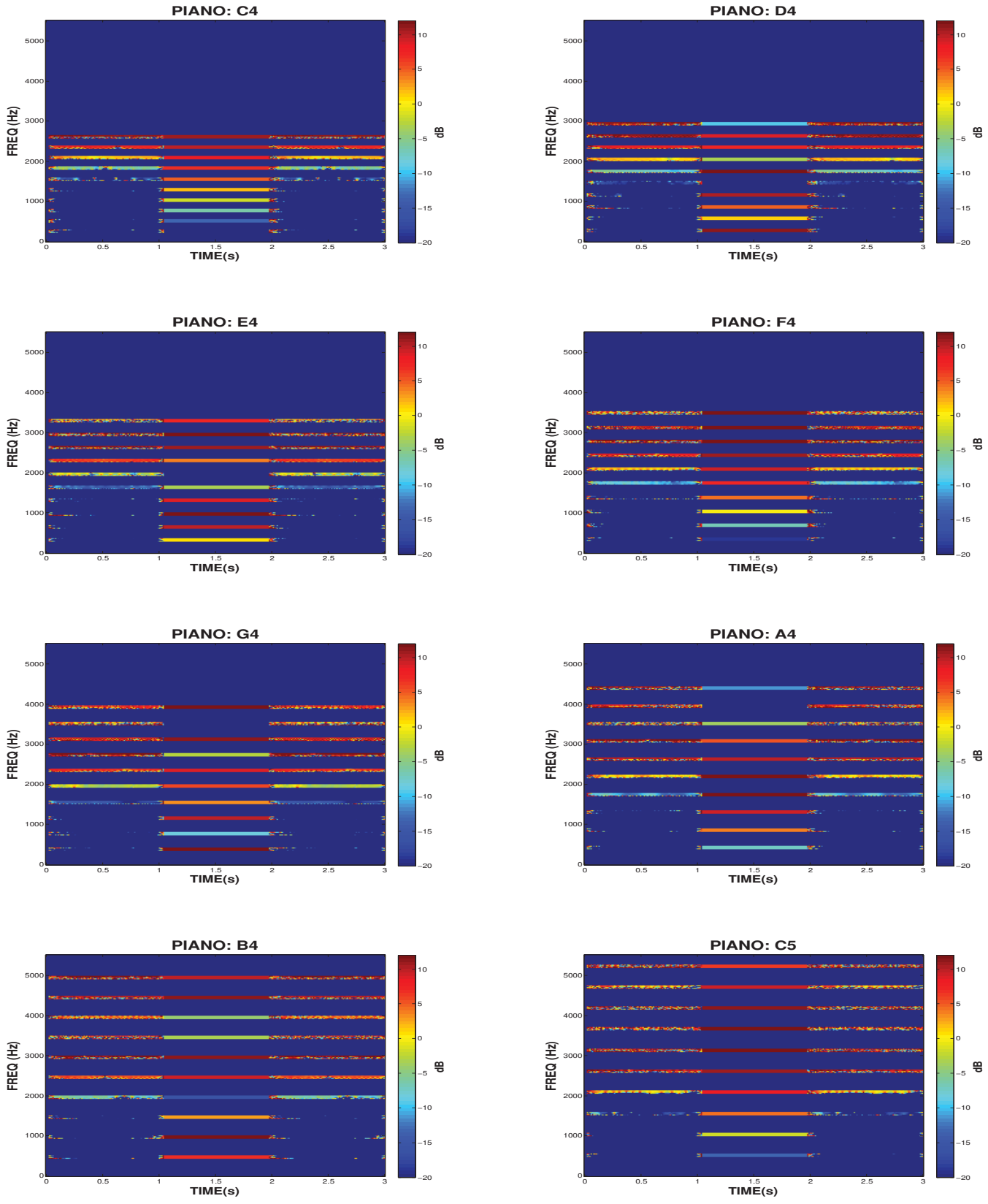

Fig. 6.3: PHASE EVOLUTION MODEL/ASSUMPTIONS VALIDITY FOR PIANO SOURCES (HIGH DISCREPANCY BETWEEN MODEL AND OBSERVED DATA INDICATED IN RED - LOW DISCREPANCY BETWEEN MODEL AND OBSERVED DATA INDICATED IN BLUE)

6.3.5 Optimization of the Phase Evolution Cost Function

The entire CMF(MP) optimization problem, now incorporating the phase evolution cost function, can be stated as follows:

$$\begin{aligned}
&\text{Given} : \mathbb{X} \in \mathbb{C}^{NxM},\ P \in \mathbb{R}_{>0} \text{ and } f_{0_p}(p = 1...P) \\
&\text{Optimize} : \min_{W,H,\Phi} C(\theta) = \sum_{n,m} |[\mathbb{X}]_{n,m} - [\widehat{\mathbb{X}}]_{n,m}|^2 + 2\lambda \sum_{n,m} |[H]_{m,n}|^g + \gamma \sum_{n,p,m} |\mathcal{F}([\widehat{C}_p])_{n,m}|^2 \\
&\qquad + \sigma \sum_{n,p,r,m} \mathbb{1}_{\mathcal{N}_{p,r}} |\exp(i[\Phi]_{n,p,m}) - \exp(i[\Phi]_{n,p,m-1}) \exp(i2\pi f_{0_p} r \tilde{H} T)|^2 \\
&\text{Subject to} : \sum_{n} [W]_{n,p} = 1\ (\forall p = 1, ..., P),\ W \in \mathbb{R}_{\geq 0}^{N \times P},\ H \in \mathbb{R}_{\geq 0}^{P \times M} \\
&\qquad \text{and } \Phi \in \mathbb{R}^{N \times P \times M}
\end{aligned} \tag{6.13}$$

Note that under assumption A-1, the number of components is equal to the number of sources ($K = P$), as such, the index k has been replaced by the index p in the above equation and the equations to follow. To be clear, the first two terms in the optimization problem described in Equation (6.13) correspond to the cost of the factorization approximation and the sparsity penalty function, respectively. The third term corresponds to the consistency constraint penalty function, and the fourth term corresponds to the newly introduced phase evolution penalty function. The scaling factor, σ, is used to weight the relative importance of the phase evolution penalty function in an analogous manner to how λ and γ are used to weight the relative importance of the sparsity penalty and the consistency penalty functions, respectively. Unlike the optimization of the cost functions established for the matrix factorization, sparsity constraint and consistency constraint, the optimization of the phase evolution cost function does not need to be approached under the MM framework. As derived mathematically in Appendix A, the minimization with respect to the phase parameter, $[\Phi]_{n,p,m}$, can be performed without the need for an auxiliary function, due to the simplicity of the structure of the phase evolution cost function. The minimization of the phase evolution cost function yields the following contribution to the phase parameter update:

$$[\Phi]_{\tilde{n},\tilde{p},\tilde{m}} = Arg\Bigg(\sigma \sum_{r} \mathbb{1}_{\mathcal{N}_{p,r}} \Bigg(\exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}-1}) \exp(i2\pi f_{0_{\tilde{p}}} r\tilde{H}T) + \exp(-i2\pi f_{0_{\tilde{p}}} r\tilde{H}T) \exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}+1}) \Bigg) \Bigg) \tag{6.14}$$

As demonstrated in Appendix A, the contribution of the phase parameter update corresponding to the phase evolution penalty can be incorporated into the global phase parameter update, which includes the contribution from the factorization cost function and the consistency constraint cost function as follows:

$$[\Phi]_{\tilde{n},\tilde{p},\tilde{m}} = Arg\Bigg(\frac{\bar{\mathbb{X}}]_{\tilde{n},\tilde{p},\tilde{m}}}{[B]_{\tilde{n},\tilde{p},\tilde{m}}} [W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}} + \gamma[\bar{\mathbb{L}}]_{\tilde{n},\tilde{p},\tilde{m}}[W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}} + \sigma \sum_{r} \mathbb{1}_{\mathcal{N}_{p,r}} \Bigg(\exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}-1}) \exp(i2\pi f_{0_{\tilde{p}}} r\tilde{H}T) + \exp(-i2\pi f_{0_{\tilde{p}}} r\tilde{H}T) \exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}+1}) \Bigg) \Bigg) \tag{6.15}$$

6.4 CMF with Phase Evolution Penalty Function

The phase evolution cost function has no dependence on any other parameters other than $[\Phi]_{n,p,m}$ and, as such, all other parameter updates will remain as previously defined in Section 3.2. The entire CMF with the phase evolution constraint is presented below. Note that the only parameter update which has changed is the $[\Phi]_{n,p,m}$ update. Also note that the required input now includes the fundamental frequency, f_{0_p}, of each of the p sources.

Algorithm 5 CMF WITH PHASE EVOLUTION CONSTRAINT

Input: $\mathbb{X} \in \mathbb{C}_{+}^{NxM}$ and $P \in \mathbb{R}_{>0}$ and $f_{0_p}(p = 1...P)$

Output: W, H, Φ s.t $[\mathbb{X}]_{n,m} \approx \sum_{p=1}^{P} [W]_{n,p}[H]_{p,m} \exp(i[\Phi]_{n,p,m})$

$W \in \mathbb{R}_{\geq 0}^{NxP}$, $H \in \mathbb{R}_{\geq 0}^{PxM}$, $\Phi \in \mathbb{R}^{N \times P \times M}$ and $\sum_{n}[W]_{n,p} = 1$ $(\forall p = 1, ..., P)$

Initialize W, H, Φ, such that $W \in \mathbb{R}_{>0}^{NxP}$, $H \in \mathbb{R}_{>0}^{PxM}$ and $\Phi = \frac{\mathbb{X}}{|\mathbb{X}|}$

while Stopping criteria not met **do**

Compute B

$$[B]_{n,p,m} = \frac{[W]_{n,p}[H]_{p,m}}{\sum_{p}[W]_{n,p}[H]_{p,m}}$$

Compute $\bar{\mathbb{X}}$

$$[\bar{\mathbb{X}}]_{n,p,m} = [W]_{n,p}[H]_{p,m} \exp(i[\Phi]_{n,p,m}) + [B]_{n,p,m}([\mathbb{X}]_{n,m} - [\widehat{\mathbb{X}}]_{n,m})$$

Compute $\bar{H}$

$$[\bar{H}]_{p,m} = H_{p,m}$$

Compute $[\bar{\mathbb{L}}]$

$$[\bar{\mathbb{L}}]_{n,p,m} = \mathcal{G}(\widehat{C}_p)_{n,m}$$

Compute Φ

$$[\Phi]_{n,p,m} = Arg\Bigg(\frac{[\bar{\mathbb{X}}]_{n,p,m}}{[B]_{n,p,m}}[W]_{n,p}[H]_{p,m} + \gamma[\bar{\mathbb{L}}]_{n,p,m}[W]_{n,p}[H]_{p,m}$$

$$+\sigma \sum_{r} \mathbb{1}_{\mathcal{N}_{p,r}}\Big(\exp(i[\Phi]_{n,p,m-1}) \exp(i2\pi f_{0_p} r\tilde{H}T) + \exp(-i2\pi f_{0_p} r\tilde{H}T) \exp(i[\Phi]_{n,p,m+1})\Big)\Bigg)$$

Compute W

$$[W]_{n,p} = \frac{\sum_{m}[H]_{p,m}\Re[(\frac{\bar{\mathbb{X}}_{n,p,m}}{[B]_{n,p,m}} + \gamma[\bar{\mathbb{L}}]_{n,p,m}) \exp(-i[\Phi]_{n,p,m})]}{(\sum_{m}[H]_{p,m}^2)(\frac{1}{[B]_{n,p,m}} + \gamma)}$$

Compute H

$$[H]_{p,m} = \frac{\sum_{n}[W]_{p,m}\Re[(\frac{\bar{\mathbb{X}}_{n,p,m}}{[B]_{n,p,m}} + \gamma[\bar{\mathbb{L}}]_{n,p,m}) \exp(-i[\Phi]_{n,p,m})]}{(\sum_{n}[W]_{kp,m}^2)(\frac{1}{[B]_{n,p,m}} + \gamma) + \lambda g([\bar{H}]_{p,m})^{g-2}}$$

Project W and H onto non-negative orthant

Normalize W **and scale** H **accordingly**

iter = iter + 1

end while

6.5 Experimental Design

An experiment was conducted, as a proof of concept, to investigate the performance of the proposed CMF(MP)-based separation procedure, as it compares to the NMF-based, CMF(EP)-based and CMF(OP)-based separation procedures. Unlike the approach of the two previous chapters, source separation tasks were conducted only for the (PN:D4 GTR:C4) note pairing. The fundamental frequency of PN:D4 was set to be $f_{0_1} \approx 293$ Hz, whereas the fundamental frequency of GTR:C4 was set to be $f_{0_2} \approx 261$ Hz[5]. The same parameter value combination of $\lambda = 0.01$, $\gamma = 0.001$, $k = 2$ and $L = 512$, as established in Chapter 5, was used. The newly introduced phase evolution weight, σ, was taken to be a function of $[W]_{n,k}$ and $[H]_{k,m}$, as follows: $\sigma = \hat{\sigma}[W]_{n,p}[H]_{p,m}$ where $\hat{\sigma}$ varied within the range of: $\hat{\sigma} \in \{0, 0.001, 0.01, 0.1, 1\}$. Although defining σ in this way is not theoretically justified as it introduces a dependence on $[W]_{n,p}$, and $[H]_{p,m}$, which was not accounted for during the optimization of the phase parameter, this definition of σ was observed to yield superior results in practice and allowed for a more intuitive scaling, relative to λ and γ, to be used. In total, $8 \times 10 \times 5 = 400$ separations were conducted for this experiment (one for each of the 8 note pairings $\times$ 10 clustered initializations $\times$ 5 phase evolution weight levels). The newly proposed algorithm was slower than any of the other factorization-based separation procedures, averaging a separation every 9 minutes (except in the case of $\sigma = 0$ which took roughly 1 minute) for a total runtime of roughly 6 hours.

6.6 Results/Analysis

6.6.1 BSS Performance Measures Analysis

The median SDR, SIR, and SAR performance measures obtained in the second experiment were recalculated so that the results reflected the measures achieved when using the NMF-based, CMF(EP)-based, and CMF(OP)-based separation procedures applied exclusively to the (PN:D4 GTR:C4) note pairing. Again, only the results for the synthesis-based source estimation method were considered. The newly determined performance measures are displayed in tables: 6.1 - 6.3. The performance measures which resulted from the third experiment, using the CMF(MP)-based separation procedure, are found in Table 6.4,

[5]Here, PN:D4 is arbitrarily labelled as the first source, $p = 1$ and GTR:C4 is arbitrarily labelled as the second source, $p = 2$, for convenience.

highlighted in red. The results of Table 6.4 indicate that the CMF(MP)-based separation procedure was able to achieve a substantially higher performance in terms of the median SIR performance measure, while still preserving higher median SDR and SAR performance measures as compared to both the NMF-based and CMF(EP)-based separation procedures, given this particular note pairing. However, the CMF(MP)-based separation procedure did not achieve as high of performance for the CMF(OP)-based separation procedure for any performance measure, which is to be expected. The results of Table 6.4 also indicate that the highest median SDR, SIR, and SAR performance measures were obtained using a phase evolution weight of $\hat{\sigma} = 0.1$ for all three performance measures. This suggests that the inclusion of the phase evolution constraint is benefitting the separation performance for all three measures. As such, the parameter value combination of $\lambda = 0.01$, $\gamma = 0.001$, $\hat{\sigma} = 0.1$, was chosen for the CMF(MP)-based separation procedure.

Table 6.1: NMF MEDIAN RESULTS FOR (PN:D4 GTR:C4) NOTE PAIRING WITH MAX MEDIAN SIR PARAMETER VALUE COMBINATION

	NMF (PN:D4 GTR:C4)		
	MEDIAN [Q1, Q3]		
NOTE RANGE	[SDR]	[SIR]	[SAR]
	9.54[9.28, 9.67] dB	**15.62[14.58, 16.14] dB**	**10.83[10.29, 11.42] dB**
PN: D4	K: 2	K: 2	K: 2
	λ: 0.001	λ : 0.001	λ :0.001
GTR: C4	γ -:	γ: -	γ: -
	N: 512	N: 512	N: 512
	SM: SYNTH	SM: SYNTH	SM: SYNTH

Table 6.2: CMF(EP) MEDIAN RESULTS FOR (PN:D4 GTR:C4) NOTE PAIRING WITH MAX MEDIAN SIR PARAMETER VALUE COMBINATION

	CMF(EP) (PN:D4 GTR:C4)		
	MEDIAN [Q1, Q3]		
NOTE RANGE	[SDR]	[SIR]	[SAR]
	8.44[7.28, 9.07] dB	**14.02[11.80, 15.54] dB**	**9.74[9.38, 10.41] dB**
PN: D4	K: 2	K: 2	K: 2
	λ: 0.01	λ : 0.01	λ :0.01
GTR: C4	γ: 0.001	γ: 0.001	γ: 0.001
	N: 512	N: 512	N: 512
	SM: SYNTH	SM: SYNTH	SM: SYNTH

Table 6.3: CMF(OP) MEDIAN RESULTS FOR (PN:D4 GTR:C4) NOTE PAIRING WITH MAX MEDIAN SIR PARAMETER VALUE COMBINATION

	CMF(OP) (PN:D4 GTR:C4)		
	MEDIAN [Q1, Q3]		
NOTE RANGE	[SDR]	[SIR]	[SAR]
	12.70[11.59, 13.82] dB	**43.45[40.49, 46.40] dB**	**12.60[11.49, 13.72] dB**
PN: D4	K: 2	K: 2	K: 2
	λ: 0.01	λ : 0.01	λ :0.01
GTR: C4	γ: 0.001	γ: 0.001	γ: 0.001
	N: 512	N: 512	N: 512
	SM: SYNTH	SM: SYNTH	SM: SYNTH

Table 6.4: CMF(MP) MEDIAN RESULTS FOR (PN:D4 GTR:C4) NOTE PAIRING WITH MAX MEDIAN SIR PARAMETER VALUE COMBINATION

	CMF(MP) (PN:D4 GTR:C4)		
	MEDIAN [Q1, Q3]		
NOTE RANGE	[SDR]	[SIR]	[SAR]
	11.61[10.54, 12.68] dB	**26.83[24.42, 28.99] dB**	**11.65[10.64, 12.69] dB**
PN: D4	K: 2	K: 2	K: 2
	λ: 0.01	λ : 0.01	λ :0.01
GTR: C4	γ: 0.001	γ: 0.001	γ: 0.001
	$\hat{\sigma}$: 0.1	$\hat{\sigma}$: 0.1	$\hat{\sigma}$:0.1
	N: 512	N: 512	N: 512
	SM: SYNTH	SM: SYNTH	SM: SYNTH

A graphical summary of the results obtained over both sources, for the synthesis-based source estimation method, are presented in the form of box plots in Figures 6.4 - 6.6, which offer a simple visualization of the spread of the results obtained for each factorization-based separation procedure. From the box plots, it can be seen that the NMF-based and CMF(EP)-based performances are fairly similar for all three performance measures. Though the median of the performance measures obtained using the NMF-based separation procedure is higher than the median of the performance measures obtained using the CMF(EP)-based separation procedure in general, the spreads of the results obtained using the CMF(EP)-based separation procedure are larger. In a similar fashion to the results of the second experiment, it can be seen that the CMF(OP)-based separation procedure produced the highest performance results for all the measures. Lastly, it can be seen from the box plots of Figures 6.4 - 6.6 that the results obtained using the CMF(MP)-based separation procedure fell between the CMF(EP)-based and the CMF(OP)-based separation procedures for all three performance measures.

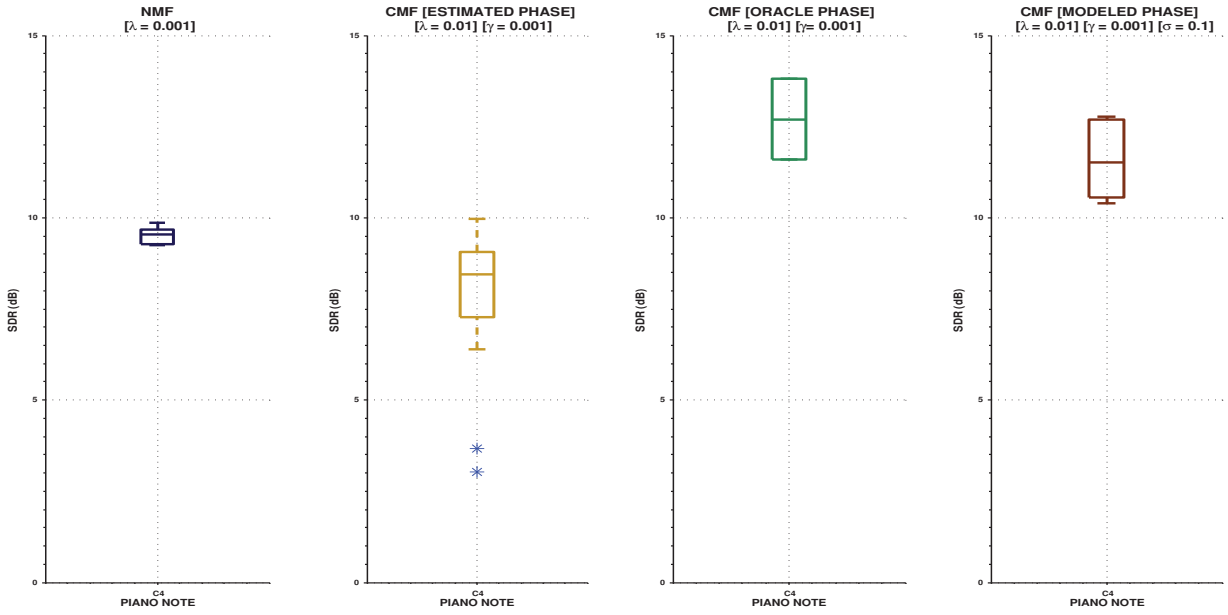

Fig. 6.4: COMBINED SDR RESULTS FOR BOTH SOURCES FOR THE NOTE PAIRING (PN:D4 GTR:C4) USING THE SYNTHESIS-BASED RECONSTRUCTION METHOD FOR NMF-BASED (BLUE), CMF(EP))-BASED (YELLOW), CMF(OP)-BASED (GREEN) AND CMF(MP)-BASED (RED) SEPARATION PROCEDURES. HIGHER VALUES ARE AN INDICATION OF BETTER PERFORMANCE

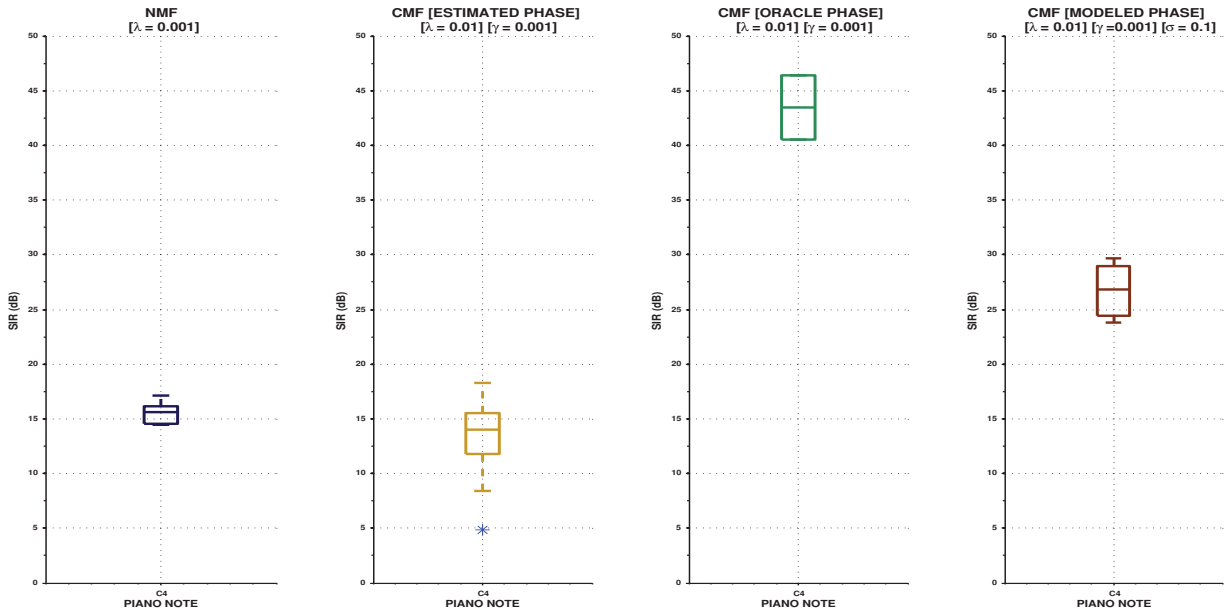

Fig. 6.5: COMBINED SIR RESULTS FOR BOTH SOURCES FOR THE NOTE PAIRING (PN:D4 GTR:C4) USING THE SYNTHESIS-BASED RECONSTRUCTION METHOD FOR NMF-BASED (BLUE), CMF(EP))-BASED (YELLOW), CMF(OP)-BASED (GREEN) AND CMF(MP)-BASED (RED) SEPARATION PROCEDURES. HIGHER VALUES ARE AN INDICATION OF BETTER PERFORMANCE

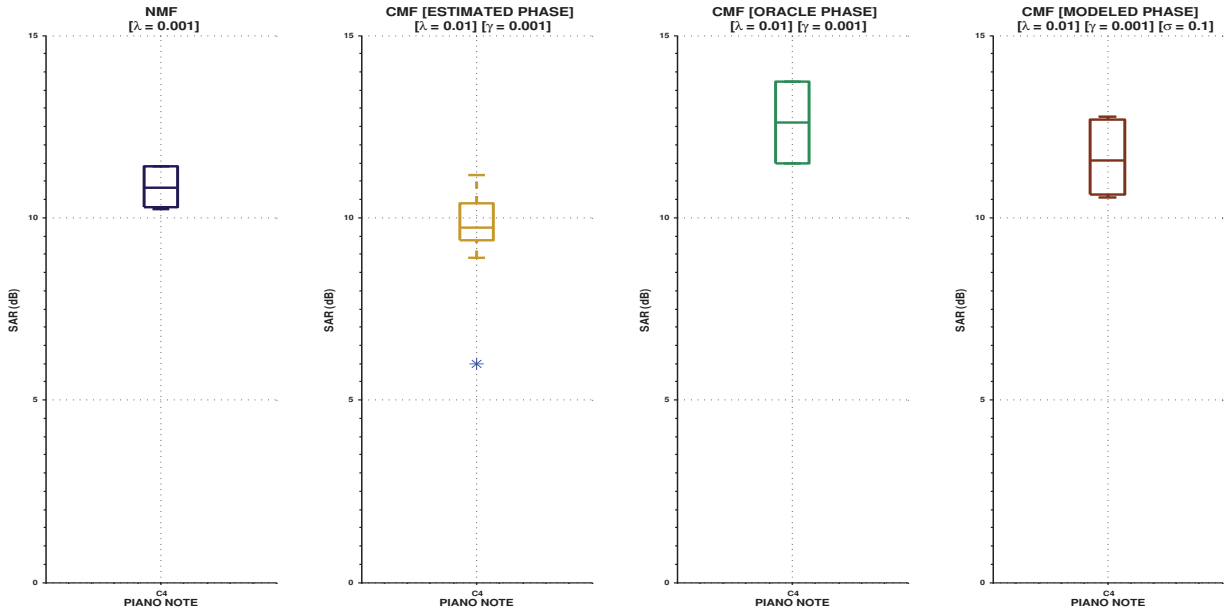

Fig. 6.6: COMBINED SAR RESULTS FOR BOTH SOURCES FOR THE NOTE PAIRING (PN:D4 GTR:C4) USING THE SYNTHESIS-BASED RECONSTRUCTION METHOD FOR NMF-BASED (BLUE), CMF(EP))-BASED (YELLOW), CMF(OP)-BASED (GREEN) AND CMF(MP)-BASED (RED) SEPARATION PROCEDURES. HIGHER VALUES ARE AN INDICATION OF BETTER PERFORMANCE

6.6.2 Instantaneous Frequency/Spectral Magnitude Analysis

Further analysis was conducted to investigate the behaviour of the CMF(MP) algorithm with respect to the estimated phase parameter, spectral templates, and activation weights, within the region corresponding to the amplitude variations associated with the overlapping first harmonic of the (PN:D4 GTR:C4) note pairing. Specifically, the instantaneous frequency estimates (defined below), and the spectral magnitude of the estimated sources, were plotted as a function of time for the frequency bin of $n = 13$, which corresponds to a centre frequency of $F_c(13) = 13 \times \frac{11025}{512} \approx 280$ Hz. The fundamental frequency of the PN:D4 note is roughly $f_{0_1} \approx 293$ Hz, whereas the fundamental frequency of the GTR:C4 note is roughly $f_{0_2} \approx 261$ Hz. As such, a high degree energy is found within the frequency bin of $n = 13$ for both sources.

The instantaneous frequency estimate, being a function of the phase relationship between bins, was used as a way of assessing the quality of the phase parameter estimate. As in (Dressler 2006), the instantaneous frequency estimate was calculated as follows:

Instantaneous Frequency Estimate:

$$\begin{aligned} F_{\text{INST}_{n,p}}(m) &= (F_c(n) + F_{\Delta_{n,p}}(m)) \\ F_{\Delta_{n,p}}(m) &= \frac{Fs}{2\pi H}\textsc{Princarg}\left([\phi]_{n,p,m} - [\phi]_{n,p,m-1} - \frac{2\pi nH}{N}\right) \end{aligned} \tag{6.16}$$

where the "principle argument" function, denoted Princarg, maps its argument into the range of $[-\pi, \pi]$. Instantaneous frequency estimates were calculated from the estimated/oracle/modelled phase parameter obtained using the CMF(EP)-based, CMF(OP)-based and the CMF(MP)-based separation procedures. To get a sense of how the phase evolution constraint was influencing the performance of the CMF(MP)-based separation procedure, the instantaneous frequency estimates were calculated for all levels of $\hat{\sigma} \in \{0, 0.001, 0.01, 0.1, 1\}$. The results of the instantaneous frequency estimates are presented below in Figure 6.7, for the piano source, and in Figure 6.8, for the guitar source. From the figures, it can be seen that, despite the relatively constant instantaneous frequency estimate of both sources, the phase interaction between the two sources produces a highly oscillatory instantaneous frequency estimate when mixed. With the phase evolution constraint set to $\hat{\sigma} = 0$, which corresponds to the CMF(EP)-based separation procedure, the phase parameters for the unmixed source estimates yield an instantaneous frequency estimate that behaves in a similar fashion to the instantaneous frequency estimate of the mixed sources. It appears that as $\hat{\sigma}$ increases, the instantaneous frequency estimates for the unmixed sources tends towards the behaviour of the instantaneous frequency estimates of the oracle phase conditions for each source. Recall that no timing information, in terms of note onsets or offsets was provided in the phase model described above in assumption A-3. This fact appears to be reflected in Figures 6.7 - 6.8, in which the spikes in instantaneous frequency of the oracle phase conditions at the location of note onsets and offsets have been smoothed out.

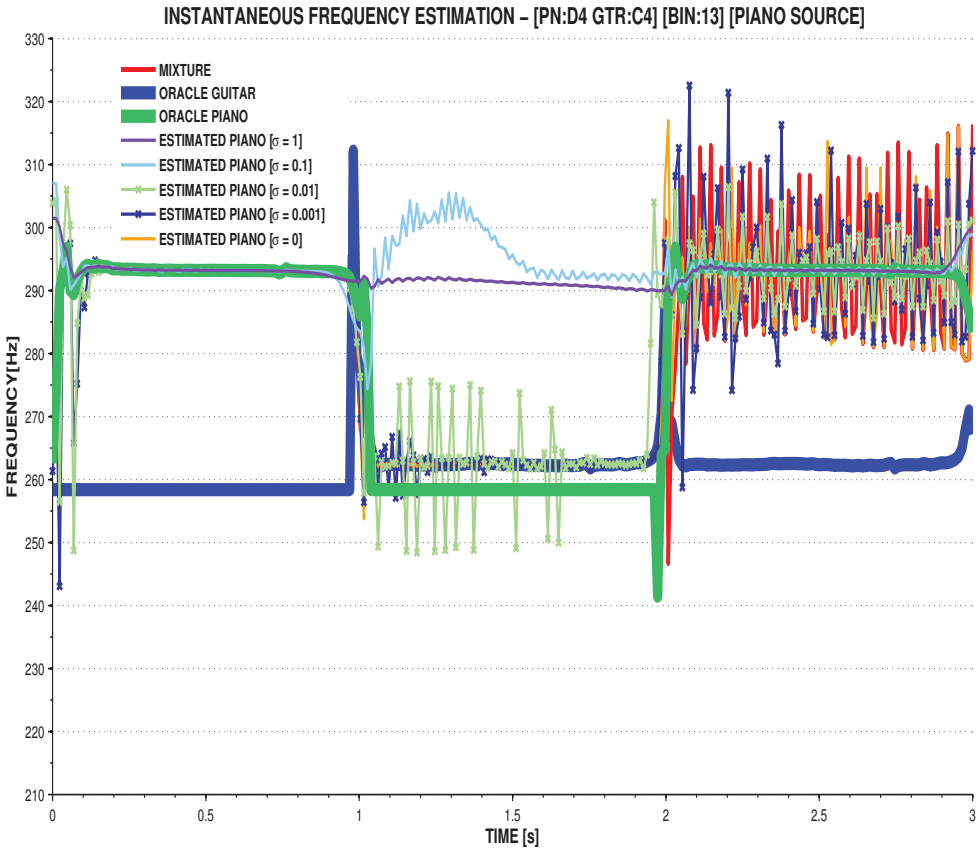

Fig. 6.7: INSTANTANEOUS FREQUENCY ESTIMATE ANALYSIS FOR THE PIANO SOURCE OF THE (PN:D4 GTR:C4) NOTE PAIRING AT THE 13^{th} FREQUENCY BIN (CORRESPONDING TO THE FIRST HARMONIC OF BOTH SOURCES)

Next, the behaviour of the spectral template and activation weight estimates for all four factorization-based separation procedures were compared through the analysis of the spectral magnitudes of the source estimates, as a function of time, for the $n = 13^{th}$ frequency bin. As previously stated, this frequency corresponds to the bin at which a high energy overlap occurs between the first harmonics of the (PN:D4 GTR:C4) note pairing. In was questioned whether the inclusion of the phase parameter, with the various degrees of refinement (EP, OP, MP), would noticeably change the shape of the magnitude spectrum of the estimated sources. If no change in shape occurred, this could possibly suggest that the phase parameter is not having a substantial effect on the estimated templates/activations, and that the incorporation of the phase refinement into the algorithm could occur independently from the templates/activations estimations. The plots of the magnitude spectrum of the source estimates can be found in Figure 6.9, for the piano source estimate, and

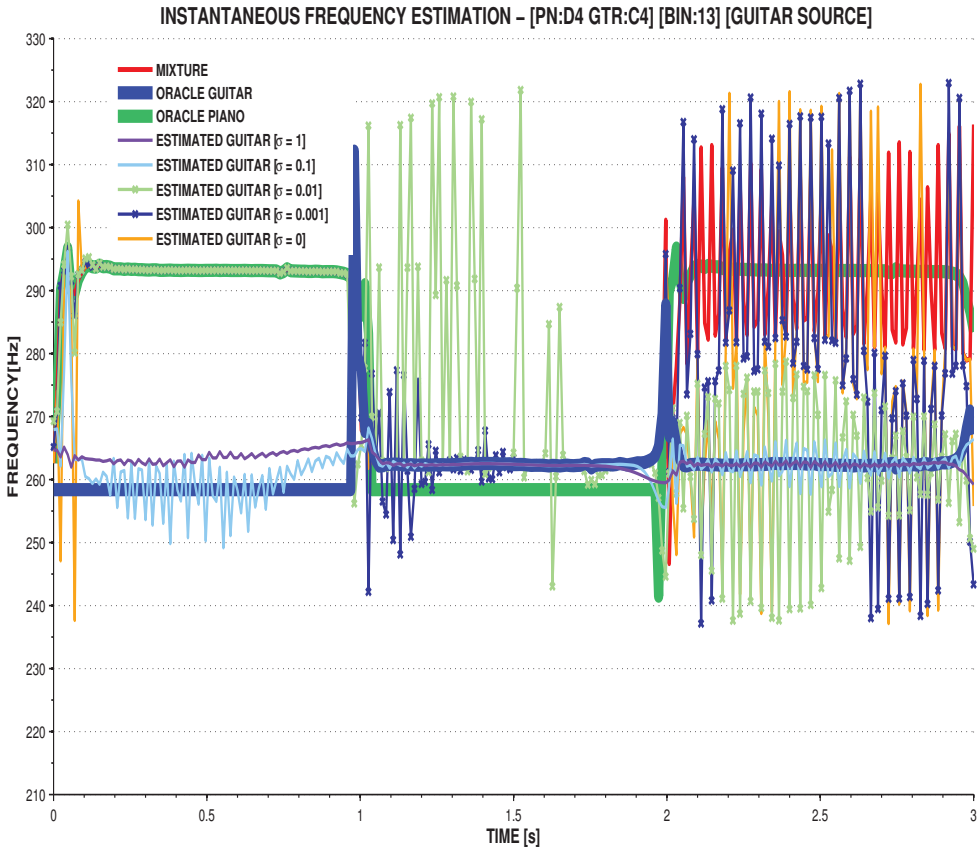

Fig. 6.8: INSTANTANEOUS FREQUENCY ESTIMATE ANALYSIS FOR THE GUITAR SOURCE OF THE (PN:D4 GTR:C4) NOTE PAIRING AT THE 13^{th} FREQUENCY BIN (CORRESPONDING TO THE FIRST HARMONIC OF BOTH SOURCES)

Figure 6.10, for the guitar source estimate. Both figures compare the spectral magnitude of the NMF-based, CMF(EP)-based, CMF(OP)-based, and CMF(MP)-based source estimates against the ground truth magnitude spectrum corresponding to the guitar source, piano source, and the mixture of the sources. Also included is the magnitude spectrum corresponding to the mixture that would occur if the first harmonic of both sources were in phase (for which the NMF mixture model assumption would hold true)[6], to get a better sense of how the magnitude of the mixture spectrum is changing due to the phase difference between the sources.

From the results presented in Figure 6.9 and Figure 6.10, the amplitude modulation

[6]Note that, in the case of the (PN:D4 GTR:C4) note pairing, it is not possible for first harmonic of both sources to be perfectly in phase over time, as both sources have a different fundamental frequency (i.e., they are changing phase at a different rate)

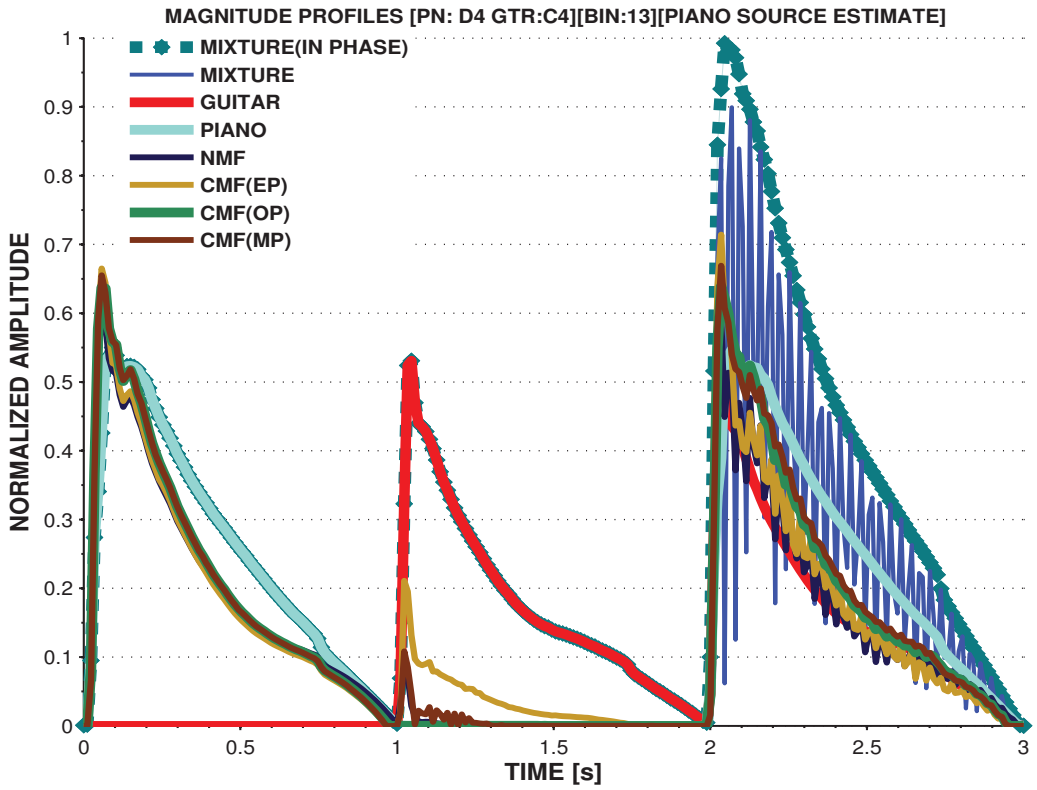

Fig. 6.9: MAGNITUDE PROFILE ANALYSIS FOR THE PIANO SOURCE OF THE (PN:D4 GTR:C4) NOTE PAIRING AT THE 13^{th} FREQUENCY BIN (CORRESPONDING TO THE FIRST HARMONIC OF BOTH SOURCES)

mentioned above can be observed in the magnitude spectrum for the mixture of the sources. Though both sources have a smoothly decaying spectral magnitude, the phase difference between the first harmonic of the two sources causes a variation in amplitude when mixed together. The magnitude spectrum of the source estimates obtained for all four factorization-based separation methods have roughly the same shape during the first two second of the mixture (for which the sources are played in isolation), though the magnitude corresponding to the spectral estimates obtained using the NMF-based and CMF(EP)-based separation procedures are slightly lower than the magnitude of the spectral estimates obtained using the CMF(OP)-based and the CMF(MP)-based separation procedures. During the last second of the mixture, the oscillating spectral magnitude of the mixture is apparent in the spectral magnitude of the source estimates obtained using all four separation procedures, however slightly more pronounced in the spectral estimates obtained using the NMF-based and CMF(EP)-based separation procedures. The spectral

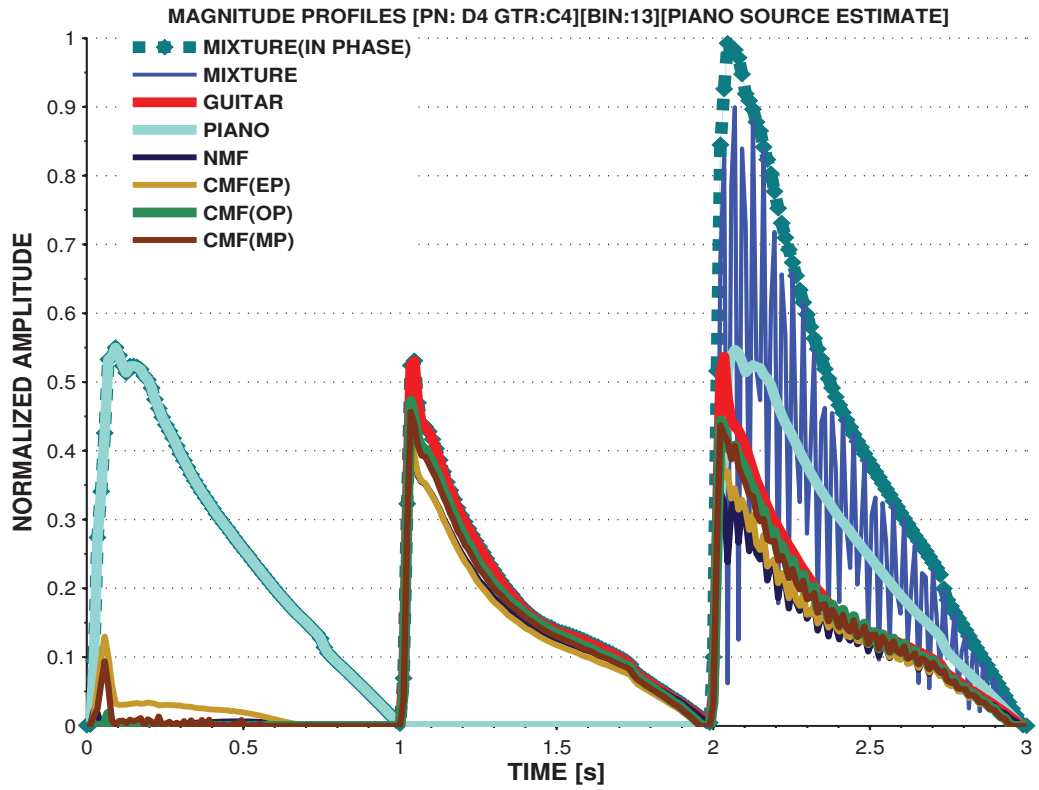

Fig. 6.10: MAGNITUDE PROFILE ANALYSIS FOR THE GUITAR SOURCE OF THE (PN:D4 GTR:C4) NOTE PAIRING AT THE 13^{th} FREQUENCY BIN (CORRESPONDING TO THE FIRST HARMONIC OF BOTH SOURCES)

magnitude of the CMF(OP)-based and CMF(MP)-based separation procedures are not as smooth as the spectral magnitude for the true guitar and piano sources, however, the results indicate an improvement both in terms of proximity to the spectral magnitude of the true sources and in terms of a reduction of the oscillating amplitude effect present over time. These results suggest that the inclusion of the phase parameter is helping to improve the estimation of the underlying source's spectral amplitudes, despite the strong violation of the NMF mixture model assumption.

Audible Differences in Performance

Audio playback of the estimated sources from all four factorization-based separation procedures confirmed a reduction in the oscillating amplitude effect using both the CMF(OP)-based and the CMF(MP)-based separation procedures. As previously stated, the spectral variability of both sources is not perfectly described using only $K = 2$ spectral templates,

one for each source. In fact, it was often observed that the transients of a given source were being described with the help of the spectral template corresponding to the other source. This fact is reflected in Figure 6.9, for example, in which a clear activation of the spectral template corresponding to the piano source is present at the onset of the guitar source, which occurs at $t = 1$ s. It was observed that the degree to which the sources were used to describe the transients of the other source varied between factorization algorithms. The CMF(EP)-based separation often produced a high degree of activation from a given source during the onset of the other source. On the other hand, it was observed that the spectral template discovered for a given source, using the CMF(OP)-based separation procedure, were not often used to describe the transient of the other source. This could suggests that the discovered templates for each source, using the CMF(OP)-based separation procedure were more distinct from one another or that the strong temporal information contained within the oracle phase (e.g., note onset/offset) could be aiding in finding a solution for which a given source is not used to describe the transients of the other source. Finally, it was observed that the CMF(MP)-based separation procedure did produce unwanted activations of a given source during the onset of the other source, however, to a lesser degree than the CMF(EP)-based separation procedure. These results were noticeable in the audio playback of the estimated sources from all four factorization-based separation procedures.

6.7 Conclusion

This chapter detailed the development of a phase evolution constraint which was incorporated into the CMF(EP) framework as a potential refinement of the phase parameter estimates. The newly proposed model, CMF(MP), was applied to the separation of the mixture consisting of the (PN:D4 GTR:C4) note pairing, as a proof of concept. It was observed that the CMF(MP)-based separation procedure offered improvements in the median SIR performance measure, while still retaining a relatively high median SDR, and SAR performance measures, as compared to the results obtained using the NMF-based and CMF(EP)-based separation procedures. Analysis of the instantaneous frequency estimates based on the phase parameter estimates obtained using the CMF(EP)-based, CMF(OP)-based and the CMF(MP)-based (with all 5 levels of $\hat{\sigma}$) separation procedures, was conducted, focusing specifically on the $n = 13^{th}$ frequency bin. The results obtain suggest that instantaneous frequency estimates using the phase parameter estimates obtained from

the CMF(EP)-based separation procedure, behaved in a similar fashion to the instantaneous frequency estimates using the phase parameter of the mixed sources. On the other hand, as $\hat{\sigma}$ increased, the behaviour of the instantaneous frequency estimates obtained using the CMF(MP)-based separation procedure were more akin to those obtained based on the oracle phase of the unmixed sources. Analysis was conducted into the behaviour of the magnitude spectrum of the estimated sources obtained from all four separation procedures. The results suggest that the phase refinement achieved using either the CMF(MP)-based and CMF(OP)-based separation procedures seems to offer improvements in the spectral template/activation weight estimates, both in terms of reduction of the oscillations in amplitude and in terms of the proximity to the true magnitude spectrum of the unmixed sources. The resolution of overlapping partials in musical sources is an important area of research, (Woodruff et al. 2008). It is acknowledged in (Li et al. 2009) that NMF is limited as a tool for resolving the true amplitudes of the underlying sources in regions of high-energy spectral overlap, for which consideration of the phase difference between overlapping sources is essential. The results presented in this chapter, however, suggest that the CMF-based separation procedures, under the two investigated phase refinements (oracle phase and modelled phase), could potentially be used as a tool for resolving the true amplitudes of the underlying sources in regions of high-energy spectral overlap.

Chapter 7

Conclusions and Future Work

7.1 Summary of Conclusions and Contributions

The main conclusions and contributions of this thesis can be summarized as follows. In Chapter 4, novel insight into the behaviour of NMF and CMF when applied as a tool for SC-MSS was gained through the analysis of highly controlled musical test mixtures. Parameter value combinations that resulted in the maximum/minimum median pooled performance measures were established, which may be used as reference for future CMF related studies. It was observed that the CMF(EP)-based separation procedure did not offer any obvious increase in performance over the NMF-based separation procedure for this particular dataset and maximal parameter value combinations.

In Chapter 5, the behaviour of the NMF-based and CMF(EP)-based separation procedures were observed in conditions for which the NMF mixture model assumptions were most violated, given the spectrogram analysis parameters considered. It was concluded that, even though the CMF(EP)-based separation procedure did not offer substantial improvements over the NMF-based separation procedure in these particular conditions, the superior results obtained using the CMF(OP)-based separation procedure provided insight into a possible upper bound of achievable separation performance under the CMF framework, for this particular dataset and parameter value combinations.

In Chapter 6, the development of the novel phase evolution constraint is presented. The CMF(MP)-based source separation procedure, armed with this newly proposed phase constraint, obtained promising results when tested using a highly controlled mixture involving the (PN:D4 GTR:C4) note pairing. It was observed that the phase/spectral magnitude

behaviour of the extracted sources using the CMF(MP)-based separation procedure more closely matched the phase/spectral magnitude behaviour of the extracted sources using the CMF(OP)-based separation procedure, which outperformed both the NMF-based and CMF(EP)-based separation procedures. It was also observed that the incorporation of phase refinement into the CMF framework appeared to have a positive effect on the spectral template/activation weights extracted, in terms of better matching the spectral magnitude of the unmixed sources. The CMF(MP)-based separation procedure, as it currently stands, offers promise as a tool for the separation of harmonic sources possessing strong spectral overlap within their highest energy partials. This task can not be previously approached using standard NMF-based separation techniques due to the strong violation of the NMF mixture model assumption.

7.2 Future Directions

Throughout the exploration of the CMF-based source separation procedure conducted in this thesis, several questions arose and further studies were conceived which would hopefully yield deeper insight into the behaviour of CMF as a tool for SC-MSS. In the third experiment, the newly proposed phase evolution constraint was only tested against the (PN:D4 GTR:C4) note pairing, as a proof of concept. More testing is required, using less confined test cases, to gain further insight into the true benefits/drawbacks of this constraint. An in-depth analysis of the variation in the shape of the learned spectral templates for various degrees of violation of the NMF mixture model assumption would also help give a better idea of the shortcomings/potential benefits of the CMF model with various degrees of phase refinement.

It was observed that the CMF(MP)-based separation procedure offered promising results when the phase evolution penalty weight, σ, was set to be equal to $\sigma = [W]_{n,p}[H]_{p,m}\hat{\sigma}$, where $\hat{\sigma} \in \{0, 0.001, 0.01, 0.1, 1\}$. However, setting σ in this manner was not theoretically justified due to the dependence on the template and activation weights, W, and H, which was not considered during the optimization of the phase evolution constraint cost function. Incorporating a template/activation weight dependent scaling factor into the phase evolution cost function would almost certainly require the development of an auxiliary function to aid in the optimization. Further research could investigate the development of a penalty function/auxiliary function of this sort. A thorough investigation of the conver-

gence behaviour of the CMF(MP) algorithm, in general, would also be beneficial in better understanding the nature of the results obtained.

It was suggested in Chapter 6 that possible sources of inaccuracies in the phase evolution model, specifically for the upper harmonics, could possibly be attributed to the fact that neither source is perfectly harmonic. Further extensions to this newly proposed phase evolution constraint could consider a refinement of the proposed sinusoidal model which accounts for inharmonic sources by more accurately specifying the location of each high energy partial other than assuming them to be located at integer multiples of the fundamental frequency. Better estimates for the fundamental frequency of each source would also help improve the modelled phase evolution of each source.

Throughout this thesis, the effect of two different source specific phase refinements were considered: oracle phase and the phase evolution constraint. By introducing source specific information into the phase parameter associated with a given component, a label is being assigned to that particular component from the very start, despite the random initialization of both the spectral templates and activation weights. The factorization results for both NMF and CMF are subject to a permutation indeterminacy of the extracted templates given a random initialization. Thus, it is possible that the component associated with a given source prior to the factorization does not remain associated to that source after the factorization is complete. In practice, the component assigned to a given source always remained associated with that source throughout the factorization using both the CMF(OP)-based and the CMF(MP)-based separation procedure for $\hat{\sigma} \geq 0.1$. These results may suggest that the source specific information incorporated into the phase matrix is aiding in preserving the component-to-source association throughout the factorization. This behaviour warrants further examination. Using the CMF(MP)-based separation procedure, an experiment could be conducted to gain insight into this association by examining the frequency with which the component-to-source grouping is preserved throughout the factorization as $\hat{\sigma}$ is varied.

The number of extracted components was set to $K = P = 2$ for all the experiments conducted throughout this thesis to ease the interpretability of the results obtained. As previously mentioned in Chapter 4, this allowed for the CMF algorithm to be treated in a form similar to CMFWISA, for which only one phase parameter is associated with a given source. Since the proposed phase constraint restricts the behaviour of the phase evolution of a source, and not of a given component in particular, this constraint could be easily

extendable to the CMFWISA framework. The CMFWISA structure would also facilitate an investigation into the performance of the CMF-based separation procedure for values of $K >> P$. As discussed in the conclusions of Chapter 4, this would allow for any potential benefits of CMF over NMF in describing the higher frequency content of the estimated sources, as observed in (King and Atlas 2011), to be explored. An in-depth investigation of CMFWISA as tool for SC-MSS, incorporating this newly proposed phase constraint, could be very beneficial to the signal processing community.

The CMF models still remains highly parameterized. The total number of independent parameters could be reduced by incorporating more source specific priors into the CMF framework. Specifically in the case of musical source separation, these source priors could come in the form of a temporal smoothness constraint, as developed in (Virtanen 2006), or a common amplitude modulation (between overlapped and non-overlapped harmonics) constraint, as detailed in (Woodruff et al. 2008), imposed on the activation matrix, H. Higher level temporal information in the form of note onsets and offsets and could also be incorporated into the initialization of both the activation matrix, H, and the phase parameter, Φ.

Lastly, the exploration conducted throughout this thesis of CMF, as a tool for SC-MSS, was approached form a blind initialization setting, without the use of any training on the spectral templates, W. As such, the musical test cases considered, consisting of the mixture of guitar and piano notes, were intentionally simplified to ease the interpretation of the CMF performance, which was deemed necessary in these early stages of exploration. Future studies, however, should consider the use of training to determine the spectral template estimates prior to the factorization procedure. This would also allow for a larger dataset of musical mixtures, with a higher degree of spectral variability, to be considered as test cases for the CMF-based separation procedure.

Appendix A

Appendix

A.1 Phase Evolution Cost Function Minimization

Consider phase evolution cost function defined as follows:

$$C(\theta)_{\text{PE}} = \sum_{n,p,r,m} \mathbb{1}_{\mathcal{N}_{p,r}} |\exp(i[\Phi]_{n,p,m}) - \exp(i[\Phi]_{n,p,m-1}) \exp(i2\pi f_{0_p} r \tilde{H} T)|^2 \tag{A.1}$$

The phase parameter update can be constructed by minimizing the associated costs which depend on the parameter $[\Phi]_{n,p,m}$. This includes, the auxiliary functions associated with the factorization (as defined in the first term in Equation (2.40)), the auxiliary function associated with the consistency constraint (as defined in Equation (3.4)) and lastly, the phase evolution cost function (as defined in Equation (A.1)). The entire auxiliary function together with the phase evolution cost function (omitting any terms which do not depend on $[\Phi]_{n,p,m}$) can be expressed as follows:

$$\begin{aligned} C(\theta, \bar{\theta})_{\phi} = & \sum_{n,p,m} \frac{|[\bar{\mathbb{X}}]_{n,p,m} - [W]_{n,p}[H]_{p,m} \exp(i[\Phi]_{n,p,m})|^2}{[B]_{n,p,m}} \\ & + \gamma \sum_{n,p,m} |[\bar{\mathbb{L}}]_{n,p,m} - [W]_{n,p}[H]_{p,m} \exp(i[\Phi]_{n,p,m})|^2 \\ & + \sigma \sum_{n,p,r,m} \mathbb{1}_{\mathcal{N}_{p,r}} |\exp(i[\Phi]_{n,p,m}) - \exp(i[\Phi]_{n,p,m-1}) \exp(i2\pi f_{0_p} r \tilde{H} T)|^2 \end{aligned} \tag{A.2}$$

where the subscript ϕ is used to indicate that the auxiliary function being optimized only corresponds to the terms that depend on the phase parameter, including the newly proposed phase evolution cost function. Expanding the square of the modulus of each summand, using $*$ to denote the conjugate of a complex number, simplifies the calculation of the derivative as follows:

$$
\begin{aligned}
C(\theta,\bar{\theta})_{\phi} = \sum_{n,p,m} \Bigg(\Big(& [\bar{\mathbb{X}}]_{n,p,m}[\bar{\mathbb{X}}^{*}]_{n,p,m} \\
& - [\bar{\mathbb{X}}]_{n,p,m}[W]_{n,p}[H]_{p,m}\exp(-i[\Phi]_{n,p,m}) \\
& - [\bar{\mathbb{X}}^{*}]_{n,p,m}[W]_{n,p}[H]_{p,m}\exp(i[\Phi]_{n,p,m}) \\
& + ([W]_{n,p}[H]_{p,m})^{2}\cancelto{1}{\exp(i[\Phi]_{n,p,m})\exp(-i[\Phi]_{n,p,m})}\Big)\frac{1}{[B]_{n,p,m}}\Bigg) \\
& + \gamma \sum_{n,p,m} \Big([\bar{\mathbb{L}}]_{n,p,m}[\bar{\mathbb{L}}^{*}]_{n,p,m} \\
& - [\bar{\mathbb{L}}]_{n,p,m}[W]_{n,p}[H]_{p,m}\exp(-i[\Phi]_{n,p,m}) \\
& - [\bar{\mathbb{L}}^{*}]_{n,p,m}[W]_{n,p}[H]_{p,m}\exp(i[\Phi]_{n,p,m}) \\
& + ([W]_{n,p}[H]_{p,m})^{2}\cancelto{1}{\exp(i[\Phi]_{n,p,m})\exp(-i[\Phi]_{n,p,m})}\Big) \\
& + \sigma \sum_{n,p,r,m} \Big(\mathbb{1}_{\mathcal{N}_{p,r}}\cancelto{1}{[\exp(i[\Phi]_{n,p,m})\exp(-i[\Phi]_{n,p,m})} \\
& - \exp(i[\Phi]_{n,p,m-1})\exp(i2\pi f_{0_p} r\tilde{H}T)\exp(-i[\Phi]_{n,p,m}) \\
& - \exp(-i[\Phi]_{n,p,m-1})\exp(-i2\pi f_{0_p} r\tilde{H}T)\exp(i[\Phi]_{n,p,m}) \\
& + \cancelto{1}{\exp(-i[\Phi]_{n,p,m-1})\exp(i[\Phi]_{n,p,m-1})}\cancelto{1}{\exp(-i2\pi f_{0_p} r\tilde{H}T)\exp(i2\pi f_{0_p} r\tilde{H}T)}]\Big)
\end{aligned}
\tag{A.3}
$$

Taking the Wirtinger derivative (Bouboulis 2010) w.r.t $[\Phi]_{\tilde{n},\tilde{p},\tilde{m}}$:

$$
\begin{aligned}
\frac{\partial C(\theta,\bar{\theta})_{\phi}}{\partial [\Phi]_{\tilde{n},\tilde{p},\tilde{m}}} &= \frac{\partial}{\partial [\Phi]_{\tilde{n},\tilde{p},\tilde{m}}} \sum_{n,p,m} \Bigg(\Bigg(- [\bar{\mathbb{X}}]_{n,p,m}[W]_{n,p}[H]_{p,m} \exp(-i[\Phi]_{n,p,m}) \\
&- [\bar{\mathbb{X}}^*]_{n,p,m}[W]_{n,p}[H]_{p,m} \exp(i[\Phi]_{n,p,m}) \Bigg) \frac{1}{[B]_{n,p,m}} \Bigg) \\
&+ \gamma \sum_{n,p,m} \Bigg(- [\bar{\mathbb{L}}]_{n,p,m}[W]_{n,p}[H]_{p,m} \exp(-i[\Phi]_{n,p,m}) \\
&- [\bar{\mathbb{L}}^*]_{n,p,m}[W]_{n,p}[H]_{p,m} \exp(i[\Phi]_{n,p,m}) \Bigg) \\
&+ \sigma \sum_{n,p,r,m} \Bigg(\mathbb{1}_{\mathcal{N}_{p,r}} [- \exp(i[\Phi]_{n,p,m-1}) \exp(i2\pi f_{0_p} r \tilde{H} T) \exp(-i[\Phi]_{n,p,m}) \\
&- \exp(-i[\Phi]_{n,p,m-1}) \exp(-i2\pi f_{0_p} r \tilde{H} T) \exp(i[\Phi]_{n,p,m})] \Bigg) \\
&= i \Bigg([\bar{\mathbb{X}}]_{\tilde{n},\tilde{p},\tilde{m}}[W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}} \exp(-i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}}) \\
&- [\bar{\mathbb{X}}^*]_{\tilde{n},\tilde{p},\tilde{m}}[W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}} \exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}}) \Bigg) \frac{1}{[B]_{\tilde{n},\tilde{p},\tilde{m}}} \\
&+ i\gamma \Bigg([\bar{\mathbb{L}}]_{\tilde{n},\tilde{p},\tilde{m}}[W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}} \exp(-i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}}) \\
&- [\bar{\mathbb{L}}^*]_{\tilde{n},\tilde{p},\tilde{m}}[W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}} \exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}}) \Bigg) \\
&+ i\sigma \sum_{r} \Bigg(\mathbb{1}_{\mathcal{N}_{\tilde{p},r}} [\exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}-1}) \exp(i2\pi f_{0_{\tilde{p}}} r \tilde{H} T) \exp(-i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}}) \\
&- \exp(-i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}-1}) \exp(-i2\pi f_{0_{\tilde{p}}} r \tilde{H} T) \exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}}) \\
&- \exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}}) \exp(i2\pi f_{0_{\tilde{p}}} r \tilde{H} T) \exp(-i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}+1}) \\
&+ \exp(-i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}}) \exp(-i2\pi f_{0_{\tilde{p}}} r \tilde{H} T) \exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}+1})] \Bigg)
\end{aligned}
\tag{A.4}
$$

Now, using the property: $i(AB^* - A^*B) = -2\Im(AB^*)$, for A, $B \in \mathbb{C}$, where A^* and B^* are the conjugate transposes of A and B, respectively:

$$
\begin{aligned}
\frac{\partial C(\theta,\bar{\theta})_{\phi}}{\partial_{[\Phi]_{\tilde{n},\tilde{p},\tilde{m}}}} &= -2\Im\Bigg(\Big[\bar{\mathbb{X}}\Big]_{\tilde{n},\tilde{p},\tilde{m}}[W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}}\exp(-i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}})\Bigg)\frac{1}{[B]_{\tilde{n},\tilde{p},\tilde{m}}} \\
&\quad - 2\gamma\Im\Bigg(\Big[\bar{\mathbb{L}}\Big]_{\tilde{n},\tilde{p},\tilde{m}}[W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}}\exp(-i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}}))\Bigg) \\
&\quad - 2\sigma\Im\Bigg(\sum_{r}\mathbb{1}_{\mathcal{N}_{p,r}}\Bigg(\exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}-1})\exp(i2\pi f_{0_{\tilde{p}}}r\tilde{H}T)\exp(-i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}}) \\
&\quad + \exp(-i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}})\exp(-i2\pi f_{0_{\tilde{p}}}r\tilde{H}T)\exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}+1})\Bigg)\Bigg) \\
&= -2\Im\Bigg[\Bigg(\Big[\bar{\mathbb{X}}\Big]_{\tilde{n},\tilde{p},\tilde{m}}[W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}} + \gamma\Big[\bar{\mathbb{L}}\Big]_{\tilde{n},\tilde{p},\tilde{m}}[W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}} \\
&\quad + \sigma\sum_{r}\mathbb{1}_{\mathcal{N}_{p,r}}\Bigg(\exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}-1})\exp(i2\pi f_{0_{\tilde{p}}}r\tilde{H}T) \\
&\quad + \exp(-i2\pi f_{0_{\tilde{p}}}r\tilde{H}T)\exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}+1})\Bigg)\Bigg)\exp(-i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}})\Bigg]
\end{aligned}
\tag{A.5}
$$

Setting the derivative to zero yields:

$$
\begin{aligned}
0 = \Im\Bigg[\Bigg(&\frac{\Big[\bar{\mathbb{X}}\Big]_{\tilde{n},\tilde{p},\tilde{m}}}{[B]_{\tilde{n},\tilde{p},\tilde{m}}}[W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}} + \gamma\Big[\bar{\mathbb{L}}\Big]_{\tilde{n},\tilde{p},\tilde{m}}[W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}} \\
&+ \sigma\sum_{r}\mathbb{1}_{\mathcal{N}_{p,r}}\Bigg(\exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}-1})\exp(i2\pi f_{0_{\tilde{p}}}r\tilde{H}T) \\
&+ \exp(-i2\pi f_{0_{\tilde{p}}}r\tilde{H}T)\exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}+1})\Bigg)\Bigg)\exp(-i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}})\Bigg]
\end{aligned}
\tag{A.6}
$$

Following the work of (Kameoka 2009), (Le Roux et al. 2009) and (King 2012), we construct the phase parameter update by taking $[\Phi]_{\tilde{n},\tilde{p},\tilde{m}}$ to be equal to the argument of the terms within the outside-most set of circle brackets of Equation (A.6) as follows:

$$
\begin{aligned}
[\Phi]^{\text{OPTIMUM}}_{\tilde{n},\tilde{p},\tilde{m}} = Arg\Bigg(& \frac{[\bar{\mathbb{X}}]_{\tilde{n},\tilde{p},\tilde{m}}}{[B]_{\tilde{n},\tilde{p},\tilde{m}}}[W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}} + \gamma[\bar{\mathbb{L}}]_{\tilde{n},\tilde{p},\tilde{m}}[W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}} \\
& + \sigma \sum_{r} \mathbb{1}_{\mathcal{N}_{p,r}} \Bigg(\exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}-1}) \exp(i2\pi f_{0_{\tilde{p}}} r\tilde{H}T) \\
& + \exp(-i2\pi f_{0_{\tilde{p}}} r\tilde{H}T) \exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}+1}) \Bigg) \Bigg)
\end{aligned} \tag{A.7}
$$

By doing so, the terms within the square brackets of Equation (A.6) become:

$$
\begin{aligned}
\frac{\partial C(\theta,\bar{\theta})_{\phi}}{\partial [\Phi]_{\tilde{n},\tilde{p},\tilde{m}}} = \Im \Bigg[\Bigg| & \Bigg(\frac{[\bar{\mathbb{X}}]_{\tilde{n},\tilde{p},\tilde{m}}}{[B]_{\tilde{n},\tilde{p},\tilde{m}}}[W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}} + \gamma[\bar{\mathbb{L}}]_{\tilde{n},\tilde{p},\tilde{m}}[W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}} \\
& + \sigma \sum_{r} \mathbb{1}_{\mathcal{N}_{p,r}} \Bigg(\exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}-1}) \exp(i2\pi f_{0_{\tilde{p}}} r\tilde{H}T) \\
& + \exp(-i2\pi f_{0_{\tilde{p}}} r\tilde{H}T) \exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}+1}) \Bigg) \Bigg) \Bigg| \cancel{\exp(i[\Phi]^{\text{OPTIMUM}}_{\tilde{n},\tilde{p},\tilde{m}}) \exp(-i[\Phi]^{\text{OPTIMUM}}_{\tilde{n},\tilde{p},\tilde{m}})} \Bigg]^{1} \\
= \Im \Bigg| & \Bigg(\frac{[\bar{\mathbb{X}}]_{\tilde{n},\tilde{p},\tilde{m}}}{[B]_{\tilde{n},\tilde{p},\tilde{m}}}[W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}} + \gamma[\bar{\mathbb{L}}]_{\tilde{n},\tilde{p},\tilde{m}}[W]_{\tilde{n},\tilde{p}}[H]_{\tilde{p},\tilde{m}} \Bigg) \\
& + \sigma \sum_{r} \mathbb{1}_{\mathcal{N}_{p,r}} \Bigg(\exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}-1}) \exp(i2\pi f_{0_{\tilde{p}}} r\tilde{H}T) \\
& + \exp(-i2\pi f_{0_{\tilde{p}}} r\tilde{H}T) \exp(i[\Phi]_{\tilde{n},\tilde{p},\tilde{m}+1}) \Bigg) \Bigg| \\
= 0 &
\end{aligned} \tag{A.8}
$$

Note that the phase parameter update could have also been taken to be $[\Phi]^{\text{OPTIMUM}}_{\tilde{n},\tilde{p},\tilde{m}} + \pi$, which would have also resulted in a derivative of zero. However, following the approach taken in (Kameoka 2009), (Le Roux et al. 2009) and (King 2012), the phase parameter update was implemented, as is defined in Equation (A.7), and observed to yield results suggestive of a local minimum stationary point in practice. Further mathematical analysis is required to uncover the true classification of the stationary point obtained.

Bibliography

Albright, R., J. Cox, D. Duling, A. N. Langville, and C. D. Meyer. 2006. Algorithms, initializations, and convergence for the nonnegative matrix factorization. Technical report, N. Carolina State University, North Carolina, USA.

Badeau, R. 2011. Gaussian modeling of mixtures of non-stationary signals in the time-frequency domain (HR-NMF). In *Proceedings of the IEEE Workshop on Applications of Signal Procesing to Audio and Acoustics (WASPAA 2011)*, New Paltz, New York, USA, 253–56.

Basu, A., I. R. Harris, N. L. Hjort, and M. C. Jones. 1998. Robust and efficient estimation by minimising a density power divergence. *Biometrika* 85 (3): 549–59.

Berry, M. W., M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons. 2007. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics and Data Analysis* 52 (1): 155–73.

Bertin, N., R. Badeau, and E. Vincent. 2009. Fast Bayesian NMF algorithms enforcing harmonicity and temporal continuity in polyphonic music transcription. In *Proceedings of the 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2009)*, New Paltz, New York, USA, 29–32.

Bertin, N., R. Badeau, and E. Vincent. 2010, March. 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): 538–49.

Bertsekas, D. P. 1999. *Nonlinear Programming*. Athena Scientific.

Bouboulis, P. 2010. Wirtingers Calculus in general Hilbert Spaces. *arXiv preprint arXiv:1005.5170*: 1–27.

Boutsidis, C., and E. Gallopoulos. 2008. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition* 41 (4): 1350–62.

Bregman, A. 1984. Auditory Scene Analysis. In *Proceedings of the 7th International Conference on Pattern Recognition*, Cambridge, Massachusetts, USA, 168–75.

Brokhorst, A. W. 2000. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica* 86 (1): 117–28.

Brown, G. J., and M. Cooke. 1994. Computational auditory scene analysis. *Computer Speech and Language* 8 (4): 297–336.

Brown, G. J., and D. Wang. 2005. Separation of speech by computational auditory scene analysis. In J. Benesty, S. Makino, and J. Chen (Eds.), *Speech Enhancement*, 371–402. Springer.

Burred, J. J. 2009. *From Sparse Models to Timbre Learning : New Methods for Musical Source Separation*. Ph.D, Berlin Technical Institue.

Cao, Y., P. B. Eggermont, and S. Terebey. 1999, January. Cross Burg entropy maximization and its application to ringing suppression in image reconstruction. *IEEE transactions on image processing* 8 (2): 286–92.

Casey, M. A., and A. Westner. 2000. Separation of mixed audio sources by independent subspace analysis. In *Proceedings of the International Computer Music Conference*, Berlin, Germany, 154–61.

Chen, T., T. Wang, Y. Kuo, and A. Su. 2012. Effective separation of low-pitch notes using NMF with non-power-of-2 short-time Fourier transforms. In *Proceedings of the International Conference on Digital Audio Effects (DAFx 2012)*, 1–6.

Chu, M., F. Diele, R. Plemmons, and S. Ragni. 2004. Optimality, computation, and interpretation of nonnegative matrix factorizations. *SIAM Journal on Matrix Analysis*: 1–18.

Cichocki, A., S. Cruces, and S. Amari. 2011, January. Generalized alpha-beta divergences and their application to robust nonnegative matrix Factorization. *Entropy* 13 (1): 134–70.

Cichocki, A., H. Lee, Y. Kim, and S. Choi. 2008. Non-negative matrix factorization with α-divergence. *Pattern Recognition Letters* 29 (9): 1433–40.

Cichocki, A., and R. Zdunek. 2007. Regularized alternating least squares algorithms for non-negative matrix/tensor factorization. In *Advances in Neural Networks (ISNN 2007)*, Nanjing, China, 793–802. Springer.

Cichocki, A., R. Zdunek, and S. Amari. 2006. Csiszárs divergences for non-negative matrix factorization: family of new algorithms. In *Independent Component Analysis and Blind Signal Separation*, 32–9. Springer.

Cichocki, A., R. Zdunek, A. H. Phan, and S.-I. Amari. 2009. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. Tokyo, Japan: John Wiley and Sons Ltd.

Daudet, L. 2012. Phase-based informed source separation of music. In *Proceedings of the International Conference on Digital Audio Effects (DAFx 2012)*, York, United Kingdom, 15–8.

Dessein, A., A. Cont, and G. Lemaitre. 2013. Real-time detection of overlapping sound events with non-negative matrix factorization. In *Matrix information geometry*, 341–71. Springer.

Donoho, D., and V. Stodden. 2003. When does non-negative matrix factorization give a correct decomposition into parts? In S. T. Schölkopf, L. Saul, and Bernhard (Eds.), *Advances in Neural Information Processing Systems 16 (NIPS 2003)*. Cambridge, Massachusetts, USA: MIT Press.

Dressler, K. 2006. Sinusoidal extraction using an efficient implementation of a multi-resolution FFT. In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx 2006)*, Montreal, Quebec, Canada, 247–52.

Duan, Z., Y. Zhang, C. Zhang, and Z. Shi. 2008. Unsupervised single-channel music source separation by average harmonic structure modeling. *IEEE Transactions on Audio, Speech, and Language Processing* 16 (4): 766–78.

Eggert, J., and E. Korner. 2004. Sparse coding and NMF. *Neural Networks* 2 (4): 2529–33.

Erdogan, H., and E. M. Grais. 2010, August. Semi-blind speech-music separation using sparsity and continuity priors. In *Proceedings of the International Conference on Pattern Recognition (ICPR 2010)*, Istanbul, Turkey, 4573–76.

Févotte, C. 2011. Itakura-saito nonnegative factorizations of the power spectrogram for music signal decomposition. In W. Wang (Ed.), *Machine Audition: Principles, Algorithms and Systems*, Chapter 11, 266–96. IGI Publishing.

Févotte, C., N. Bertin, and J.-L. Durrieu. 2009, March. Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. *Neural computation* 21 (3): 793–830.

Févotte, C., R. Gribonval, and E. Vincent. 2005. BSS_EVAL toolbox user guide. Technical report, IRISA, Rennes, France.

Févotte, C., and J. Idier. 2011. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation* 23 (9): 2421–56.

Finesso, L. 2006. Nonnegative matrix factorization and I-divergence alternating minimization. *Linear Algebra and its Applications* 2 (416): 270–87.

Finesso, L., and P. Spreij. 2004. Approximate nonnegative matrix factorization via alternating minimization. In *arXiv preprint math/0402229*, Leuven, Belgium, 1–10.

Fitzgerald, D., M. Cranitch, and E. Coyle. 2009. On the use of the beta divergence for musical source separation. In *Proceedings of Signals and Systems Conference (ISSC 2009), IET Irish*, Dublin, Ireland, 1–6.

Fletcher, H., E. Donnell Blackham, and R. Stratton. 1962. Quality of piano tones. *The Journal of the Acoustical Society of America* 34 (6): 749–61.

Frederic, J. 2008. *Examination of initialization techniques for nonnegative matrix factorization*. Ph. D. thesis, Georgia State University.

Gillis, N. 2011. *Nonnegative matix factorization complexity, algorithms and applications*. Ph. D. thesis, Université Catholique de Louvain.

Gillis, N. 2012. Sparse and Unique Nonnegative Matrix Factorization Through Data Preprocessing. *arXiv preprint arXiv:1204.2436*: 1–34.

Gillis, N., and F. Glineur. 2012. On the geometric interpretation of the nonnegative rank. *Linear Algebra and its Applications* 437 (11): 2685–2712.

Gonzalez, E. F., and Y. Zhang. 2005. Accelerating the Lee-Seung algorithm for non-negative matrix factorization. Technical report, Department of Computational and Applied Mathematics, Rice University, Houston, Texas.

Griffin, D., and J. Lim. 1984, April. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32 (2): 236–43.

Hennequin, R., R. Badeau, and B. David. 2011. NMF with time-frequency activations to model nonstationary audio events. *IEEE Transactions on Audio, Speech, and Language Processing* 19 (4): 744–53.

Hennequin, R., B. David, and R. Badeau. 2011. Beta-divergence as a subclass of Bregman divergence. *IEEE Signal Processing Letters* 18 (2): 83–6.

Ho, N.-D. 2008. *Nonnegative matrix factorization algorithms and applications*. Ph. D. thesis, École Polytechnique de Louvain.

Hoyer, P. O. 2004. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning* 5: 1457–69.

Hunter, D. R., and K. Lange. 2003. A Tutorial on MM Algorithms. *The American Statistician* 58 (1): 30–7.

Itoyama, K. 2011. *Source Separation of Musical Instrument Sounds in Polyphonic Musical Audio Signal and Its Application*. Ph. D. thesis, Kyoto University.

Jaiswal, R., E. Coyle, D. Fitzgerald, and S. Rickard. 2012. Shifted NMF with group sparsity for clustering NMF basis functions. In *Proceedings of the International Conference on Digital Audio Effects (DAFx 2012)*, York, United Kingdom, 17–21.

Jutten, C., and P. Comon. 2010. Introduction. In *Handbook of Blind Source Separation*, Chapter 1, 1–22. Access Online via Elsevier.

Kameoka, H. 2009, April. Complex NMF: a new sparse representation for acoustic signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, Taipei, Taiwan, 3437–40.

Kim, H., and H. Park. 2008. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications (SIMAX)* 30 (2): 713–30.

King, B. 2012. *New methods of complex matrix factorization for single-channel source separation and analysis*. Ph. D. thesis, University of Washington.

King, B., and L. Atlas. 2010. Single-channel source separation using simplified-training complex matrix factorization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2010)*, Dallas, Texas, USA, 4206–9.

King, B., and L. Atlas. 2011. Single-channel source separation using complex matrix factorization. *IEEE Transactions on Audio, Speech, and Language Processing* 19 (8): 2591–97.

King, B., and L. Atlas. 2012. Complex matrix factorization toolbox version 1.0 for Matlab.

King, B., C. Févotte, and P. Smaragdis. 2012. Optimal cost function and magnitude power for NMF-based speech separation and music interpolation. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, Satander, Spain, 1–6.

Kirchhoff, H., S. Dixon, A. Klapuri, and M. E. Road. 2012. Derivation of update equations for multiple-template shift-variant non-negative matrix deconvolution based on β-divergence. Technical report, Queen Mary University of London, London, United Kingdom.

Klingenberg, B., J. Curry, and A. Dougherty. 2009. Non-negative matrix factorization: Ill-posedness and a geometric algorithm. *Pattern Recognition* 42 (5): 918–28.

Kompass, R. 2007, March. A generalized divergence measure for nonnegative matrix factorization. *Neural computation* 19 (3): 780–91.

Langville, A. N., C. D. Meyer, and R. Albright. 2006. Initializations for the nonnegative matrix factorization. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, Pennsylvania, USA, 23–6.

Laurberg, H., M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen. 2008, January. Theorems on positive data: on the uniqueness of NMF. *Computational Intelligence and Neuroscience* 2008: 1–9.

Lawton, W. H., and E. A. Sylvestre. 1971, September. Self modeling curve resolution. *Technometrics* 13 (3): 617–33.

Le Roux, J. 2009. *Exploiting regularities in natural acoustical scenes for monaural audio signal estimation, decomposition, restoration and modification*. Ph. D. thesis, University of Tokyo.

Le Roux, J., K. Hirokazu, K. Kunio, O. Junki, and S. Shigeki. 2011. Signal Analyzer, Analytical Method, Program, and Recording Medium. *Patent. Nippon Telegraph Company. University of Tokyo*.

Le Roux, J., H. Kameoka, and E. Vincent. 2009. Complex NMF under spectrogram consistency constraints. In *Proceedings of the Acoustical Society of Japan Autumn Meeting*, Tokyo, Japan, 4–5.

Le Roux, J., E. Vincent, Y. Mizuno, H. Kameoka, N. Ono, and S. Sagayama. 2010. Consistent Wiener filtering: generalized time-frequency masking respecting spectrogram consistency. In *Proceedings of the 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2010)*, Tokyo, Japan, 89–96.

Lee, D., and H. Seung. 1999, October. Learning the parts of objects by non-negative matrix factorization. *Nature* 401 (675): 788–91.

Lee, D. D., M. Hill, and H. S. Seung. 2001. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems* 13: 556–62.

Leeuw, J. D. 1994. Block-relaxation algorithms in statistics. In H. H. Bock, M. M. Richter, and W. Lenski (Eds.), *Information Systems and Data Analysis*, 308–24. Springer.

Lefevre, A., F. Bach, and C. Févotte. 2011. Itakura-Saito nonnegative matrix factorization with group sparsity. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, 21–4.

Li, Y., J. Woodruff, and D. Wang. 2009. Monaural Musical Sound Separation Based on Pitch and Common Amplitude Modulation. *IEEE Transactions on Audio, Speech and Language Processing* 17 (7): 1361–71.

Lin, C. 2007a. On the convergence of multiplicative update algorithms for non-negative matrix factorization. *IEEE Transactions on Neural Networks* 18 (6): 1589–96.

Lin, C. 2007b, October. Projected gradient methods for nonnegative matrix factorization. *Neural computation* 19 (10): 2756–79.

Mysore, G. J., P. Smaragdis, and B. Raj. 2010. Non-negative hidden Markov modeling of audio with application to source separation. In *Latent Variable Analysis and Signal Separation*, 140–8. Springer.

Nakano, M., H. Kameoka, J. L. Roux, Y. Kitano, N. Ono, and S. Sagayama. 2010. Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010)*, Kittala, Finland, 283–8.

Nakano, M., Y. Kitano, N. Ono, and S. Sagayama. 2010. Monophonic instrument sound segregation by clustering NMF components based on basis similarity and gain disjointness. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR 2010)*, Utrecht, Netherlands, 375–80.

O'Grady, P. D., and B. A. Pearlmutter. 2007. Discovering Convolutive Speech Phones using Sparseness and Non-Negativity Constraints. In *Independent Component Analysis and Signal Separation*, 520–7. London, UK: Springer.

Opolko, F., and J. Wapnick. 1987. McGill University master samples. [Compact Disks].

Paatero, P., and U. Tappert. 1994. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* 5 (2): 111–26.

Parry, R., and I. Essa. 2007a. Incorporating phase information for source separation via spectrogram factorization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, Honolulu, Hawaii, USA, 661–4.

Parry, R., and I. Essa. 2007b. Phase-aware non-negative spectrogram factorization. In *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation (ICA 2007)*, Atlanta, Georgia, USA, 536–43. Springer.

Razaei, M., R. Boostani, and M. Rezaei. 2011. An efficient initialization method for nonegative matrix factorization. *Journal of Applied Sciences* 11 (2): 354–59.

Schmidt, M. N. 2008. *Single-channel speech separation using sparse non-negative matrix factorization*. Ph. D. thesis, Technical University of Denmark.

Schmidt, M. N., and M. Morup. 2006. Nonnegative matrix factor 2-D deconvolution for blind single channel source separation. In *Proceedings of the 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2006)*, 700–7. Charleston, South Carolina, USA: Springer.

Schmidt, M. N., and R. K. Olsson. 2006. Single-channel speech separation using sparse non-negative matrix factorization. In *Proceedings of International Conference on Spoken Language Processing (ICSLP 2006)*, Pittsburgh, Pennsylvania, USA, 2614–17.

Shashanka, M. 2003. *Latent variable framework for modeling and sparating single-channel acoustic sources*. Ph. D. thesis, Boston University.

Siamantas, G. 2009. *An Iterative, Residual-Based Approach to Unsupervised Musical Source Separation in Single-Channel Mixtures*. Ph. D. thesis, University of York.

Slaney, M., D. Naar, and R. Lyon. 1994. Auditory model inversion for sound separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 1994)*, Adelaide, South Australia, Australia, 77–80.

Smaragdis, P. 2004. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In C. Puntonet and A. Prieto (Eds.), *International Symposium on Independent Component Analysis and Blind Signal Separation*, 494–99. Springer Berlin Heidelberg.

Smaragdis, P. 2007, January. Convolutive speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech and Language Processing* 15 (1): 1–12.

Smaragdis, P., and J. C. Brown. 2003. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, 177–80.

Smaragdis, P., M. Shashanka, and B. Raj. 2009. A sparse non-parametric approach for single channel separation of known sounds. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Eds.), *Proceedings of Advances Neural Information Processing Systems (NIPS 2009)*, Vancouver, Canada, 1705–13.

Vavasis, S. A. 2009. On the complexity of nonnegative matrix factorization Nonnegative matrix factorization. *SIAM Journal on Optimization* 20 (2): 1364–77.

Vincent, E., R. Gribonval, and C. Févotte. 2006. Performance measurement in blind audio source separation. *IEEE Transactions on Speech and Audio Processing* 14 (4): 1462–69.

Virtanen, T. 2004. Sound source separation using sparse coding with temporal continuity objective. In *Proceedings of the ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, Jeju Island, Korea, 231–3.

Virtanen, T. 2006. *Sound Source Separation in Monaural Music Signals*. Ph. D. thesis, Tampere University of Technology.

Wang, B., and M. Plumbley. 2005. Musical audio stream separation by non-negative matrix factorization. In *Proceedings of the UK Digital Music Research Network (DMRN 2005) Summer Conference*, Glasglow, United Kingdom, 23–4.

Wang, B., and M. D. Plumbley. 2006. Investigating single-channel audio source separation methods based on non-negative matrix factorization. In *Proceedings of the ICA Research Network International Workshop*, Liverpool, United Kingdom, 17–20.

Wang, Y., and Y. Zhang. 2012. Non-negative matrix factorization: a comprehensive review. *IEEE Transactions on Knowledge and Data Engineering* 25 (6): 1336–1353.

Weiss, R. J., D. P. W. Ellis, and E. Eng. 2006. Estimating single-channel source separation masks: Relevance vector machine classifiers vs. pitch-based masking. In *Proceedings of the Workshop on Statistical Perceptual Audition (SAPA 2006)*, Pittsburgh, Pennsylvania, USA, 31–6.

Wild, S. 2003. *Seeding non-negative matix factorization with the spherical k-means clustering*. Ph. D. thesis, University of Colorado.

Woodruff, J., Y. Li, and D. Wang. 2008. Resolving overlapping harmonics for monaural musical sound separation using pitch and common amplitude modulation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR 2008)*, Pittsburgh, Pennsylvania, USA, 538–43.

Yoshii, K., and M. Goto. 2012. Infinite composite autoregressive models for music signal analysis. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR 2012)*, Porto, Portugal, 79–84.

Zhao, L. 2008. Facial expression recognition based on PCA and NMF. In *7th World Congress on Intelligent Control and Automation (WCICA 2008)*, Chongqing, China, 6826–29.

Zheng, Z., J. Yang, and Y. Zhu. 2007, February. Initialization enhancer for non-negative matrix factorization. *Engineering Applications of Artificial Intelligence* 20 (1): 101–10.