Audio Source Separation for Music in Low-latency and High-latency Scenarios

Ricard Marxer Piñón

TESI DOCTORAL UPF / 2013

Directors de la tesi

Dr. Xavier Serra i Dr. Jordi Janer
Department of Information and Communication Technologies

This dissertation is dedicated to my family and loved ones.

Acknowledgments

I wish to thank the Music Technology Group (MTG) at the Universitat Pompeu Fabra (UPF) for creating such a great environment in which to work on this research. I especially want to thank my supervisors Xavier Serra and Jordi Janer for giving me this opportunity and their support during the whole process. I also want to express my deepest gratitude to the Yamaha Corporation and their Monet research team formed by Keita Arimoto, Sean Hashimoto, Kazunobu Kondo, Yu Takahashi and Yasuyuki Umeyama, without whom this work would not have been possible. I also want to thank MTG's signal processing team composed of Merlijn Blaauw, Jordi Bonada, Graham Coleman, Saso Musevic and Marti Umbert with whom we had many fruitful discussions that became the seeds of the research conducted here.

Another special thanks goes to the Music Cognition Group at the ILLC / University of Amsterdam with Henkjan Honing, Leigh Smith and Olivia Ladinig who invited me to stay and do research with them for a while. There I learned a lot about how we humans perceive music and how this can be taken into account when processing these types of signals.

Many thanks to all the researchers with whom I have collaborated, discussed and shared great moments over these years. Some of these are Eduard Aylon, Andreas Beisler, Dmitry Bogdanov, Òscar Celma, Maarten de Boer, Juan José Bosch, Ferdinand Fuhrmann, Jordi Funollet, Cristina Garrido, Emilia Gómez, Perfecto Herrera, Piotr Holonowicz, Enric Guaus, Martín Haro, Cyril Laurier, Oscar Mayor, Owen Meyers, Waldo Nogueira, Hendrik Purwins, Gerard Roma, Justin Salamon, Joan Serrà, Mohamed Sordo, Roberto Toscano, Nicolas Wack and José R. Zapata. I'm sorry if I have

forgotten anyone of you who have aided me so much.

Many thanks to my father, Jack Marxer, for his valuable feedback, proofreading and corrections of this document.

Finally I would like to thank my family and Mathilde for being there when I most needed them.

Abstract

Source separation in digital signal processing consists in finding the original signals that were mixed together into a set of observed signals. Solutions to this problem have been extensively studied for musical signals, however their application to real-world practical situations remains infrequent. There are two main obstacles for their widespread adoption depending on the scenario. The main limitation in some cases is high latency and computational cost. In other cases the quality of the results is insufficient. Much work has gone toward improving the quality of music separation under general conditions. But few studies have been devoted to the development of low-latency and low computational cost separation of monaural signals, as well as to the separation quality of specific instruments.

We propose specific methods to address these issues in each of these scenarios independently. First, we focus on methods with low computational cost and low latency. We propose the use of Tikhonov regularization as a method for spectrum decomposition in the low-latency context. We compare it to existing techniques in pitch estimation and tracking tasks, crucial steps in many separation methods. We then use the proposed spectrum decomposition method in low-latency separation tasks targeting singing voice, bass and drums. Second, we propose several high-latency methods that improve the separation of singing voice by modeling components that are often not accounted for, such as breathiness and consonants. Finally, we explore using temporal correlations and human annotations to enhance the separation of drums and complex polyphonic music signals.

Resum

En el camp del tractament digital de senyal, la separació de fonts consisteix en l'obtenció dels senyals originals que trobem barrejats en un conjunt de senyals observats. Les solucions a aquest problema s'han estudiat àmpliament per a senyals musicals. Hi ha però dues limitacions principals per a la seva aplicació generalitzada. En alguns casos l'alta latència i cost computacional del mètode és l'obstacle principal. En un segon escenari, la qualitat dels resultats és insuficient. Gran part de la recerca s'ha enfocat a la millora de qualitat de separació de la música en condicions generals, però pocs estudis s'han centrat en el desenvolupament de tècniques de baixa latència i baix cost computacional de mescles monoaurals, així com en la qualitat de separació de instruments específics.

Aquesta tesi proposa mètodes per tractar aquests temes en cadascun dels dos casos de forma independent. En primer lloc, ens centrem en els mètodes amb un baix cost computacional i baixa latència. Proposem l'ús de la regularització de Tikhonov com a mètode de descomposició de l'espectre en el context de baixa latència. El comparem amb les tècniques existents en tasques d'estimació i seguiment dels tons, que són passos crucials en molts mètodes de separació. A continuació utilitzem i avaluem el mètode de descomposició de l'espectre en tasques de separació de veu cantada, baix i percussió. En segon lloc, proposem diversos mètodes d'alta latència que milloren la separació de la veu cantada, gràcies al modelatge de components específics, com la respiració i les consonants. Finalment, explorem l'ús de correlacions temporals i anotacions manuals per millorar la separació dels instruments de percussió i dels senyals musicals polifònics complexes.

Resumen

En el campo del tratamiento digital de la señal, la separación de fuentes consiste en la obtención de las señales originales que han sido mezcladas en un conjunto de señales observadas. Las soluciones a este problema se han estudiado ampliamente para señales musicales. Hay dos limitaciones principales para su adopción generalizada. En algunos casos la alta latencia y coste computacional es el mayor obstáculo. En un segundo escenario, la calidad de los resultados es insuficiente. Gran parte de la investigación se ha enfocado en la mejora de la calidad de separación de la música en condiciones generales, pero pocos estudios se han centrado en el desarrollo de técnicas de baja latencia y bajo coste computacional de mezclas monoaurales, así como en la calidad de separación de instrumentos específicos.

Esta tesis propone métodos para tratar estos temas en cada uno de los casos de forma independiente. En primer lugar, nos centramos en los métodos con un bajo coste computacional y baja latencia. Proponemos el uso de la regularización de Tikhonov como método de descomposición del espectro en el contexto de baja latencia. Lo comparamos con las técnicas existentes en tareas de estimación y seguimiento de los tonos, que son pasos cruciales en muchos métodos de separación. A continuación utilizamos y evaluamos el método de descomposición del espectro en tareas de separación de voz cantada, bajo y percusión. En segundo lugar, proponemos varios métodos de alta latencia que mejoran la separación de la voz cantada, gracias al modelado de componentes que a menudo no se toman en cuenta, como la respiración y las consonantes. Finalmente, exploramos el uso de correlaciones temporales y anotaciones manuales para mejorar la separación de los instrumentos de percusión y señales musicales polifónicas complejas.

Résumé

Dans le domaine du traitement du signal, la séparation de source consiste à obtenir les signaux originaux qui ont été mélangés dans un ensemble de signaux observées. Les solutions à ce problème ont largement été étudiés pour l'application à la musique. Il existe deux principales limitations. La première est une latence et un coût de calcul élevés. La seconde est une qualité insuffisante des résultats. Une grande partie de la recherche actuelle s'est concentrée à améliorer, dans le cas générale, la qualité de séparation de la musique. Peu d'études ont été consacrées à minimiser la latence et le coût de calcul des techniques, ainsi qu'à la qualité de séparation d'instruments spécifiques.

Cette thèse propose des méthodes pour aborder ces deux questions indépendamment. Dans un premier temps, nous nous concentrons sur les méthodes à faibles coût de calcul et de latence. Pour cela, nous proposons d'utiliser la régularisation de Tikhonov en tant que méthode de décomposition spectrale. Nous la comparons à des techniques existantes dans le cadre de l'estimation et de suivi des tons, qui sont des étapes cruciales pour de nombreux procédés de séparation. Nous avons aussi utilisé et évalué la méthode de décomposition spectrale pour la séparation de voix chantée, de basse et de percussion. Dans un second temps, nous proposons des méthodes à latence élevée. La modélisation des composants qui ne sont souvent pas pris en compte, comme la respiration et les consonnes, nous permet d'améliorer la séparation de voix chantée. Nous explorons l'utilisation de corrélations temporelles et annotations manuelles pour améliorer la séparation des instruments de percussion et les signaux musicaux polyphoniques complexes.

Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

Abbreviation	Description
AASP	Audio and Acoustic Signal Processing
AC	Autocorrelation
ACF	Autocorrelation Function
APS	Artifact-related Perceptual Score
AQO	Audio Quality Oriented
ASR	Automatic Speech Recongnition
BASS	Blind Audio Source Separation
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
CHIME	Computational Hearing in Multisource Environments
COG	Center of gravity
DCT	Discrete Cosine transform
DFT	Discrete Fourier transform
DUET	Degenerate Unmixing Estimation technique
EEG	Electroencephalography
EM	Expectation-Maximization
EMD	Empirical Mode Decomposition
ERB	Equivalent rectangular bandwidth
FASST	Flexible Audio Source Separation Toolbox
FFT	Fast Fourier transform

Abbreviation	Description
FIR	Finite Impulse Response
FP	False Positive
GCC	Generalized Cross-Correlation
GMM	Gaussian Mixture Model
HHT	Hilbert–Huang transform
HMM	Hidden Markov Model
HPSS	Harmonic-Percussive Sound Separation
HSA	Hilbert Spectral Analysis
HSE	Harmonic Spectral Envelope
ICA	Independent Component Analysis
IDFT	Inverse Discrete Fourier Transform
IF	Instantaneous Frequency
IIR	Infinite Impulse Response
IMF	Intrinsic Mode Function
IMM	Instantaneous Mixture Model
IPD	Inter-channel Phase Difference
IPS	Interference-related Perceptual Score
IS	Itakura-Saito
ISTFT	Inverse Short-time Fourier transform
KL	Kullback-Leibler
LSM	Least Squares Means
LVA	Latent Variable Analysis
MAP	Maximum a posteriori
MFCC	Mel-frequency cepstrum
MIR	Music information retrieval
MIREX	Music information retrieval evaluation exchange
ML	Maximum Likelihood
MP	Matching Pursuit
MUSIC	Multiple Signal Classification
MVBF	Minimum Variance Beamforming
NMF	Nonnegative Matrix Factroization
NNMA	Nonnegative Matrix Approximation
OMP	Orthogonal Matching Pursuit
OPS	Overal-related Perceptual Score
PEASS	Perceptual Evaluation methods for Audio Source Separation
PLCA	Probabilistic Latent Component Analysis
PLSI	Probabilistic Latent Semantic Indexing
REPET	Repeating Pattern Extraction Technique
RWC	Real World Computing (music database)

Abbreviation	Description
SAR	Signal-to-Artifact Ratio
SDR	Signal-to-Distortion Ratio
SIMM	Smoothed Instantaneous Mixture Model
SIR	Signal-to-Interference Ratio
SNMF	Sparse Nonnegative Matrix Factorization
SNR	Signal-to-Noise Ratio
SO	Significance Oriented
STFT	Short-time Fourier transform
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TCG	Temporal Center of Gravity
TDOA	Time Difference of Arrival
TF	Time-Frequency
TN	True Negative
TP	True Positive
TPS	Target-related Perceptual Score
TR	Tikhonov Regularization

Mathematical symbols

General

Example	Symbol type	Description
$\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$	Bold uppercase letters	Matrices, bidimensional arrays.
A, B, C	Uppercase letters	Single numbers: constants, fixed values, etc.
$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$	Bold lowercase letters	Vectors, unidimensional arrays.
a, b, c	Lowercase letters	Single numbers: indices, variables, etc.

Specific

Symbol	Description
$\boldsymbol{X}$	STFT magnitude of signal x
$\widetilde{\boldsymbol{X}}$	STFT of signal x
ω	Frequency index

Symbol	Description
N_ω	Number of frequency bins
n	Sample index
N_T	Number of samples
t	Time index
N_T	Temporal length
w	Basis index
N_W	Number of basis
$\boldsymbol{W}$	Basis matrix
$\boldsymbol{H}$	Gains matrix
$v(t)$	Mixture signal
o	Mixture index
N_O	Number of mixtures
$x(t)$	Source signal
m	Source index
N_M	Number of sources
j	Candidate index
N_J	Number of candidates
p	Peak index
N_P	Number of peaks
l	Pitch index
N_L	Number of pitches
i	Band index
N_I	Number of bands
$\mathcal{P}(X)$	Probability of the random variable X
$\boldsymbol{B}$	Noise matrix
$\boldsymbol{m}$	Time-frequency mask
$\boldsymbol{L}$	Pitch likelihood
$\hat{\boldsymbol{X}}$	Estimation of $\boldsymbol{X}$
$\boldsymbol{X}^\top$	Transpose of $\boldsymbol{X}$
$\boldsymbol{X} \otimes \boldsymbol{Y}$	Elementwise multiplication between $\boldsymbol{X}$ and $\boldsymbol{Y}$
$\boldsymbol{X}/\boldsymbol{Y}$	Elementwise division between $\boldsymbol{X}$ and $\boldsymbol{Y}$

Part I

Chapter 1

Introduction

During the past few decades we have witnessed exponential growth in computing power, motivating the development of digital audio systems and new audio processing techniques and algorithms. This technological development provided professional audio engineers with new tools that go beyond their counterparts in the traditional analog domain such as filters or compressors. There has also been an enormous decrease in the cost of computational devices which translates into a democratization of these tools. This is giving birth to a community of audio enthusiasts who search for a more intuitive interaction with digital audio. The roles and the relations between musicians and audience are changing, leading to proactive listening habits in contrast to the traditional passive music-listening experience. These circumstances have led to a significant amount of research in high-level representations, descriptions, generation and manipulation of audio signals, and more specifically, of music.

The field of psychoacoustics has focused on understanding and mimicking the human auditory system. Bregman (1990) presented a series of psychoacoustical studies that set the basis for understanding human listening capabilities in sound segregation. Wang (1998) introduced the concept of Computational Auditory Scene Analysis (CASA), systems designed to separate mixtures of sound sources in the same way that human listeners do. Research in CASA systems aims to explain how the human auditory system works and how the human brain "makes sense" of the binaural audio stream.

Research in the domain of mathematics and statistics has been targeting the separation or isolation of signals, developing a field known as Blind Signal

Separation or Blind Source Separation (BSS). Jutten and Herault (1991) and Comon (1994) were the pioneer contributors with the proposal of Independent Component Analysis (ICA), which allowed separating signals based on two assumptions: statistical independence and not being Gaussian. Since then there have been multiple methods with the same goal, but focusing on specific types of signals such as audio and music. Recently the focus has shifted towards factorization techniques like Non-negative Matrix Factorization (NMF) (Smaragdis and Brown, 2003) that exploit the non-negativity of the factors.

In recent years we have seen enormous growth in the number of people using mobile devices. Mobile platforms impose constraints such as low computational power and limited memory on digital audio processing methods. Furthermore new forms of human-computer interaction such as touch screens, depth cameras and others are promoting the use of manual intervention in the source separation process. Finally, recent developments in the world of robotics are motivating the use of source separation methods in embedded computing. These facts translate into a growing interest in low-latency and realtime source separation methods. While these types of methods have been extensively researched in communications signals, they have only rarely targeted music signals (Vinyes et al., 2006; Ono et al., 2008a).

Low-latency source separation methods proposed until now were usually based on multiple sensors. In robotics and communications this is feasible because the capture process can be controlled by adding more microphones. Music mixtures, on the other hand, are created by sound engineers and producers, and the consumer typically does not have access to or information about the mixing process. However a large number of music mixtures are available in stereo format and therefore the separation process can often access two different channels. This has led the research in low-latency realtime music separation to focus on multiple-sensor scenarios. These music separation methods have limited scope since the assumptions about the mixing processes and the requirement of multiple channels are not always valid. The sources in music mixtures are not always panned differently in the stereo image and there is still a significant amount of music in mono format, where only a single channel is available.

On the other hand, the increasing computational power of desktop computers and their presence in many households is motivating the use of computationally expensive source separation methods in consumer products. In this scenario the main limitation is the quality of the results. Recent

methods for music signals have significantly improved separation quality and accuracy over previous techniques. However current state-of-the-art methods are still unacceptable for many commercial uses. By narrowing the scope of the task and the range of target instruments, one may exploit further assumptions or previous knowledge and improve the quality of existing high-latency, computationally expensive techniques.

1.1 Applications

Advancements in computational models of human sound segregation have seen many applications over the years. Audio compression and hearing aid systems have benefited from findings in the field of psychoacoustics and from systems such as CASA. Blind source separation methods have been applied in many diverse fields, ranging from medical imaging and computer vision to speech enhancement.

In this work we concentrate on blind source separation methods for audio signals or Blind Audio Source Separation (BASS). Working towards the creation of a BASS evaluation framework, Vincent et al. (2003) presented a typology of the different BASS tasks. This study proposed a distinction between two main groups of applications. Audio Quality Oriented (AQO) tasks are those whose output consists of a set of extracted sources intended for listening. Significance Oriented (SO) tasks, on the other hand, use the extracted sources or mixing parameters to obtain other information at an abstract level.

We present several applications of blind audio source separation methods that serve as motivation to work on this problem. These possible applications are also interesting for defining future lines of work and further delimiting the constraints under which the problem is framed.

We first briefly present applications for general audio signals. We then describe separation applications for music signals, and finally discuss a series of tasks related to the specific context of low-latency and realtime scenarios.

Audio source separation

Source separation applied to audio signals was first used in the field of communications. One of the most popular problems in this domain is known as the *cocktail party effect*. Cherry (1953) describes it as the ability of a human to follow the conversation of a single speaker in a highly noisy en-

vironment. The noisy environment may include other talkers, background noise or music, and humans are able to understand the conversation even when the interfering sources have an energy level similar to that of the source on which the listener is focusing. Methods for source separation have often been used to enhance the audio quality of target speech in communication systems. These sorts of tasks would fall into the AQO category described above. The enhancement of audio signals has also been used in non-communication tasks. For example, the restoration of old audio recordings has used source separation techniques to achieve better results. Audio imputation consists in restoring missing or corrupted regions in the time-frequency representation of an audio signal. Recently we have seen some researchers (Han et al., 2012) focusing on this task from a source separation perspective.

On the other hand, several applications target the extraction of source-specific information other than the audio signal from the mixtures. Researchers have long been working on Automatic Speech Recognition (ASR) systems that are able of transcribing speech and identifying speakers from audio recordings. In line with the *cocktail party effect* problem, recently we have observed an increasing interest in ASR from audio mixtures (Di Persia et al., 2007; Persia et al., 2008; Marti et al., 2012). These efforts have not been limited to speech signals, lately we have seen a growing interest in other common environmental sounds, such as the detection and classification of acoustic scenes and events (Aucouturier et al., 2007; Giannoulis et al., 2013).

Motivated by the questions raised in the domain of psychoacoustics and needs arising in the field of robotics, there has been extensive work on acoustic source localization (Blauert, 1983; Knapp and Carter, 1976; Schmidt, 1986; Asano et al., 1999). These methods are often based on source separation methods such as beamforming.

Music source separation

Music signals are especially interesting targets for source separation methods. We often find many more sources (instruments) than mixtures (channels) in music which renders the task quite difficult. Due to harmony and rhythm there is a large amount of structure and *a priori* knowledge in the signals. However there is also a lot of overlap between sources which makes the problem yet more challenging.

One of the most difficult applications of music source separation is to isolate the sounds produced by the different instruments of a song. This process is often called "unmixing" or "demixing", and the goal is to produce a multitrack recording from a mono or stereo version of a recorded piece of music. In the optimal scenario, the recovered audio tracks would have the same quality as if they were recorded separately. This task clearly belongs to the AQO category proposed by Vincent et al. (2003). In some cases the task consists in recovering the dry sources, without any applied effects such as reverb or delays. "Unmixing" is a complex task and current methods are far from achieving satisfactory results, except for some very specific situations.

A more attainable goal from a quality point of view is the process of "remixing". This task consists in performing transformations to the individual unmixed sources and mixing them back together. When the recovered and transformed sources are remixed, many of the artifacts and interferences may be masked, reducing significantly the audible errors. This process has many applications in the context of audio post-production. Recently, products that provide this solution, such as the Direct Note Access extension for Melodyne (Celemony Software GmbH, 2009), are generating a lot of interest from music producers and sound engineers.

The "remixing" process is also used to create specific mixes for different speaker configurations when the original multitrack version is not available. This task is named "downmixing" or "upmixing" depending on whether the speaker configuration has fewer or more channels than the original. With the increase in 3-D audio setups, the use of audio source separation for these tasks is gaining traction (Shim, 2009; Fitzgerald, 2011). Some applications only require removing one specific instrument from the mix, resulting in so-called "minus-one" mixes. This kind of process is used to create *karaoke* versions from mono or stereo recordings or practice versions of songs for musicians.

There are also several source separation applications in the SO category related to the field of music. Music Information Retrieval (MIR) research focuses on automatically extracting meaningful information such as instrument recognition, beat detection or polyphonic transcription from music signals. Very often researchers target the analysis of music signals with multiple simultaneous instruments, in such cases source separation methods may be used as a preprocessing step. Burred (2008) takes a source separation approach to polyphonic instrument recognition. Zapata and Gómez

(2013) remove the singing voice in order to improve accuracy of beat estimation and tracking methods.

Another field benefiting from music source separation is audio compression (Vincent and Plumbley, 2005). Audio compression seeks to achieve a faithful representation of the audio signal in a condensed form. Source separation techniques usually lead to sparse representations of the music signals, which help in reducing redundancies.

Realtime/low-latency separation

Realtime audio processing is becoming an important research topic in the fields of communications, robotics and music. Realtime processing on most devices often requires low latency and low computational costs. Additionally audio techniques on certain embedded devices need to work under limited memory conditions. Many recent applications impose these requirements on the use of source separation methods.

In communication technologies, many researchers are exploring the advantages of employing realtime source separation methods for denoising and speech enhancement (Joder et al.; Duan et al., 2012). With the boom in home automation and humanoid robots, robot audition is becoming an important field of research. In scenarios where multiple sound sources are present, such as office or outdoor environments, source separation is often used to detect, localize and recognize acoustic events (Asano et al., 2001; Nakadai et al., 2006; Valin et al., 2007). In most cases the response time of the robot must be short and therefore low-latency processing is often necessary. Hearing aids present another application area where realtime and low-latency source separation is crucial. For a long time, hearing aids have been exploiting beamforming techniques to improve the listening of humans. Recently we have seen increasing interest in improving cochlear implants in contexts where music signals are present by means of other source separation methods (Hidalgo, 2012).

Low computation cost and low-latency source separation also has many applications specifically for music signals. The widespread availability of music from online sources in streaming mode is motivating the creation of source separation techniques which can quickly process large collections and can operate on the data as it arrives (Dessein et al., 2010). Interactive music editing applications often require low-latency responses to user input. In these cases low-latency source separation can be used to achieve immediate previews while editing (Celemony Software GmbH, 2009).

High-latency/computationally expensive separation

There are many music source separation applications where response time is not a restriction. In these cases the main goal is often to perform the separation that will lead to results with the highest possible perceptual quality while keeping interference between the sources to a minimum. Examples of this type of application can be found in Digital Audio Workstations (DAW). These are starting to include functions to perform pitch and temporal modifications of individual notes in polyphonic music mixtures (Celemony Software GmbH, 2009). Fast response time in rendering the modified audio is less important than high quality separation in this type of application. Karaoke system developers are also exploring the use of blind source separation methods to automatically generate versions of commercially available songs without vocals. Since the version can be generated before the reproduction, the process can take a long time for computation. The main goal is to completely remove the voice source without producing perceivable artifacts. This type of "minus-one" process can also be applied in other contexts such as in music performance training. Another field interested in high-quality separation regardless of computational cost is the entertainment industry. Restoration or dubbing of old films often requires processing separately signals of different audio sources which are only available in a mixture (Burred and Leveau, 2011; Pedone et al., 2011).

1.2 Motivation and Objectives

This dissertation focuses on source separation methods for western commercial monaural and stereo music. The main motivation is the use of source separation methods for musical signals in a wide spectrum of practical situations. Currently source separation methods are used mainly in research and rarely exploited in real-world applications. Several companies have pointed out that in some situations the quality of music source separation methods may be acceptable but the methods are not used because they are too computationally expensive and too slow. In other situations the computational cost and latency may be acceptable, however the separation quality is not high enough.

Nowadays low latency and memory restrictions constraints are becoming common due to the increase in low cost computing devices. The growing availability of music streamed via the Internet and the desire to manipulate it on mobile or embedded devices makes achieving fast, low-computational-

cost source separation methods for music more necessary.

In settings where computation and memory constraints are not an issue, current source separation methods have often targeted general solutions to accommodate most signal types. In real world applications we may reduce the scope of the problem by targeting specific types of music and sources. By sacrificing generality and restricting the set of targeted sources we may further exploit a priori knowledge and increase separation quality.

The objective of this thesis is to design, develop and evaluate methods for separating musical components found in commercial western music which are practical for use in a wide variety of real-world applications. This work belongs in the context of AQO (Audio Quality Oriented) applications as described above, where the property of interest is the audio quality of the isolated sources.

The strategy used to achieve the objective is to take different research paths for the low-latency and high-latency scenarios. In low-latency scenarios the signal must be processed as it arrives and often has constraints on memory consumption and computational cost. If the processing is fast enough it may lead to real-time processing that can have a wide range of applications. In high-latency situations we assume access to large blocks of data which often requires higher memory consumption and greater computational complexity. However the quality of the separation can be significantly improved, since more information is available at the moment of processing.

1.3 Context of the Study

This work was done in the Music Technology Group (MTG), Universitat Pompeu Fabra (UPF) in Barcelona. The research presented in this dissertation was conducted under the umbrella of the Monet project, a 3-year joint research project with the Yamaha Corporation. The goal of the project was to develop practical methods to produce minus-one mixes of commercially available western popular music signals. Minus-one mixes are versions of music signals where all instruments except the targeted one are present.

1.4 Presentation of Contributions

This dissertation comprises three main contributions. First, we perform an extensive review of source separation methods with a special focus on

those applied to musical signals and based on Non-negative Matrix Factorization. Second, we propose a new spectrum factorization technique based on Tikhonov regularization, which is simple, intuitive and computationally less expensive than NMF, making it specially interesting for contexts in which low-latency processing is sought. This technique is evaluated in a series of applied tasks comprising single and multiple pitch estimation and singing voice, bass and drums separation. Third and last, we propose a set of enhancements to state of the art source separation techniques in high-latency and semi-supervised scenarios. Due to the context of this study as a collaboration with a corporation, all the developments in this dissertation focus principally on western popular music widely available commercially and on practical real-world situations.

1.5 Organization

This thesis is divided into four main parts: Part I provides an introduction, context and a review of the state of the art. Part II is devoted to low-latency scenarios. Part III is dedicated to high-latency cases. And Part IV contains conclusions and suggestions for future research directions. Figure 1.1 shows the thematic dependencies between the chapters.

Part I: Chapter 1 is this Introduction.

Chapter 2 is a definition of the context and the problem. It starts by presenting a formal definition of the source separation problem and the elements involved. It then reviews the possible representations of signals involved in source separation tasks. Next it discusses the different types of mixing processes, the generated mixtures and their nature. Chapter 2 also provides an overview of the types of sources and their properties. It contains a definition of the target conditions, constraints and scenario. Lastly, it describes the typical evaluation frameworks and datasets employed in the field.

Chapter 3 is an exhaustive review of state of the art techniques in the field of source separation. The presentation is structured by order of generality and applicability to our target scenario. It begins with general statistics-based methods and continues with a review of beamforming techniques. Special attention is dedicated to music signal modeling and spectrum decomposition methods which are commonly applied in our context.

Part II: Chapter 4 introduces Tikhonov regularization as an alternative to

NMF under low-latency and low computational costs constraints, and considers the tasks of pitch and multipitch estimation as application domains.

Chapter 5 focuses on source separation applications of the Tikhonov regularization spectrum decomposition method. It presents experiments covering three instruments: singing voice, bass and drums.

Part III: Chapter 6 is dedicated to improvements of current high-latency separation methods with a focus on singing voice, drums and mixes with multiple monophonic sources. Regarding the singing voice, it considers two common issues with current separation methods. First a technique to estimate and separate the breathiness component of the voice is presented. Second, a method to estimate and separate the unvoiced fricative consonants is proposed. With respect to drums, this chapter explores the use of regularizations and constraints based on transient estimation and manual annotations to enhance the isolation of percussive sources. Finally, the chapter contains a section dedicated to the use of multiple pitch annotations for the separation of music mixes with multiple harmonic sources, often encountered in classical music pieces.

Part IV: Chapter 7, presents the conclusions and a summary of the contributions and results. It also suggests potential future directions for research. Finally, it contains a compilation of the outcomes of this work.

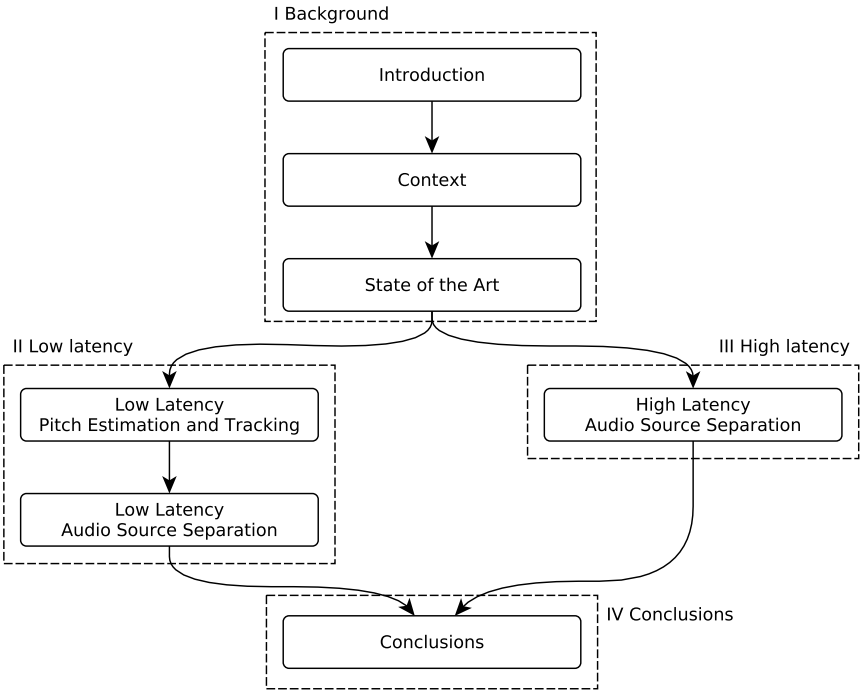

Figure 1.1: Overview of thematic dependencies.

Chapter 2

Audio Source Separation Basics

The best choice of Blind Audio Source Separation (BASS) methods is highly dependent on the *a priori* information we have and the assumptions we can make regarding the properties of the sources and the mixing process. In this chapter we define a global framework for the source separation problem. We review the different types of mixtures and sources involved and how these affect the difficulty of the task.

2.1 Problem Definition and Classification

The problem of source separation consists in retrieving one or more source signals given a set of one or more observed signals in which the sources are mixed. The observed signals are sometimes referred to as observations, sensors or mixtures and the sources are sometimes referred to as components. The concept of mixture is very general and can include any function of the set of original source signals. We can express a given set of observed signals in the following way:

$$(v_o[n])_{1\leq o\leq N_O} = f\left((x_m[n])_{1\leq m\leq N_M}\right) \tag{2.1}$$

where $(v_o[n])_{1\leq o\leq N_O}$ are the observed signals (or mixture) and $(x_m[n])_{1\leq m\leq N_M}$ are the source signals we want to retrieve. f is a mixing function that generates a set of observed signals given a set of source signals.

Formulated in such way the problem is very general and open. However, depending on the field of application and the task at hand, there are numerous assumptions, approximations and/or bits of prior knowledge that can be used to render the problem more tractable.

Four main characteristics of the problem are determinant to the difficulty and approach chosen:
– The properties and knowledge of the mixing process f
– The properties and knowledge of the sources $x_m[n]$
– The ratio between the number of mixtures and sources $\frac{N_M}{N_O}$
– Conditions and constraints on the separation procedure
O'Grady et al. (2005) presented a survey of different methods with a classification depending on several of these properties. We will now present these characteristics in somewhat more detail.

2.2 Signal Representation

The sources and mixtures in our context are acoustic signals. The most common representation of an acoustic signal is the audio waveform. The waveform is a function in time of the sound pressure level. For most common applications the waveform is sampled at regular time intervals, which results in a digital signal or sampled waveform. An acoustic signal $s(t)$ and the sampled waveform $s[n]$ can be expressed as:

$$s(t) \qquad \forall t \in \mathbb{R}^+ \tag{2.2}$$
$$s[n] \quad = s(n * T_S) \quad \forall n \in \mathbb{Z}^+ \tag{2.3}$$

where T_S is the sampling period $T_S = 1/F_S$. When the signal is finite in time, then the sampled waveform is only defined for a given limited range $n \in [0, N_T]$.

The sampled waveform representation of an acoustic signal is used for many applications. However in most tasks other representations have proven more useful. Temporal signals can be represented in different domains or supports. The representation of the signal has a large influence on its properties and the techniques that we can use. A generic way to express a temporal signal is known as *additive expansion*. This consists in representing the signal as the weighted sum of a set of expansion functions.

$$s[n] \quad = \quad \sum_{\omega=0}^{N_\omega} c_\omega b_\omega[n] \tag{2.4}$$

where c_ω are the expansion coefficients and $b_\omega[n]$ are expansion functions or basis components.

If our signal $s[n]$ has finite length N_T and our basis expansion set has a finite number N_ω of basis functions that are also limited in the time support $b_\omega[n] \quad for \quad n \in [0, N_T - 1]$, Equation 2.4 can be expressed as a matrix expression:

$$\boldsymbol{s} \quad = \quad \boldsymbol{Bc} \tag{2.5}$$

where $\boldsymbol{s}$ is a column vector containing the values of our signal $\boldsymbol{s} = (s[0], s[1] \ldots s[N_T - 1])$. $\boldsymbol{c}$ is a column vector containing the expansion coefficients $\boldsymbol{c} = (c[0], c[1] \ldots c[N_\omega - 1])$ and $\boldsymbol{B}$ is an $N_T \times N_\omega$ matrix whose column vectors $\boldsymbol{b}_\omega$ are the vector representations of the basis functions $\boldsymbol{b}_\omega = (b_\omega[0], b_\omega[1] \ldots b_\omega[N_T - 1])$.

Using the *additive expansion* expression, the temporal representation of our signal can be rewritten as an expansion over time-shifted *impulses*:

$$s[n] \quad = \quad \sum_{k=0}^{N_T-1} c_\omega^t \delta[n-k] \tag{2.6}$$

where the expansion coefficients are simply the waveform samples $c_\omega^t = s[\omega]$. This representation is often referred to as a *time localized* representation, because the energy of the basis components is localized in the time domain.

Similarly, the *frequency localized* representation is also widely used in audio tasks. This consists in expanding the signal as a weighted sum of functions with their energy localized in the frequency domain. The Discrete Fourier Transform (DFT) is probably the most well known frequency-localized representation:

$$s[n] \quad = \quad \sum_{k=0}^{N_T-1} c_\omega^{dft} e^{i2\pi \frac{k}{N_T} n} \tag{2.7}$$

In the DFT representation, the expansion functions are single frequency complex exponentials and the expansion coefficients are computed as:

$$c_\omega^{dft} \quad = \quad \frac{1}{N_T} \sum_{k=0}^{N_T-1} s[n] e^{-i2\pi \frac{k}{N_T} n} \tag{2.8}$$

The signal models most commonly used in digital audio processing are *time-frequency localized* representations, in which the expansion functions are

localized in both the time and the frequency domains where the expansion functions and coefficients are indexed in two dimensions. One index is for the time localization and the other for the frequency localization. However this can still be seen as the addition of a set of expansion functions. One of the most well known representations is the short-time Fourier transform (STFT). In this case, the expansion functions are single frequency complex exponentials multiplied by a time localized windowing function:

$$s[n] = \sum_{t=0}^{M-1} \sum_{\omega=0}^{N_T-1} c_{t,\omega}^{stft} w[n - tH] e^{i2\pi \frac{\omega}{N_T} n} \tag{2.9}$$

where $w[n] \quad for \quad n \in [0, L-1]$ is a windowing function of length L, and H is the hop size of the windowing function. In the STFT representation, the expansion coefficients are computed as:

$$c_{t,\omega}^{stft} = \frac{1}{N_T} \sum_{\omega=0}^{L-1} s[n - tH] w[n] e^{-i2\pi \frac{\omega}{N_T} n} \tag{2.10}$$

Since the STFT is a signal representation that we often use, we will refer to the STFT coefficients of a signal $s[n]$ as $\widetilde{\boldsymbol{S}}(\omega, t) = c_{t,\omega}^{stft}$ where t is the index of the time axis and ω the index of the frequency axis. Throughout this text we will use $\boldsymbol{S}(\omega, t)$ as a simplified notation for either the magnitude $|\widetilde{\boldsymbol{S}}(\omega, t)|$ or the power spectrum $|\widetilde{\boldsymbol{S}}(\omega, t)|^2$ depending in the context.

One limitation common to all signal representations is analogous to the uncertainty principle. In the context of signal processing it is often called the Gabor limit, which states that a function cannot be both time limited and band limited. In other words there is a limit to the joint temporal and frequency localization of our signals. The limitation is common to all signal representations, however different representations have different temporal and frequency resolutions. There is a tradeoff between time and frequency resolution, the higher the frequency resolution the lower the temporal one and vice versa.

One of the main strengths of the DFT and STFT is that a very fast and efficient algorithm, the Fast Fourier Transform (FFT), is used to calculate them. Furthermore the FFT is implemented on most available platforms and even in some hardware implementations, which often permit fast, energy-efficient computation. However the DFT and STFT also have their shortcomings. Their main limitation in the field of audio processing is that the resolution of their frequency axis is constant on the Hertz scale, while

humans perceive frequency differences linearly on a logarithmic Hertz scale Stevens et al. (1937). In practice this means that when choosing a frequency resolution for our STFT we must decide between having good frequency resolution or separability for the low frequencies or good temporal resolution for the high frequencies. This choice determines the size of the window used for the analysis, the larger the window the higher the frequency resolution and the lower the temporal resolution.

In Appendix A we review some of the signal representations found in the literature that tackle the frequency-temporal localization limitation. It remains as future work to evaluate these alternative representations for use as a replacement for the STFT representation that used throughout our work.

The expansion of a signal into a set of basis components allows us to exploit the linearity property of some systems. In such cases we can understand how the system affects our signal by studying how it affects the individual basis components. There are several reasons for choosing a specific signal representation. Some properties of a signal are obscured or revealed depending on the representation that we use. Features that could be advantageous for a given task are more prominent in certain representations. On the other hand, in some cases we have *a priori* information about the source in a specific representation. The mixing process is often simpler and easier to estimate using a specific representation of the signals being mixed. Computing certain representations of a signal is often not a trivial task. The process of transforming a signal from one representation to another may impose some conditions on the availability of the data and affect the complexity of the process. Therefore the conditions imposed on the separation tasks also influence the choice of the representations used. These cases will be reviewed in the following sections 2.3, 2.4, 2.6.

2.3 Mixing Processes

The mixing process in music signals is a practical, aesthetic, and/or creative treatment by which multiple recorded sounds are combined into one or more channels in order to produce a mix that is more appealing to listeners. From a signal processing point of view, it consists in creating a set of observed mixture signals from a set of source signals. Information about this process can significantly help the separation task. In this section we review the most common mixing processes and the way they are modeled. We also discuss the availability of knowledge of the model parameters depending on the

scenario. First we present mixing situations that are general and common to many different type of signals. Next we focus on mixing processes specific to music signals.

General mixing processes

Instantaneous

The simplest model of a mixing process consists in assuming that the mixtures are linear combinations of the sources:

$$v_o[n] = \sum_{m=1}^{N_M} a_{om} x_m[n] \tag{2.11}$$

where a_{om} is the contribution gain of the m^{th} source to o^{th} mixture. Mixing processes so modeled are called **instantaneous mixtures**. The name reflects the fact that the mixture is performed instantaneously for every input sample of the sources, without requiring information from previous input.

Given a specific time index, this model can be expressed in the following matrix notation:

$$\boldsymbol{v} = \boldsymbol{A}\boldsymbol{x} \tag{2.12}$$

where $\boldsymbol{v}$ is a column vector containing the values of our mixtures at a given time frame $\boldsymbol{v} = (v_1[n], v_2[n]...v_{N_O}[n])$, $\boldsymbol{A}$ is a $N_O \times N_M$ matrix containing the mixing gains a_{om}, and $\boldsymbol{x}$ is a column vector containing the values of the sources at a given time frame $\boldsymbol{x} = (x_1[n], x_2[n]...x_{N_M}[n])$. The matrix $\boldsymbol{A}$ is often called the mixing matrix.

If the coefficients of the mixing matrix $\boldsymbol{A}$ are known, the problem is a basic linear problem with many solutions available in the literature. However in most source separation cases the coefficients of this matrix are not known and must be estimated. This renders the problem much harder and specific solutions must be used. As we will see in section 2.5, the ratio between the number of sources N_M and mixtures N_O will determine the availability of a unique solution.

Attenuated and delayed

A common scenario in the fields of communications and signal processing is to have a sensor array when capturing signals. This configuration creates a mixture signal for each sensor, in which the source signals are mixed.

In most cases the source signals will be captured by all or most of the sensors. However the positions of the sources with respect to each sensor and the travel speed of the signals will affect the delay and attenuation with which the source signal will contribute to each mixture. Given the previous situation, if we consider the sources static in space we may express the mixing process as:

$$v_o[n] = \sum_{m=1}^{N_M} g_{om} x_m[n - d_{om}] \tag{2.13}$$

where g_{om} is the attenuation with which the m^{th} source contributes to the o^{th} mixture, and d_{om} is the delay in sample units with which the m^{th} source arrives at the o^{th} sensor. This is a very common scenario when we have control over the capturing process and are able to use sensor arrays. However in many cases the mixtures have not been created in such a way and using this model would lead to incorrect results.

A matrix representation of such model would consist in:

$$\boldsymbol{v} = \boldsymbol{A} * \boldsymbol{x} \tag{2.14}$$

where $*$ denotes the elementwise convolution. In this case, the mixing matrix A would be composed of the following elements $\boldsymbol{A}_{om} = g_{om}\delta[n - d_{om}]$, where $\delta[n]$ is a Kronecker delta whose value is 1 at $n = 0$ and 0 elsewhere.

Convolutive

A more general model for many mixing processes is to consider the mixtures as a sum of filtered sources. In each mixture, instead of having a single attenuated delay of each source, we have multiple attenuated delays of each source:

$$v_o[n] = \sum_{m=1}^{N_M} \sum_{k=0}^{L} h_{om}[k] x_m[n - k] \tag{2.15}$$

where $h_{om}[k]$ are the coefficients with which the m^{th} source is filtered when contributing to the o^{th} mixture and L is the maximum length of the filters. This model is often referred as **convolutive mixtures**, since the sources are convoluted into the mixtures.

The matrix representation of 2.15 is the same as for 2.13:

$$\boldsymbol{v} = \boldsymbol{A} * \boldsymbol{x} \tag{2.16}$$

However in this model, the mixing matrix $\boldsymbol{A} \in N_O \times N_M \times L$ is composed of the filter coefficients $\boldsymbol{A}_{omk} = h_{om}[k]$.

It can be easily seen that 2.13 is a specific case of 2.15. An attenuation and a delay mixing process is a linear system that can be reformulated as a convolution by choosing the filter as:

$$h_{om}[k] = \begin{cases} g_{om} & \text{if } k = d_{om} \\ 0 & \text{if } k \neq d_{om} \end{cases} \tag{2.17}$$

Music mixing processes

Since in this study we concentrate on music signals, we now present some of the properties common to most music mixing processes.

Non-physical delays

The general mixing models presented in the previous section (Sections 2.3 and 2.3) are rarely directly applicable to the music signals we target in this study. Our main focus in this work is on methods that can be applied to widely available commercial western popular music. This type of music is often mixed in studios using a Digital Audio Workstation (DAW) or a multitrack mixer. Using such devices for mixing normally invalidates the assumptions made in the Attenuated and Delayed mixture model.

In stereo or other multichannel music signals, attenuations and delays can be artificially added during the mixing process using the DAW. This allows producers to give a sense of source localization. In some cases only attenuation of the different sources is used in order to move the sources in the stereo image. This is often known as the panning control of the sources in the mixing process.

A musical signal may comply with the assumptions of the Attenuated and Delayed mixture model when the excerpt has been recorded with a microphone pair. An extensive study and discussion of such recordings in the context of source separation can be found in the work by Burred (2008). Such spatial information is valuable and we will show how to use it when it is available. However, it is not always present and our goal is to study methods that can also handle situations in which multiple sensors are not available, such as monophonic music signals.

Spatially Static sources

One main characteristic of music signals in contrast to other audio mixtures is that most often the sources mixed are statically or almost statically positioned in the spatial image. For instance, in a music ensemble recording the musicians stay in the same place during the whole piece and therefore the attenuation and delay of that instrument will remain constant. In studio mixtures this is often also true, that is, the instruments stay in the same or a similar position in the spatial image throughout the song.

In popular western music the assumption can be taken a step further. Most pop and rock songs maintain a very similar distribution of the instruments in the stereo space. The drums and lead vocals are typically panned in the center channel, the bass is often panned slightly to the right while other accompaniment instruments such as rhythm guitars get panned to the sides. This disposition replicates the distribution of the musicians in a scene during a live performance. By panning the instruments in this manner the producer tries to recreate the sonic situation we are used to when attending a live show.

These properties make it possible to perform separation without the need for tracking the positioning of the sources. This situation led to the proposal of a series of simple methods based on simple spectral bin classification derived from the clustering of the spatial positions of the bins (see Section 3.2).

Even though these assumptions cannot always be made, they can often be used as *a priori* information during the source separation process.

Reverberation

Another common effect used when mixing popular western music is reverberation. Reverberation is the persistence of sound in a particular space after the original sound is produced. It is often caused by the multiple reflections of the sound waves on the boundaries of the space.

Music producers often replicate this effect to recreate the acoustic sensation of being in a certain room, stadium, chapel, concert hall or other space. The reverberation effect can be applied in many different ways, it can be applied to each source individually and/or applied to the mixture itself. The amount and type of reverberation are highly related to the music genre and style. For instance while classical music is often recorded with all the instruments playing live in a concert hall, in pop and rock music the instruments are often

recorded individually in a studio under dry conditions and the reverberation effects are synthetically added afterward.

In the field of signal processing this effect is often modeled as a filter with a very long tail that can easily surpass 1 second in length. Reverberation has been widely studied in the literature. In our particular scenario reverberation time, decay and/or how it is applied to the different sources and the mixture is unknown. Therefore using the characteristics of reverberation as *a priori* information to source separation methods can be a challenging task. The lack of proposed BSS methods exploiting or targeting reverberation in musical signals is a hint to how challenging the task is.

Non-linear Effects

One of the last steps in the music production chain consists in mastering the final mix. This process may be performed using non-linear effects. The use and nature of non-linear effects largely depend on the type of music. Modern rock and pop make a wide use of compression and distortion, while jazz and classical music often avoid such effects. There are a large number of non-linear effects that can be applied and they cannot be easily taken into account when modeling the signal, therefore they are often considered as noise in the signal. Recent work (Gorlow and Reiss, 2013) has begun to target the inversion of some of these non-linear operations on music signals. However these studies are restricted to scenarios in which the parameters and the procedure of the effect are known in advance.

2.4 Source Properties

Defining the target source to be recovered is one of the first tasks in the source separation problem. In the context of music it is common to define the target sources as the vibrant physical entities producing the sounds.

However, depending on the task there may be exceptions to this definition. For instance, it is quite common to treat the vocal chorus as one single source or the whole group of violins that are playing the same voice as one single source. Another exception where a target source does not map directly to a physical entity is with the drum kit. Drum kits are composed of several drums and other percussive instruments, but in most source separation tasks these are treated as a single *drums* source.

The definition of the target source will greatly influence the assumptions and prior knowledge that we can include in our models. Therefore, and in order to avoid confusion, throughout this research we define clearly in each task and experiment which sources we target. Once the target source is defined we can start making some assumptions about the signals in order to have a model with fewer free variables, which will then be easier to learn from the data. In this section we will describe properties of the individual sources that will allow us to make certain assumptions in order to render the problem more approachable. We will also define some properties that are applicable to sets of sources. These refer to relations between the sources and allow us to assess their separability given a mixture.

General Properties

At first we will only consider properties common to all signals. These are generic properties that are independent of the nature of the source and that derive mostly from the mathematical definition of a signal as a temporal sequence of values.

Statistical Independence Sometimes we face the problem of not having any prior knowledge about the sources. This can happen in situations where we do not know the nature of the sources that we are studying or when the sources do not present any specific or interesting features. In those cases the most common assumption is to view target sources as independent random variables. As we will see in Section 3.1, these assumptions lead to the development of the first set of well-known source separation methods.

The statistical independence of the sources is defined as:

$$p(x_1, \ldots, x_{N_M}) = \prod_{m=1}^{N_M} p(x_m) \tag{2.18}$$

This condition implies that for any $m_1 \neq m_2$, t and τ:

$$E[x_{m_1}(t)x_{m_2}(t+\tau)] = E[x_{m_1}(t)]E[x_{m_2}(t+\tau)] \tag{2.19}$$

Which means that the occurrence of one source does not affect the occurrence probability of the other. In terms of variables, it implies that the observed value of one does not affect the probability distribution of the other. This applies to multiple statistics of different orders such as the variance or the kurtosis. While the statistical independence of the sources is

often a necessary condition or assumption in most methods, we often also require a more restrictive assumption such as the difference between these statistics.

Stationarity The stationarity of a signal often refers to the invariance with respect to time of the joint probability distribution of it's values. If we consider the sequence of signal coefficients $c_{t_1}, \ldots, c_{t_T}$ as a discrete-time random process $\{\boldsymbol{C}_t\}$ and let $F_{\boldsymbol{C}}(c_{t_1+\tau}, \ldots, c_{t_k+\tau})$ represent the cumulative distribution function of the joint distribution of $\{c_t\}$ at times $t_1+\tau, \ldots, t_k+\tau$. We can express the stationarity condition as:

$$F_{\boldsymbol{C}}(c_{t_1+\tau}, \ldots, c_{t_k+\tau}) = F_{\boldsymbol{C}}(c_{t_1}, \ldots, c_{t_k}) \quad \forall \tau, k, t_1, \ldots, t_k \tag{2.20}$$

When the condition only evaluates true for the first two statistical moments it is called weak–sense stationarity, wide–sense stationarity (WSS) or covariance stationarity. In this case the mean and variance of the stochastic process is constant over time.

In some source separation situations the mixture signals can be explained as a linear superposition of stationary and nonstationary sources. The assumption that the target source is nonstationary in contrast to the other sources in a mixture, is often used to derive source separation methods. This is the case for Subspace Stationarity Analysis (SSA). This type of method has been used to separate sources principally in the fields of Brain Computer Interfacing and Speech Separation.

Correlation Correlation between sources is another property that can be exploited in the separation techniques. Given a mixture that can be modeled using Equation 2.14, we can easily show that the correlation matrix of the mixtures $\boldsymbol{R_{vv}} = E[v_i[n]v_j[n]]$ maintains the following relation with the correlation matrix of the sources $\boldsymbol{R_{xx}} = E[x_i[n]x_j[n]]$:

$$\boldsymbol{R_{vv}} = \boldsymbol{A}\boldsymbol{R_{xx}}\boldsymbol{A}^\top \tag{2.21}$$

When the sources are uncorrelated, the matrix $\boldsymbol{R_{xx}}$ will be a diagonal matrix since all the cross correlation terms between the sources will be 0. As we will see in Section 3.1 the problem of estimating the mixing matrix $\boldsymbol{A}$ is reduced to a linear algebra diagonalization problem. Of course, special care must be taken when working with a different number of mixtures and sources. In real world cases where the mixing process adds noise or the sources are not fully uncorrelated, this property can still be used as

an approximation. In a more general case we can exploit not only the correlation but also the temporal cross correlation of the sources. In such situations equation 2.21 can be reformulated using the cross correlation matrices $\boldsymbol{C_{xx}}(\tau) = E[x_i[n]x_j[n-\tau]]$ and the extra dimension of τ turns the problem into a tensor diagonalization problem.

Disjointness Another key aspect influencing the separation process is the disjointness of the sources in the mixture. Disjointness refers to the non-overlapping of the energy of a set of signals in a given representation, see Jourjine et al. (2000) and Yilmaz and Rickard (2004). We must note that disjointness depends not only on the individual properties of the sources or the mixing process but on their mutual properties. On the other hand the signal representation has a high impact on the degree of disjointness of a set of signals. As an example, lets consider two pure sinusoids of very different frequencies occurring at the same time. The time domain mixture signal presents very low disjointness since both signals overlap in energy at most instants. However in a time-frequency representation these signals will barely overlap. Thus, the representation of the signal can greatly influence the apparent disjointness of a set of signals.

Sparsity In practice, when choosing a good signal representation we may not have information about the specific source sets that will be present in each mixture. Therefore we will not be able to know for sure the representation that will maximize their disjointness. However we may select a signal representation that is more likely to produce highly disjoint source sets. This is done by considering the sparsity of the representations. Sparsity is the property of a signal that relates to the amount of non-zero coefficients in a given representation. A sparse representation of a signal is one in which there are few non-zero coefficients.

Although high sparsity of the components does not guarantee high disjointness, low sparsity does guarantee low disjointness. In the extreme case where two sources have energy in all their coefficients they will definitely overlap, and therefore have low disjointness.

Several criteria for sparsity have been proposed in the literature. The measure based on the strict definition of sparsity corresponds to the L^0-

"norm" [1]:

$$\kappa_{L^0}(c_\omega) = \sum_i^{N_\omega} |c_i|^0 \tag{2.22}$$

This measure corresponds to the number of non-zero coefficients of the signal.

However solving a problem that involves minimizing the L^0-"norm" of a signal is often hard. To overcome this problem, people use the L^1-norm:

$$\kappa_{L^1}(c_\omega) = \sum_i^{N_\omega} |c_i| \tag{2.23}$$

The use of the L^1-norm leads to well-known efficient linear programming solutions (Rolewicz, 1985). In the context of Compressive Sampling, Candes and Wakin (2008) have proven that when the solution is sufficiently sparse, minimizing the L^1-norm is equivalent to minimizing the L^0-"norm".

Hoyer and Dayan (2004) propose a sparseness criteria based on the relation between the L^1 and L^2 norms:

$$\kappa_{L^1,L^2}(c_\omega) = \frac{\sqrt{N_\omega}\sum_i^{N_\omega} |c_i| / \sqrt{\sum_i^{N_\omega} c_i^2}}{\sqrt{n} - 1} \tag{2.24}$$

The measure evaluates to 1 if and only if c_ω contains a single non-zero component, takes the value of 0 if all components are equal (ignoring the signs), and interpolates smoothly between the two extremes. This measure has been used for the development of Non-negative Matrix Factorization with sparseness constraints, see 3.4.

Burred (2008) presented an extensive study about the influence of signal representation on the sparsity of sources and the disjointness of mixtures. He evaluated different time-frequency representations on two different datasets: speech and music. The representations tested were: STFT, Constant-Q (CQ), Equal Rectangular Band (ERB), Bark Bands and Mel Bands. The methods were compared in terms of sparsity of the individual sources and the disjointness between them in given mixtures. The experiments showed that representations well adapted to the signals in play have a great impact on sparsity and disjointness measures.

1. Note the quotes around the word norm. The measure, proposed by Donoho is not a proper F-norm, because it is not continuous with respect to scalar-vector multiplication (as the scalar approaches zero).

Acoustic Sources

All the sources comprising music mixtures in this work are acoustic signals. We may therefore exploit knowledge about the properties and characteristics of acoustic sources in performing the separation.

Harmonic Structure A characteristic common to many acoustic signals is that for long periods of time they are periodic or quasiperiodic. This is often the case for sounds produced by sustained vibrating bodies such as many musical instruments (guitar, flute, saxophone...) or speech and singing voice.

The main property of such signals is that their waveform can be decomposed into a summation of a set of sine waves called the fundamental and the harmonics, partials, or overtones.

These sounds have a harmonic frequency spectra. Usually, the lowest frequency present is the fundamental, and is the frequency at which the entire wave vibrates. The overtones vibrate faster than the fundamental, but must vibrate at integer multiples of the fundamental frequency in order for the total wave to be exactly the same each cycle. In some cases the partial corresponding to the fundamental frequency is missing, however the harmonic structure remains and the fundamental frequency can be estimated from this structure.

The energy of these sounds are concentrated in narrow regions around the partial positions in the spectra. These types of signals are called narrow band or frequency-localized components.

Transients In contrast to narrow band harmonic structured acoustic signals, we often encounter time localized sounds. These are often referred to as transient sounds. This type of sound is often generated from impacts such as in percussive musical instruments or fast changes of airflow found, for example, in plosive consonants.

The distinguishing characteristic of this type of sound is that the energy of the signal is concentrated in a short segment of time. These signals have their energy distributed in a wide range of frequencies and therefore are often called wideband signals. An example is the sound produced by snare drums, a very noisy wave shortly after they are struck. The double headed nature of this instrument means that the harmonics are extraordinarily complex since there are, theoretically, an infinite number of resonant

modes, and thus harmonics. In practice, though, the sound produced is not infinitely harmonic, rather it is extremely dissonant, due to the large number of constructive and destructive interfaces.

Noise In acoustics noise refers to any unwanted sound. It is clearly a context or task dependent definition. In the context of source separation noise refers to sounds other than the ones targeted for isolation. Given such a definition any type of sound could be a noise, if it is an unwanted component of the mix. In some cases noise refers to both unwanted and random components of the signal. Here we present several types of sounds that are often considered as noise in many audio tasks. Most audio mixtures are composed of audio recordings. The recording system and electronics are responsible for a random additive component. However in most professional audio recordings this component is very small, and can often be discarded when comparing to other noise sources.

Another source of noise that often appears in audio recordings is the ambient noise. These are audio sources that are not the target source of the recording and are mixed during the recording. In a recording of a live concert, the sum of the audience voices and sounds composes such noise. In this case the individual noise signals are not completely random since speech signals are highly predictable and deterministic. However the level at which they are recorded and the fact that they are all summed together and interfere with each other renders them less modelable.

While noise often refers to random components in the signal, not all random components are unwanted. In source separation tasks, we often want to isolate the drums. The sounds generated by drums are inherently noisy by nature. The spectra of such sounds do not present a harmonic structure and in some cases they are not time localized, but are rather sustained during a long period of time. An example of such sounds are those produced by cymbals.

Human auditory perception An important factor to take into account when processing acoustic signals, is the way these are perceived by humans. One characteristic of human auditory perception is its large dynamic range. The difference between the threshold of hearing and the threshold of pain is around 100 dB (Rossing (1990)). Therefore low-energy components in audio are important from a perceptual point of view. This property of human perception will play an important role in selecting a good divergence function between spectra (see 3.4). It is well known that humans perceive pitches

equally spaced in an approximately logarithmic fundamental frequency scale (Stevens et al., 1937). This has often motivated the use of logarithmic distributions of spectral bins (Constant-Q), spectral bands (Mel scale) and pitch candidates. Another important property of the human auditory perception is that of masking when listening to complex stimuli. This has motivated the development of a subdivision of the audible frequency range into critical bands such as the Bark bands (Zwicker, 1961). Many other properties of auditory perception (Purwins et al., 2008b,a) may be taken into account when processing acoustic signals targeted at humans.

Music Sources

The sources mixed in musical audio signals are highly structured in the time and frequency dimensions. Additionally, depending on the music genre and culture, they often present a limited set of timbres, pitches and pitch ranges.

In this section we review some of this *a priori* knowledge that can be useful in the design of source separation methods.

Tonality In Western music of the major-minor tonality period, roughly from Bach's birth (1685) to Wagner's Tristan (1857), the tonality of a piece of music is defined by the kinship of tone centers (local keys during a short time frame).

Furthermore, the sources present in western music almost always use a limited set of pitches or concentrate the pitches used around a fixed set of frequencies from an equal temperament system. Equal temperament is a musical tuning system in which every pair of adjacent notes has an identical frequency ratio. As pitch is perceived roughly as the logarithm of frequency, this means that the perceived "distance" from each note to its nearest neighbor is the same for every note in the system.

In particular western popular and classical music is mainly based on the 12 tone equal temperament scale in which an interval of an octave is divided into a series of 12 equal steps, also known as semitones. This known distribution of notes can be of great utility when creating methods for pitch estimation of sources in western music.

Frequency Correlation Harmony is the use of simultaneous pitches or chords. The study of harmony has highly influenced western music since the 17th century. Theories of harmony provide examples of how to subsume

chords in different harmonic categories (Riemann, 1877; Schenker, 1935). Some traditions of Western music performance, composition, and theory have specific rules of harmony. These rules are often held to be based on natural properties such as the Pythagorean tuning law's whole number ratios or harmonics and resonances.

From a spectral point of view these properties translate to high correlation values at specific lags of the spectra between the sources in a music mixture. These correlation values are due to the fact that the notes in chords often used in pop/rock western music have overlapping partials. This quality is quite specific to music signals, in contrast to other acoustic mixtures such as speech that do not show such large correlations.

Temporal Correlation Rhythm is another very important aspect of Western music. Rhythm is the temporal structure of music and the organization of the different musical events in the music sequence. In Western music rhythm is often composed of highly regular and periodic structures. Pop and rock western music often present a section of instruments called the rhythm section (Randel (1999)). This section provides the accompaniment of the music, giving the music its rhythmic texture and pulse, also serving as a rhythmic reference for the rest of the band.

In Blind Audio Source Separation (BASS) we often encounter the task of separating the main melody or leading voice from the accompaniment. The use of rhythm or temporal correlation information is of great value in these cases, as we will see in Section 3.3.

Timbre Information Timbre is the quality of a musical sound that distinguishes different types of sound production, such as voices, string instruments, wind instruments, percussion instruments or other musical instruments. Timbre is the property that allows us to distinguish and recognize the different instruments in a musical mixture. This property allows us to distinguish the sound of a trumpet from the sound of a clarinet even if they are playing the exact same pitch at the same loudness.

There is a long history of research on musical timbre (Grey (1977); Wessel (1978)) focused mostly on studying the properties of a sound that contribute to its timbral qualities. Klapuri (2006) empirically found a spectral envelope that could best represent most of the pitches found in a given music dataset. Haro et al. (2012) showed that the most frequent spectral energy distributions in music, speech and environmental sounds are quite different.

In musical signals, both western and non-western, the energy of the most common spectral shapes is mostly concentrated in the first bark bands. In speech signals the energy is mostly concentrated in the first or middle bark bands, while in environmental sounds, the energy is often concentrated in the high frequency bark bands.

In the field of Music Information Retrieval (MIR) we have seen (Fuhrmann and Herrera (2011); Fuhrmann (2012)) a growing interest in developing automatic instrument recognition systems based on appropriate timbral characteristics. This type of work led to the creation of a set of features and machine learning algorithms used to classify different timbres. The study of sound production of musical instruments and the proposal of sound production models such as the source-filter model have been quite successful at decoupling to some extent the timbre information from the pitch information (Noll, 1967).

The timbral information of the sources can be a very important aid in Blind Audio Source Separation (BASS) of musical mixtures. When the pitch information is not sufficient to distinguish between sources, timbre may be the only alternative. Furthermore, in contrast to pitch, timbre information often changes slowly in time, and this temporal correlation of the sources can be exploited in the segregation of musical mixtures.

2.5 Ratio of Sources and Sensors

The ratio between the number of sources N_M and mixtures N_O is an important and determinant factor in the choice of source separation method. In this section we will review several common situations, with a special focus on music audio recordings and how these affect the selection of methods.

Determined or Over/under-determined

Many source separation techniques are based on a two-stage approach. The first step is estimating the mixing matrix $\boldsymbol{A}$ or the unmixing matrix $\boldsymbol{U} = \boldsymbol{A}^{-1}$ and the second step consists in unmixing the mixture signals. This is often the case with the statistics-based approaches (see Section 3.1). In such cases the existence of $\boldsymbol{U}$ and the possibility of computing it is sometimes necessary. Taking into account that $\boldsymbol{A} \in N_O \times N_M$, the existence of the unmixing matrix requires the rank of the mixing matrix to be higher than the number of sources $rank(\boldsymbol{A}) \geq N_M$. If we assume linear independence

of the mixtures and the sources, then this condition comes down to having $N_O \leq N_M$.

When a source separation problem has the same number of sources as mixtures $N_O = N_M$ it is labeled **determined**. If the mixture count is larger than the source count $N_O > N_M$ it is named **overdetermined**. When there are fewer mixtures than sources it is called **underdetermined**.

The most common formats for the audio signals in western pop/rock music are mono or stereo. This means we have $N_O = 1$ or $N_O = 2$. Additionally we can often expect to find 3 instruments or more playing simultaneously ($N_M > 3$). Therefore we can safely assume that in our work we are targeting an **underdetermined** problem.

Knowledge of the number of sources

As we have seen above the number of sources is a critical aspect for selecting a source separation method. Prior knowledge of the number of sources present in the mixture will play a fundamental role in the source separation process. Many speech enhancement applications, which can be viewed as source separation tasks, assume only two sources in the mixture: the speech and the noise. This assumption is rarely true for musical signals, since the number of instruments in a song is quite variable.

However an option that is often used is to separate the instruments one at a time or to separate a specific source given some *a priori* knowledge. In these situations knowing or estimating the number of sources in the mixture is not necessary. Instead we may estimate for each target instrument whether it is present or not and treat the mix as being composed of two sources, the target instrument and the rest.

2.6 Separation Conditions and Constraints

Apart from the properties that we have presented here, the source separation problem may be further qualified by special conditions or constraints on the separation process. These conditions and constraints are often due to the task or application as well as the availability of the data. In this section we will review some of the more common constraints and how they affect the choice of method.

Blind or Informed Separation

In the literature we often find a distinction between uninformed (or blind) and informed source separation. This distinction is based on the amount and type of knowledge we have about the sources being processed in a given task. **Informed** source separation refers to situations in which we have information about the specific sources that are being retrieved (Parvaix and Girin, 2011; Liutkus et al., 2011). On the other hand **blind** source separation (BSS) consists in retrieving sources about which we don't have any specific knowledge. We must note that even though we may not have information about the specific sources being retrieved, we may have prior knowledge or make assumptions about the nature of the target sources. For instance, if the task consists in separating speech signals we can make certain assumptions about their spectral structure or temporal properties, but if we do not have information about the specific speech signal that we are separating then it could be considered a blind source separation problem.

In this work we assume that we do not have any information about the specific sources being separated. However, we will make and use a significant number of assumptions about the nature and types of sources and mixtures based on the properties presented above.

Online

Another factor that will affect the method used to solve the problem is the way in which the mixtures are made available and the way we must output the separated sources. If the mixture signal is available in its totality at the time the separated sources are required we may apply methods that require the full signal. These type of methods are often referred to as **batch** source separation or batch processing.

On the other hand a more restrictive situation is when we must output the separated sources as the mixture signal becomes available. Methods doing this are often called **online** methods. An increasing number of applications in source separation impose this constraint. Interactive applications often impose this condition. In many interactive tasks the mixture signal does not have a determined duration and is continuously lengthening such as in machine listening and robotics. Another situation in which this type of processing is preferable is when memory is restricted or costly. With the increasing availability of mobile and memory-limited devices, online methods are often more advantageous and sometimes the only acceptable solutions.

Interest in online methods is growing as the number of recent publications on this topic demonstrate (Mairal et al., 2010; Lefevre et al., 2011; Marxer et al., 2011; Duan et al., 2012).

Low-latency

Even though online is a necessary constraint on methods used for interactive applications, it is not the only requirement. Every source separation method necessarily creates a small delay between input of the mixtures and output of the separated sources. This delay is called **latency**. Source separation tasks in the field of communication or hearing aids, often require a very low latency.

In the field of music this constraint may also appear when the user input requires immediate response from the system. The low-latency constraint also occurs when the mixture data is available only in streaming and buffering is limited. These situations are more and more common with the arrival of popular streaming music services.

Realtime

Finally another important constraint in interactive source separation tasks is that the computation of each block of mixture data must be performed in a deterministic amount of time. Furthermore the implementation of the method should compute the block of data faster than the presentation time of the data. In other words the implementation of the method should be fast enough to present the resulting data to the user as it is being computed. These constraints basically define the concept **realtime**.

Even though this study does not focus on the implementation details of the source separation methods, one of our goals is to limit the computational cost of solutions to facilitate implementations operating under realtime constraints. We also consider the realtime nature of the methods by studying whether they can be performed in a deterministic number of operations.

2.7 Target Scenario

In this section we present the target scenario that serves as context for this study. The main task in our work is to isolate the singing voice and the percussion or drums in commercially available western music audio recordings under low-latency and/or realtime constraints. We review the nature

and types of sources and mixtures we focus on as well as the constraints and restrictions we impose on ourselves. Finally, we present the reasons for these choices.

The choice of studying blind audio source separation of western popular music is motivated by the large number of such mixtures commercially available. Furthermore, our focus on low-latency methods is driven by the growing availability of audio in streaming distribution form and of embedded devices with strong memory restrictions. Given the diversity of western music, in terms of genre, styles and instruments, we decided to narrow our goals to a set of specific instruments and music styles. However, throughout this work we propose methods that are general enough to be applied to other instruments and styles.

The main reason for targeting singing voice and percussion is the fact that these instruments are present in most western popular music.

The choice of target instruments was made to cover a wide range of source properties. We chose the singing voice and percussion because they represent two very different types of sources. The singing voice produces a very harmonic narrowband signal while percussion instruments produce wideband signals highly localized in time. Singing voice sources have components that present deterministic properties and temporal and frequency domain correlations. Drum sources are dominated by stochastic signals, such as the sound of cymbals.

Additionally the singing voice has a very wide range of pitch values, pitch contour behaviors and timbres. These variations exist not only between different singers, but also between the singing voice and musical instruments or even between the different vocalizations and realizations of a given singer. Such a variety of parameters makes the task interesting for the development of techniques that can be used to isolate other harmonic instruments. Similarly, techniques used for the percussion section can be easily applied to other transient sounds such as the attacks of most instruments.

In the following sections we will present specific characteristics of these two instruments.

Singing Voice

The sounds produced by the human voice have been the focus of a great deal of research. Domains such as speech recognition and speech identification have become very important with the arrival of desktop computers and

mobile devices. The knowledge that we have gained about singing voice and speech production can be extremely useful in source separation scenarios where voice is one of the sources to be separated. This section presents a brief introduction to the mechanisms involved in human singing generation, a review of the different sounds that are produced in human singing, and a discussion of some of the signal characteristics of these sounds. Finally it introduces some of the mathematical models developed over the years which will be used in source separation techniques.

Human singing is the act of producing musical sounds with the human voice, which are the sounds made by humans using mainly the vocal folds(vocal cords). Even though humans are able to make sounds in many ways, the human voice refers to sounds generated using the vocal folds as the primary sound source. The mechanism for generating the human voice can be subdivided into three parts: the lungs, the vocal folds within the larynx, and the articulators. The lungs must produce adequate airflow and air pressure to vibrate vocal folds. The vocal folds are a vibrating valve that chops up the airflow from the lungs into audible pulses that form the laryngeal sound source. The muscles of the larynx adjust the length and tension of the vocal folds to modify the pitch and tone. The articulators are the parts of the vocal tract above the larynx consisting of tongue, palate, cheek, lips, etc. They are in charge of articulating and filtering the sound emanating from the larynx and to some degree can interact with the laryngeal airflow to strengthen it or weaken it as a sound source. This mechanism is capable of producing a large variety of different sounds (Titze, 1994; Stevens, 1998; Titze and Alipour, 2006).

As stated above, the vibration of the vocal folds is the process responsible for the most energetic component of the voice, because of the resonance in the vocal tract. However in human voice there are other components that do not require the vibration of the vocal cords. For instance, in some situations the airflow going through the space between the vocal cords produces a sound that also gets filtered by the vocal track. This component is often referred to as the breathiness of the singing voice since it is similar to the sound produced when breathing. Other components that do not require the vibration of the vocal cords are the sounds corresponding to the voiceless plosives (such as the English /p/, /t/, /k/, etc.) or the voiceless fricatives (such as the English /f/, /s/, etc.).

Depending on the domain of study, the different sounds produced by a human being during singing can be categorized in different ways: production

mechanism (vocal cord vibration or non-vibration), signal characteristics (quasi-stationary or non-stationary, wideband or narrowband, harmonic or non-harmonic), phonetically, timbrically, etc. Given our focus on source separation, we will concentrate on the signal properties of the different sounds present in the singing voice.

Harmonic components are usually present during vowels and voiced consonants. These components are mainly non-stochastic and narrow band. These components appear in the spectrum as periodic structures where most energy is concentrated in bins which are located at positions which are integer multiples of the fundamental frequency. These energy concentrations are also referred to as partials or harmonics. While harmonic components are present in most western music instruments, the ones present in singing voice are characterized by their flexibility. Different singing effects, such as vibratos, growl, pitch bends and glissandos, can significantly modify the spectrum of this component. The harmonic component of the voice is mainly due to the glottal excitation or vibration of the vocal cords, which occurs mainly during vowels that are predominant in singing voice and carry most of the musical information.

Another major part of the human voice consists of the wideband sustained component. This component is stochastic and smoothly modulated over time. It is often due to the breathiness in the voice. Breathiness is of special importance in western music where it is often used for aesthetic purposes. Breathiness is produced by the turbulent air that flows through the vocal folds at opening instants. Since this sound is filtered by the vocal tract it is highly correlated to the harmonic component. There are other voice production mechanisms, such as the fricative phonemes, that also generate wideband sustained component. However in this case of fricatives, since the turbulence is not produced at the vocal folds level, their sound is not filtered by the same filter as the glottal excitation of the harmonic component.

The human voice can also generate a large and varied set of wideband transient components. These are usually produced by sudden opening or closing of air channels by the use of the tongue, lips or other human speech organs. These components are characterized by their wideband and time localized nature.

There have been many works targeting the modeling and parametrization of the human voice. We will briefly introduce some that are of special interest for our work. The glottal source is probably one of the components that have drawn most attention over the years due to its fundamental role in

voice production. Rosenberg (1971) proposed several glottal pulse models created using inverse filtering. The best matching model is composed of 2 polynomial parts. Klatt and Klatt (1990) extended the previous model by adding a low pass filter to smooth the pulse shape and control the spectral tilt. The model was controlled by two main parameters: the opening quotient and the spectral tilt. Many other models have been proposed (Fujisaki and Ljungqvist, 1986; Fant et al., 1985; Fant, 1995; Doval et al., 2003), however their study is out of the scope of this work. The glottal aspiration noise produced by the turbulence of the airflow, responsible for a wideband stochastic component in human voice, has also been studied and modeled (Stevens, 1971; Flanagan, 1972; Liljencrants, 1985; Mehta and Quatieri, 2005). Finally the filtering parts of the human voice production system, such as the vocal tract and the radiation produced at the mouth and nostrils, have also received extensive research (Fujimura and Lindqvist, 1971; Flanagan, 1972; Maeda, 1982; Liljencrants, 1985; Lim and Lee, 1993; Niu et al., 2005).

In the context of our work, we are mostly interested in the modeling of the singing voice as a whole, taking into account all the processes involved. In the field of source separation the possibilities for estimating singing voice components from audio recordings are most interesting. Degottex (2010) and Degottex et al. (2011) proposed a glottal source and vocal tract model as well as the means for estimating its parameters from audio recordings of isolated singing voice.

Due to the importance of vocals in western music culture, the singing voice has become a widely targeted source in music transcription and separation tasks (Ryynänen, 2006; Li and Wang, 2007; Hsu and Jang, 2010b; Durrieu et al., 2010; Rafii and Pardo, 2012; Gómez et al., 2012).

Percussion

Percussion instruments have gotten a lot of attention in the Music Information Retrieval community. This focus is due to the role of rhythm in western music and to the fact that sounds produced by percussive instruments are significantly different from those produced by other instruments such as guitar, piano or saxophone. The percussion family is also believed to include the oldest musical instruments, following the human voice.

Percussion instruments are often categorized into two groups: pitched percussion instruments, which produce notes with an identifiable pitch, and unpitched percussion instruments, which produce notes without an identi-

fiable pitch. Most drums and cymbals belong to the latter category and are typically used to maintain or to provide accents, and their sounds are unrelated to the melody and harmony of the music. In western popular music percussion is commonly interpreted by a single player using a drum kit or drum set. A standard modern drum kit, as used in popular music, taught in many music schools and similar institutions consists of: a snare drum, a bass drum, a hi-hat stand and cymbals, one or more tom-tom drums and one or more cymbals.

In this work we mainly focus on unpitched percussion, since in most cases the separation of pitched percussion can be done the same way as for other pitched instruments or even the singing voice. From now on we will use percussion or drums to refer to unpitched percussion instruments.

From a signal point of view percussion or drums sounds are often highly characterized by their temporal evolution, usually with a short attack and different decay lengths depending on the instrument. Different drums present different frequency profiles. The bass drum presents a low frequency resonance, the tom-toms usually have different mid-band resonances, the hi-hat presents a high-band burst, and the snare-drum is a wideband short noise burst. Cymbals usually display a highly stochastic and slowly decaying wideband spectral component.

Over the years, there has been extensive work on modeling drums and the sounds produced by them (Rossing, 1990; Van Duyne and Smith, 1993; Trautmann et al., 2001; Marogna and Avanzini, 2009). Given the stochastic nature of drums they have usually been approached using physical models.

The task of analysis, recognition and classification of drums and other unpitched percussive instruments has also received much attention from the MIR community (Schloss, 1985; Herrera et al., 2002; Gillet and Richard, 2004; Yoshii et al., 2005b; Paulus and Virtanen, 2005; Tanghe et al., 2005; Hazan, 2005), in most cases using spectrotemporal pattern matching and adaptation techniques.

Finally, drums separation from music mixes has also become of special interest in the past decade (FitzGerald, 2004; Helén and Virtanen, 2005; Ono et al., 2008b; Schuller et al., 2009; Kim et al., 2011).

Chapter 3

Review of BASS Methods

In this Chapter we present the state of the art in Blind Source Separation (BSS) methods. We will be reviewing a wide spectrum of approaches, from the well-established statistical methods to more recent signal decomposition and factorization approaches. We will focus especially on methods used in low-latency and real-time conditions, as well as techniques that specifically target music mixes. There are many ways to categorize the different BSS techniques. From a filtering point of view we have methods that filter the estimated sources using multichannel time-invariant filtering and other methods that perform Time-Frequency (TF) masking. In the TF masking category we can further divide the methods into those that apply binary masking and those that use soft masks. In this work we organize the BSS methods categorized by the *a priori* information that they exploit. Section 3.1 first briefly introduces separation methods based on the statistics of the sources and the assumptions made about them. Section 3.2 presents the so called Beamforming techniques which exploit prior knowledge of the geometric configuration of the mixing process. Section 3.3 discusses using musical knowledge and information to aid the separation process. Section 3.4 covers methods based on signal decomposition techniques which exploit knowledge of signals' structure properties, especially important in music and audio source separation. Finally, Section 3.5 presents evaluation methods specifically developed to assess the performance of BSS methods.

3.1 Statistical Based Separation

The first efforts in Blind Source Separation (BSS) were focused mainly on using knowledge or assumptions about the statistics of the sources to perform the separation. These methods are important due to the weak assumptions they require, and have proven successful on synthetic tests and simulations. However they often rely on the problem being determined or overdetermined (see Section 2.5). In our target scenario we focus on musical signals with at most one or two mixtures and many sources and therefore these methods are not directly applicable.

Decorrelation Techniques

In cases where we have very little knowledge about the set of sources, one of the simplest assumptions that we can make is about the correlation of the sources.

One of the first methods to exploit the decorrelation of multiple sources was proposed by Weinstein et al. (1993). The authors also proposed the use of *a priori* knowledge about the sources to increase the performance of the method.

Belouchrani et al. (1997) proposed a method based on the time coherence of the sources. It relies only on stationary second-order statistics based on a joint diagonalization of a set of covariance matrices.

TDSEP is another BSS algorithm using time structure. Proposed by Ziehe and Müller (1998), this method is based on a set of time-delayed second-order correlation matrices. The unmixing matrix is built by performing a whitening step followed by the joint diagonalization of the whitened matrices.

Another example of decorrelation-based separation was proposed by Parra and Spence (2000). By assuming decorrelated second order statistics of the sources the authors show that the problem becomes a Singular Value Decomposition (SVD) task. Parra et al. (1998) and Parra and Sajda (2004) used decorrelation between sources in their unified framework of BSS by Generalized Eigenvalue Decomposition.

Independent Component Analysis

Independent Component Analysis (ICA) is one of the most popular BSS techniques. It is based on the assumption that the sources of the mixture are

statistically independent. Unlike Principal Component Analysis (PCA), the measure of independence used in ICA is not the variance (decorrelation) but rather higher order statistics. When employing the 3rd and 4th moments of the sources, they must be assumed to be non-Gaussian.

Several measures of independence have been proposed in the literature (e.g. skewness, kurtosis, mutual information, entropy, maximum likelihood). Usually the separation process is based on a gradient descent (or ascent) of these independence measures given the unmixing matrix as parameter.

The assumption of statistical independence between sources allows further exploiting other statistical properties in order to perform source separation. For instance the components being non-Gaussian leads to the well known Independent Component Analysis (ICA) method (Jutten and Herault (1991); Comon (1994)). The non-Gaussian assumption is necessary because ICA assumes that the mixtures will necessarily be more Gaussian than the components due to the central limit theorem.

On the other hand, different autocorrelations of the components lead to decorrelation methods (Weinstein et al., 1993). Finally, smoothly changing nonstationary variances of the sources can also be exploited using a log-likelihood maximization approach as in Pham and Cardoso (2001).

Similar or equivalent properties and assumptions can be found in a probabilistic or information theory framework. Bell Anthony J. and Sejnowski Terrence J. (1995) proposed an information theory framework to derive methods similar to ICA where the mutual information between the sources is exploited.

Cardoso (1998) reviewed the different approaches to the Blind Source Separation task under the assumption of mutual independence of the sources. The author explored the different higher order statistics or measures used in different separation algorithms and how these relate to each other. He also explored the different probabilistic interpretations of such methods.

Parra and Sajda (2004) showed that when further assumptions about the sources are made, the linear blind source separation problem is equivalent to a Generalized Eigenvalue Decomposition. The additional assumptions consist in considering independent sources non-Gaussian, non-stationary or non-white.

In audio these methods can be applied in the time domain (Time Domain ICA or TD-ICA) by interpreting the waveform of the sources as random

variables. However these methods often require the availability of at least as many mixture signals as sources. In the case of music, this is rarely possible, since music is mostly available in stereo format and more than 2 instruments are often present in the mix. As an alternative, we can assume a certain degree of time-frequency disjointness, and apply these methods in the frequency domain (Frequency Domain ICA or FD-ICA). This is done by treating the spectral bins as independent random variables. In this case an additional step is necessary where we must group together the estimated sources of each of the bins that belong to a given instrument. This step is needed due to the permutation and scaling ambiguity of ICA methods.

3.2 Beamforming Techniques

In some situations we may be dealing with live signals and/or we may have access to the capturing process and to the sensors. In such cases there is a set of assumptions that can be exploited to enhance the separation process. In a musical context this is often the case with multiple microphone studio mixtures or the recording of live concerts.

The use of geometrical assumptions about the sources and sensors has produced a set of methods known as beamforming techniques. These make use of interferences between the mixture signals to filter the sources coming from some specific direction or position.

Degenerate Unmixing Estimation Technique (DUET)

Probably one of the simplest methods based on geometrical assumptions about the sources/mixtures is the Degenerate Unmixing Estimation Technique (DUET) method. It classifies the bins of the spectrum depending on the panning (magnitude ratio) and phase difference between two given channels. This method is especially interesting for commercial stereo music.

Jourjine et al. (2000) proposed a mixture model where we consider the measurements of the direct path from N_M sources to 2 sensors. If we dismiss the absolute position of the sources, the attenuation and delay parameters of one of the mixtures can be absorbed into the definition of the source. In other words, the mixture of one of the two channels is considered as the simple addition of all the sources. The mixture in the other source is the sum of the sources weighted and delayed. This leads to a simplified version

of the attenuated and delayed mixture model presented in Section 2.3:

$$v_1[t] = \sum_{m=1}^{N_M} x_m(t) + \boldsymbol{B}_1[t] \tag{3.1}$$

$$v_2[t] = \sum_{m=1}^{N_M} g_m x_m[t - d_m] + \boldsymbol{B}_2[t] \tag{3.2}$$

where $\boldsymbol{B}_o[t]$ is the noise at the o^{th} sensor (Yilmaz and Rickard, 2004) and (Vinyes et al., 2006).

This bin classification technique has been extended to other measures such as the frequency value of each bin in order to separate non-pan-disjoint sources (Li and Wang, 2008).

Time Difference of Arrival (TDOA) Estimation

These types of methods use geometric assumptions about the relative positions of sources and sensors and about the traveling sound wave in order to perform the separation. These methods can be split into two different steps. The first step consists in estimating the relative location or direction of the source. The second step uses the estimated location to perform the separation.

The first step, also referred to as source localization, has been widely studied over the years due to its wide range of applications. Knapp and Carter (1976) proposed the Generalized Cross-Correlation (GCC) approach, a Maximum Likelihood estimator for determining time delays between two signals by prefiltering them before computing correlation. Reed et al. (1981) proposed a method called Least Mean Squares (LMS) Adaptive Filter. The authors estimated the time delay by optimizing the coefficients of a Finite Impulse Response (FIR) filter that minimizes the mean square difference between the two inputs. Carter (1987) proposed using the Magnitude Square Coherence (MSC) function to estimate the bearing, range and position of a source given the response of a sensor array. The coherence function between two random processes is defined as the crosspower spectrum divided by the square root of the product of the two auto power spectra. Omologo and Svaizer (1997) introduced a variation of the GCC method where the correlation prefiltering is chosen to be the inverse of their respective magnitude. Called Crosspower Spectrum Phase (CSP), it leads to a sharpening of the cross correlation peaks, which in turn achieves better results, especially under low SNR conditions.

Cross-spectrum

An interesting set of methods to perform beamformer-based audio source localization and separation comes from the field of robotics. Robotics methods are often required to be online and realtime. Even though these methods may not be directly applicable to musical signals, they are of special interest in our work due to the low-latency constraints they assume.

One of the beamforming techniques coming from robotics relevant to our work was proposed by Valin et al. (2004). The method is based on the frequency domain approximation of the delay-and-sum beamformer:

$$r_{\boldsymbol{\tau}}[t] = \sum_{o}^{N_O} v_o[t - \tau_o] \tag{3.3}$$

Assuming a single source the energy of this measure over a given length L will be maximal when the delays τ_o are set in such a way that the mixture signals v_o are in phase. The energy of this measure can be seen as a localization spectrum:

$$E_{\boldsymbol{\tau}}[t] = \sum_{l=0}^{L-1} \left(\sum_{o}^{N_O} v_o[t + l - \tau_o] \right)^2 \tag{3.4}$$

Expanding the energy expression we appreciate an almost constant term which will not play a role in the maximization $K = \sum_{l=0}^{L-1} \sum_{o}^{N_O} v_o^2[t - \tau_o]$ and a variable term containing the cross factors:

$$E_{\boldsymbol{\tau}}[t] = K + 2 \sum_{l=0}^{L-1} \sum_{o_1}^{N_O} \sum_{o_2}^{o_1-1} v_{o_1}[t + l - \tau_{o_1}] v_{o_2}[t + l - \tau_{o_2}] \tag{3.5}$$

which can be rewritten in terms of the cross-correlation as:

$$E_{\boldsymbol{\tau}}[t] = K + 2 \sum_{o_1}^{N_O} \sum_{o_2}^{o_1-1} R_{v_{o_1}, v_{o_2}}[\tau_{o_1} - \tau_{o_2}] \tag{3.6}$$

This approach works correctly under the assumption of a single source. When multiple sources are present, the frequency domain version allows computing this measure for each frequency bin, for which the single source assumption has more chances to hold true, given the disjointess of the sources in this representation domain:

$$R_{o_1 o_2}(\tau, t) \approx \sum_{l=0}^{L} \boldsymbol{V_{O_1}}(t + l) \boldsymbol{V_{O_2}}(t + l)^* e^{j 2\pi k \tau / L} \tag{3.7}$$

where $\boldsymbol{V_O}$ is the Discrete Fourier Transform of $v_o[t]$ and $(\cdot)^*$ denotes the complex conjugate. Other benefits of the frequency domain approach are less computational complexity and the possibility of weighting the different bins when computing their contribution to the localization spectra which leads to an enhanced cross-correlation measure:

$$R^e_{o_1 o_2}(\tau, t) = \sum_{l=0}^{L} \frac{\zeta_{o_1}(t+l)\boldsymbol{V_{O_1}}(t+l)\zeta_{o_2}(t+l)\boldsymbol{V_{O_2}}(t+l)^*}{|\boldsymbol{V_{O_1}}||\boldsymbol{V_{O_2}}|} e^{j2\pi k\tau/L} \quad (3.8)$$

This use of the cross-spatial correlation has often been incorporated in multiple sensor beamforming source localization/separation. The definition of the weight ζ_o is often what varies among methods. However another interesting contribution of Valin et al. (2007) is how the source tracking is performed using the proposed beamforming spectrum. While most tracking techniques in audio and music analysis are based on Hidden Markov Models (HMM) and dynamic programming approaches such as Viterbi algorithms (Durrieu et al., 2010), Valin et al. (2007) proposed using a sequential Monte Carlo method (SMC) or particle filtering to track the position of the sources. This technique is a model estimation method based on simulation. It assumes a Bayesian model with latent variables connected in a Markov chain. It is similar to HMM, however the state space of the latent variables is typically continuous rather than discrete.

The authors of this work propose a set of possible mappings from candidates to source tracks. The candidates are computed as the peaks of the beamforming spectra. The source track assignment can be selected from a set of possible simultaneous tracks. Additionally a source track may exist but be inactive or be active but unobserved. Finally the latent state space is defined by the speed and position of the sources and an excitation-damping dynamics model is used in order to predict next states from current states. This configuration allows for a low latency estimation of the current simultaneous source tracks. It also allows a flexible framework for assigning probabilities to the different source behaviors, such as source activation/deactivation or start/end of source track probability.

MUSIC and MVBF

Another approach to the TDOA estimation problem consists of exploiting a subspace of the spatial cross-correlation matrix between the spectra of the channels.

Multiple Signal Classification (MUSIC) MUSIC (Multiple Signal Classification) is a method for performing sound source localization when various sensors are available. This type of method stems from the assumption that the mixing process can be modeled as in Section 2.3. The process exploits the phase delays and magnitude attenuations at the different sensors from the different sources. In the frequency domain the input at the sensors can be modeled as:

$$\boldsymbol{v}(\omega,t) = [\boldsymbol{V}_1(\omega,t) \ \ldots \ \boldsymbol{V}_{N_O}(\omega,t)]^\top = \boldsymbol{A}_k\boldsymbol{x}(\omega,t) + \boldsymbol{n}(\omega,t) \tag{3.9}$$

where the vector $\boldsymbol{x}(\omega,t)$ represents the sources' outputs as:

$$\boldsymbol{x}(\omega,t) = [\boldsymbol{X}_1(\omega,t) \ \ldots \ \boldsymbol{X}_{N_M}(\omega,t)]^\top \tag{3.10}$$

The vector $\boldsymbol{n}(\omega,t)$ corresponds to the noise. In this case the noise refers to the non-directional or less-directional components of the signal. The matrix $\boldsymbol{A}_k$ is the transfer function matrix, whose element (o,m) is the transfer function of the k^{th} bin from the m^{th} source to the o^{th} sensor.

The first step consists in calibrating the system beforehand by recording the impulse responses or computing the transfer function from each possible source to each sensor. Depending on the application and the assumptions that can be made about the sources, different sets of parameters will be used to characterize each possible source. In the domain of audio source separation the parameters are usually spatial properties of the sources. For instance if the sensors are close together in relation to their distances from the sources, we can safely assume the sound to act as a plane wave. Under this assumption and disregarding the attenuation of the wave, the sources can be characterized by just their angles of incidence or directions to the sensor array. In these situations the space of possible source parameters would be defined by (θ,ϕ) where θ and ϕ are the elevation and azimuth angles of incidence from the source respectively. In other situations, for example when the sources are near the sensors, the distance of the source r to the sensor array may also be used. This happens when the sound wave cannot be considered as a plane wave but rather as a spherical wave. In all cases the parameter space of the sources is discretized into a finite set of parameter combinations. For simplicity, from now on we consider the case where the sources are parametrized with (θ,ϕ), however the method can easily be generalized to other parameterizations of the sources. The calibration results in a set of vectors $\boldsymbol{a}_\omega(\theta,\phi)$, called location vectors which will be used later in the calculation of the spatial spectrum.

The localization process is performed by taking the STFT at each sensor and calculating a spatial correlation matrix of the bins of these multichannel spectra:

$$\boldsymbol{R}_{vv}(\omega, t) = E\left[\boldsymbol{v}(\omega, t)\boldsymbol{v}^H(\omega, t)\right] \tag{3.11}$$

In order to have a stable estimation of the correlation matrices we must average along H frames in the time dimension:

$$\overline{\boldsymbol{R}}_{vv}(\omega, t) = \sum_{h=-H/2}^{H/2} \boldsymbol{v}(\omega, t+h)\boldsymbol{v}^H(\omega, t+h) \tag{3.12}$$

In order to simplify the following explanation, we will focus on a single frame t and the averaged matrix correlation will be referred to as $\boldsymbol{R}_\omega = \overline{\boldsymbol{R}}_{vv}(\omega, t)$

The method proceeds by computing the eigenvalue decomposition of the correlation matrices:

$$\boldsymbol{R}_\omega = \boldsymbol{E}_\omega \boldsymbol{\Sigma} \boldsymbol{E}_\omega^{-1} \tag{3.13}$$

The eigenvectors are split as $\boldsymbol{E}_\omega = [\boldsymbol{E}_\omega^x | \boldsymbol{E}_\omega^n]$ where $\boldsymbol{E}_\omega^x$ and $\boldsymbol{E}_\omega^n$ denote the sets of eigenvectors corresponding to the N_M dominant eigenvalues and the rest of the eigenvalues, respectively.

We use $\boldsymbol{E}_\omega^n$ and the transfer function matrix $\boldsymbol{A}_\omega$ to the calculate the MUSIC spatial spectrum:

$$\boldsymbol{P}(\theta, \phi, \omega) = \frac{1}{\left|\tilde{\boldsymbol{a}}_\omega^H(\theta, \phi)\boldsymbol{E}_\omega^n\right|^2} \tag{3.14}$$

where $\tilde{\boldsymbol{a}}_\omega(\theta, \phi)$ is the normalized location vector for the scanning point (θ, ϕ) defined as:

$$\tilde{\boldsymbol{a}}_\omega(\theta, \phi) = \frac{\boldsymbol{a}_\omega(\theta, \phi)}{\|\boldsymbol{a}_\omega(\theta, \phi)\|} \tag{3.15}$$

The MUSIC spatial spectrum gives us a distribution for each bin over the parameter space. This spectrum is then averaged over a frequency range in order to get a spatial spectrum:

$$\overline{\boldsymbol{P}}(\theta, \phi) = \frac{1}{\omega} \sum_{\omega_L}^{\omega_H} \boldsymbol{P}(\theta, \phi, \omega) \tag{3.16}$$

where ω_L and ω_H are the indices for the lower and upper boundaries of the frequency range, and $\omega = \omega_H - \omega_L + 1$.

The location of the parameter values (in our case the direction of arrival) of the sources can be estimated from the peak locations of the spatial spectrum $\overline{\boldsymbol{P}}(\theta, \phi)$.

This method was first proposed by Schmidt (1986). These kinds of methods are also called subspace methods.

Asano et al. (1999, 2001) successfully applied this method to automatic speech recognition under real-time conditions. This first use of the MUSIC technique in the context of robotics audition was followed by many variations focusing mainly on speech signals.

In this study the authors also proposed a way to locate more sources, as well as a method to estimate the number of sources. This method consists in performing a preliminary estimation of the source count and location from the narrow band MUSIC spatial spectrum $\boldsymbol{P}(\theta, \phi, \omega)$ before aggregating it into the broadband spatial spectrum $\overline{\boldsymbol{P}}(\theta, \phi)$.

Minimum Variance Beamforming (MVBF) Once the source localization has been performed, the next step is to separate it from the mixture. The MUSIC source localization technique is often coupled with the Minimum-Variance Beamformer (MVBF) technique for source separation developed by Johnson and Dudgeon (1993). This consists of recovering the spectrum of the m^{th} source by filtering the multichannel spectrum in the following way:

$$\hat{\boldsymbol{X}}_m(\omega, t) = \boldsymbol{w}^H(\omega)\boldsymbol{v}(\omega, t) \tag{3.17}$$

where

$$\boldsymbol{w} = \frac{\boldsymbol{R}_\omega^{-1}\hat{\boldsymbol{a}}_{m,\omega}}{\hat{\boldsymbol{a}}_{m,\omega}^H \boldsymbol{R}_\omega^{-1}\hat{\boldsymbol{a}}_{m,\omega}} \tag{3.18}$$

The vector $\hat{\boldsymbol{a}}_{m,\omega}$ is the location vector of the m^{th} source estimated in the source location step previously presented.

There exist several variations of this method. The different approaches vary in the use of the correlation matrix in the calculation of the filter $\boldsymbol{w}$. Asano et al. (2001) showed the use of two other methods named MV1 and MV2.

In MV1 the filter is computed as:

$$\boldsymbol{w} = \frac{\boldsymbol{K}_\omega^{-1}\hat{\boldsymbol{a}}_{m,\omega}}{\hat{\boldsymbol{a}}_{m,\omega}^H \boldsymbol{K}_\omega^{-1}\hat{\boldsymbol{a}}_{m,\omega}} \tag{3.19}$$

where $\boldsymbol{K}_\omega$ is the spatial correlation matrix computed during a period when the target source $\boldsymbol{X}_m(\omega, t)$ is not present in the mixtures. This method largely reduces the presence of noise in the estimation, however it requires some means of detecting the absence of the target source in the mixture.

The other variant, named MV2 consists in defining the filter as:

$$\boldsymbol{w} = \frac{\boldsymbol{Q}_\omega^{-1}\hat{\boldsymbol{a}}_{m,\omega}}{\hat{\boldsymbol{a}}_{m,\omega}^H \boldsymbol{Q}_\omega^{-1}\hat{\boldsymbol{a}}_{m,\omega}} \tag{3.20}$$

where

$$\boldsymbol{Q}_\omega = \hat{\boldsymbol{A}}_\omega \hat{\boldsymbol{A}}_\omega^H + \gamma \boldsymbol{I} \tag{3.21}$$

The matrix $\hat{\boldsymbol{A}}_\omega = [\hat{\boldsymbol{a}}_{1,\omega} \ldots \hat{\boldsymbol{a}}_{m,\omega}]$ is built using the estimates of the source locations. The γ parameter controls the directivity of the beamformer filter. The larger the value of γ is, the larger the reduction of directional interference will be. A smaller value of γ will result in less reduction of non-directional noise. This method does not require large amounts of data since it does not rely on the correlation matrices. This reduces even more the latency of the algorithm. However it is more sensitive to errors in the localization of the sources.

The MUSIC and MVBF methods are widely used in the fields of robotic audition and conferencing. They present very interesting low-latency and realtime properties, and perform quite well on speech signals even sometimes with the presence of environmental sounds. However in the context of music signals, their application has been less successful. One of the main reasons is that the geometric assumptions about the sources' positions and the mixing process do not apply to many musical signals, especially musical recordings mixed and mastered in studios, which comprise the majority of commercially available western music. In studio mixes, the different instruments are often recorded separately and then mixed artificially without necessarily respecting the traveling sound wave models. In many cases, to achieve the sensation of different locations of the instruments in the mix, panning and reverb effects are used. These effects invalidate the assumptions of the MUSIC and MVBF methods, significantly degrading their performance.

Another shortcoming of these methods in music is the ratio between the number of sources and mixtures. In musical signals we often have many more sources than mixtures, and in these cases subspace methods do not work. Even though some solutions have been proposed when there are more sources than mixtures, these depend on high disjointness in the time-frequency representation of the signal.

3.3 Music-specific Signal Model Methods

The solutions based only on statistical knowledge of the sources and/or the mixing process often require that the number of mixture signals be larger than or equal to the number of sources (determined or overdetermined problem). This condition can sometimes be relaxed if we make some assumptions about the individual sources themselves.

Several techniques make use of *a priori* information about the actual sources which may be extracted from the mixture signals. This situation is quite common in music sources separation. The extensive research in Music Information Retrieval (MIR) has been applied to extract information about the target sources in order to simplify the separation.

Pitch Estimation and Tracking

Pitch is a perceptual property that allows the ordering of sounds on a frequency-related scale (Klapuri, 2004). The technical term for this property is fundamental frequency (f_0). Pitch is a major auditory attribute of musical tones, along with duration, loudness, and timbre. Pitch may be quantified as a frequency, but pitch is not a purely objective physical property; it is a subjective psychoacoustical attribute of sound. Historically, the study of pitch and pitch perception has been a central problem in psychoacoustics.

The estimation of pitch is highly related to source separation in the field of music. Many music source separation methods use pitch estimation as a previous step. Additionally pitch estimation approaches often share the same techniques as those of source separation. Therefore we briefly review here some pitch estimation techniques proposed in the literature.

In the field of pitch estimation several tasks are often differentiated. Monophonic pitch estimation consists in estimating the pitch line of an audio recording where a single pitched sound is present at any given time. It

has often been used in speech analysis, audio effects or transformations and other applications in which the individual instrument tracks are available. Predominant pitch, bass line or melody estimation often refers to the estimation of one of the pitch lines in a polyphonic recording, where the selection of the pitch line depends on the application. Multiple pitch estimation consists in extracting all the pitch lines in a polyphonic recording. These two last families of methods are the ones of interest in the field of source separation.

Probably the most intuitive and well-known method for pitch estimation of monophonic signals consists in computing the Auto Correlation Function (ACF) and finding the second highest peak (Rabiner, 1977). In the literature we find multiple variations of this method, such as de Cheveigné (1991) computing the ACF per band and aggregating the results. de Cheveigné and Kawahara (2002) propose a fast computation based on the DFT. Another relevant monophonic pitch estimation method is the cepstrum (Noll, 1967) which consists in performing the DFT of the logarithm of the spectrum. This calculation allows us to decouple the filter component from the pitched source when assuming a smooth filter source-filter model. Finally the so called harmonic summation or Harmonic Product Spectrum (HPS) methods (Schroeder, 1968) are based on a histogram that counts the contribution of each spectral peak to the pitch candidates that are common divisors of its frequency.

The methods for monophonic pitch estimation usually present limitations when applied to polyphonic signals. These are mostly due to not taking into account the overlapping partials and background spectral noise when multiple instruments are present. Several authors have proposed variations of the monophonic methods to overcome these problems, including heuristics to reduce the influence of other instruments (Goto and Hayamizu, 1999; Klapuri, 2003; Salamon and Gomez, 2012). These methods often use heuristics such as iterative spectral substraction or pitch contour-based selection in order to achieve satisfactory results in single pitch estimation of polyphonic recordings.

Most recent research on the task of multipitch estimation has been focusing on generative models (Yeh et al., 2010) or signal decomposition approaches. Deconvolution in the logarithmic frequency domain (Sagayama et al., 2004; Saito et al., 2008) and factorization techniques such as NMF (Smaragdis and Brown, 2003; Kameoka et al., 2007; Raczyński et al., 2007) and PLCA (Smaragdis, 2011; Benetos and Dixon, 2011b, 2013) have been widely used

for this purpose. Semi-supervised techniques that exploit *a priori* knowledge about the instrument timbres (Quesada et al., 2010) have also been used to perform multipitch estimation.

Li and Wang (2008) use pitch annotations to create a binary time frequency mask that exploits the harmonic structure of pitched instruments for separation purposes. Hu and Wang (2004) and Li et al. (2009) use the estimated pitch and correlations between the amplitude modulations and phase evolutions of the partials to separate monaural mixes of monophonic instruments. Durrieu et al. (2011) applies predominant pitch-based constraints to the factorization of a source-filter plus accompaniment spectrum model. The predominant pitch estimation and the separation are both performed using NMF and the same spectrum model. Carabias-Orti et al. (2011) and Rodriguez-Serrano et al. (2012) assume instrument-specific harmonic envelopes that, once learned from isolated instrument recordings, can be used to perform multipitch estimation and separation. The pitch estimation and instrument separation is performed using NMF on source-filter spectrum models.

Transients

In music, there is no strict definition of what a transient is, but it usually refers to sudden changes in the statistics of the signal. Note onsets or high-amplitude, short-duration sounds that occur at the beginning of a steady state waveform are commonly referred to as transients. Transients are found in many musical sources. Percussive instruments or the note attacks of certain instruments concentrate a lot of their energy in a localized time region. These regions can be exploited in source separation tasks. Estimating transient locations has been the target of many MIR studies. Transients are especially useful in drums isolation or removal. Most transient estimation algorithms are based on an onset (or transient) detection function, which is a one dimensional function that represents the saliency or intensity of change in the input signal. In a post-processing stage a discretization algorithm is applied to the onset detection function in order to detect and localize the transients. Many onset detection functions have been proposed. The spectral flux introduced by Masri (1996) consists in calculating the change rate of the power of the spectrum by means of the STFT by comparing the current frame to the previous one. Duxbury et al. (2002) and Bello et al. (2005) studied the use of the L_1-norm and L_2-norm when comparing the frames. In some cases logarithmic magnitudes (relative or normalized) of

the STFT frame are used in order to make the onset detection function less dependent on the global energy of the signal (Klapuri, 1999).

The High Frequency Content (HFC) method (Masri, 1996; Brossier et al., 2004) applies a linear weighting when comparing frames, emphasizing the higher range of the spectrum, usually associated with percussive sounds. Duxbury et al. (2003) and Bello and Sandler (2003) proposed the use of spectral phase time-derivatives in order to compute an onset detection function that relates to large changes in the instantaneous frequency. Phase deviation methods have also been used in combination with magnitude changes (Dixon, 2006). In order to capture harmonic transients, Macleod and Malcolm (2003) proposed the use of the Kullback-Leibler divergence between spectral frames.

Röbel (2003, 2005, 2009) proposed an onset detection function related to the center of gravity of the spectrum in the analysis window. The computation of the center of gravity is based on the spectral reassignment work done by Auger and Flandrin (1995).

Some of these transient detection techniques have been used as a preprocessing step in many drum transcription and separation methods (Gillet and Richard, 2004; Barry, 2005; Yoshii et al., 2005a; Gillet, 2007; Gillet and Richard, 2008). In other cases the transient estimation and transcription steps are performed jointly by performing pattern matching and adaptation (Zils et al., 2002; Yoshii et al., 2005a).

Beat Detection and Tracking

As explained in Section 2.4 music sources are characterized by their high temporal correlations. The beat, the basic unit of time in music, has received special attention for decades. The beat level is a metric level of rhythm that is often used as a reference.

Automatic beat analysis has often been done in Music Information Retrieval (MIR) and it comprises two main tasks: beat detection and beat tracking. Beat detection consists in finding the beat rate of an excerpt, while beat tracking consists in finding the actual beat positions. In blind source separation tasks, beat positions can be of great help for isolating or separating the rhythmic section of the music mixture.

Goto and Muraoka (1994) developed a real-time algorithm for beat detection and tracking based on detected onsets and multiple agent architecture. Smith (1996) proposed the use of wavelets and phase congruence to find

the beat positions and account for beat variations. Dixon (1997) derived a beat estimation algorithm based on onset times and inter-onset intervals. Scheirer (1998) estimated the beat by creating a small number of bandpass filters coupled to banks of parallel comb filters. This method is causal and able to predict future beat positions, allowing it to work under low latency conditions. Desain and Honing (1999) created a rule-based method to detect and track the beat and other higher level metric structures. Lang and de Freitas (2004) attempted a probabilistic approach to the beat tracking problem. Finally Hazan et al. (2007) performed beat tracking and prediction under real-time and low-latency constraints.

Identifying the beat has recently been used to separate the melody from the background music. The work done in the MIR field has led to recent methods of regularity-based source separation. These methods mainly focus on the separation of the accompaniment or background music from the main melody or lead voice. The accompaniment often presents a highly repetitive structure while the main melody tends to be less predictable.

Rafii and Pardo (2011) proposed a method named Repeating Pattern Extraction Technique (REPET) based on a model for the repeating segments in a music excerpt. The beat rate and positions of the music piece are first estimated and then the musical signal is segmented into repeating sections and an average of the repeating segments is taken to create the model of the repeating segment. Finally a binary mask is computed from the similarities/dissimilarities of frequency bins to the model. This binary mask is applied to the original spectrogram and the inverse STFT is performed to resynthesize the audio waveform.

Liutkus et al. (2012) developed an extension of the REPET technique. REPET is based on the beat rate and the repeating model of the entire music piece. Liutkus et al. (2012) proposed an adaptive method that allows variations in the beat and in the repeating model. The goal of this method is to correctly handle the different sections of the music piece, such as verse and chorus. Additionally the authors used a soft mask in contrast to a binary mask, which should lead to less musical noise due to smoother frequency domain filters and in consequence lower temporal aliasing.

Kameoka et al. (2012) proposed the use of constraints on Non-negative Matrix Factorization to exploit this type of musical information. The authors proposed a constraint specifically tailored to capture the beat structure of the music mixture in the gains of the factorization.

Score Informed

In cases where the score of a given musical piece is known, we may use this signal-specific information to greatly increase the performance of existing source separation methods. Woodruff et al. (2006) combine prior information from an aligned score to improve the performance of a music source separation system based on panning and time-frequency masking. The system is evaluated on synthetic music performances with and without score alignment to assess its influence in the separation. Raphael (2008) proposes a time-frequency and panning classification-based system that relies on aligned score information.

Due to the widespread availability of musical transcriptions of popular music, these techniques are lately drawing more attention from the scientific community. Ganseman et al. (2010) use score synthesis to initialize a signal decomposition monaural source separation system. The score is aligned to the music recording using Dynamic Time Warping (DTW) and then used to create the priors of a Probabilistic Latent Component Analysis (PLCA) model. Similarly, Hennequin et al. (2011) use the musical score to perform an initialization of a parametric Nonnegative Matrix Factorization (NMF) of the mixture spectrum.

Ewert and Müller (2011) and Ewert and Müller (2012) use the score to impose activity and harmonicity constraints on the components of a factorization-based source separation method. This approach is applied to piano recordings and allows the authors to achieve high quality separation results. The score of the pieces must be previously globally aligned to the recording and the constraints must be loose to permit local misalignments. Bosch et al. (2012) present a method in which Musical Instrument Digital Interface (MIDI) scores improve the lead voice pitch estimation of audio recordings. The estimated pitch is then used in a singing voice separation task. As in the previous work the score must be first globally aligned to the recording. The aligned score is then used to derive the state probabilities of the Hidden Markov Model (HMM) used in the pitch tracking stage.

3.4 Signal Decomposition Approaches

The music signal model-based approaches are very practical since we are able to easily reuse much of the work done in Music Information Retrieval (MIR). On the other hand they are very dependent on the model itself, and mismatch between the data and the model can be the source of artifacts

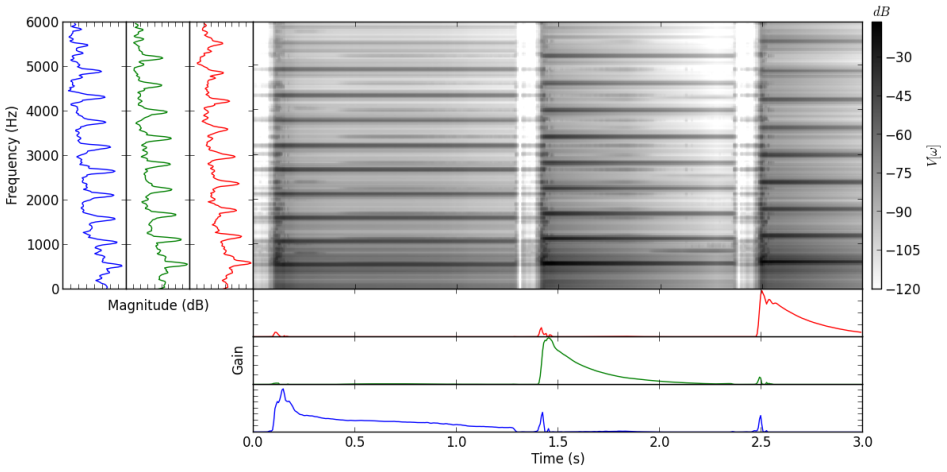

Figure 3.1: Example of the NMF decomposition of 3 piano notes, with the spectrogram of the mixture $\boldsymbol{V}$ *top right*, 3 basis components $\boldsymbol{W}$ *top left*, gains $\boldsymbol{H}$ *bottom* of the 3 components.

and/or noise. Alternatives to these types of solutions are those based on signal decomposition methods. These methods are based on finding the components that form a given signal and mixing them into different groups so that an estimate of the sources can be computed.

Non-negative Matrix Factorization (NMF)

Currently some of the most popular methods for performing audio source separation use Non-negative Matrix Factorization (NMF). The main assumption in NMF based methods is that the spectra (magnitude or power of the STFT) of the audio mixture signal, $\boldsymbol{V}$ can be modeled as linear combinations of N_W elementary non-negative spectra (also called basis components). This can be expressed as $\boldsymbol{V} \approx \hat{\boldsymbol{V}} = \boldsymbol{W}\boldsymbol{H}$ where $\boldsymbol{V}$ is the observed spectrum of the mixture signal at a given frame and $\hat{\boldsymbol{V}}$ is the modeled spectrum. $\boldsymbol{W} \in \mathbb{R}^{N_\omega \times N_W}$ is the matrix whose columns are the basis components, it is also referred to as the basis matrix. $\boldsymbol{H} \in \mathbb{R}^{N_W \times N_T}$ is a vector of component gains for the current frame.

The factorization of a positive matrix into two non-negative factor matrices was first proposed by Paatero and Tapper (1994) and fast convenient algorithms were developed by Lee and Seung (2001). The algorithms proposed by Lee and Seung (2001) consisted of two rules based on the gradient

descent formulation. When using a specific step, the update steps become multiplicative rules under which the targeted cost functions were shown to be non-increasing. The objective functions they studied were the Euclidean distance and I-divergence [1]:

$$\Phi_{euc}(\boldsymbol{H},\boldsymbol{W}) = \sum_{t=1}^{N_T}\sum_{\omega=1}^{N_\omega} \frac{1}{2}\left([\boldsymbol{WH}]_{t,\omega} - [\boldsymbol{V}]_{t,\omega}\right)^2 \tag{3.22}$$

$$\Phi_{kl}(\boldsymbol{H},\boldsymbol{W}) = \sum_{t=1}^{N_T}\sum_{\omega=1}^{N_\omega} [\boldsymbol{V}]_{t,\omega} log\frac{[\boldsymbol{V}]_{t,\omega}}{[\boldsymbol{WH}]_{t,\omega}} - [\boldsymbol{V}]_{t,\omega} + [\boldsymbol{WH}]_{t,\omega} \tag{3.23}$$

where $[X]_k$ is the k^{th} element of vector X.

The proposed update rules for the Euclidean distance cost function are:

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \otimes \frac{\boldsymbol{W}^\top \boldsymbol{V}}{\boldsymbol{W}^\top (\boldsymbol{WH})} \tag{3.24}$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \otimes \frac{\boldsymbol{H}^\top \boldsymbol{V}}{(\boldsymbol{WH})\,\boldsymbol{H}^\top} \tag{3.25}$$

where $\otimes$ is the Hadamard product (an elementwise multiplication of the matrices) and all divisions are elementwise. $\boldsymbol{H}$ and $\boldsymbol{W}$ are initialized as random positive matrices.

For the Kullback-Leibler divergence cost function, the update rules are:

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \otimes \frac{\boldsymbol{W}^\top \frac{\boldsymbol{V}}{(\boldsymbol{WH})}}{\boldsymbol{W}^\top \mathbf{1}} \tag{3.26}$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \otimes \frac{\frac{\boldsymbol{V}}{(\boldsymbol{WH})}\boldsymbol{H}^\top}{\mathbf{1}\boldsymbol{H}^\top} \tag{3.27}$$

where $\mathbf{1}$ is a matrix of ones with the same size as $\boldsymbol{V}$.

The resulting algorithm to perform the non-negative matrix factorization can be resumed as:

where the initialization matrices $\boldsymbol{W}_0$ and $\boldsymbol{H}_0$ are commonly random positive matrices. The normalization step is optional, however it is often performed to avoid the scale ambiguity of the solution.

1. They are also referred to as generalized Kullback-Leibler divergence or simply Kullback-Leibler if the basis and gains sum to 1.

Initialize the basis $\boldsymbol{W} = \boldsymbol{W}_0$
Initialize the gains $\boldsymbol{H} = \boldsymbol{H}_0$
while continue criterion **do**
  Compute the estimation $\hat{\boldsymbol{V}} = \boldsymbol{W}\boldsymbol{H}$
  Update the basis $\boldsymbol{W}$
  Update the gains $\boldsymbol{H}$
  Normalize the basis $\boldsymbol{W}[\omega, w] = \boldsymbol{W}[\omega, w] / \sum_i^{N_\omega} \boldsymbol{W}[i, w]$
end while

Smaragdis and Brown (2003) were the first to apply this factorization method to music in the context of polyphonic music transcription. They focused on music passages comprised of notes exhibiting a harmonically fixed spectral profile, such as piano notes. This allowed them to express the spectrogram of the signal as a linear basis transformation and use non-negative matrix decomposition techniques to find the spectral basis and the activations over time of this basis.

NMF Divergences

While the objective functions (also called loss functions) based on Euclidean distance and Kullback-Leibler divergence have served well in many applications, there has been a continuous search to find divergences that better suit each specific task and/or type of data. For instance if the error between the model and the data is Gaussian, then the Euclidean distance will be the most appropriate divergence. If we are dealing with histograms or probability distributions the Kullback-Leibler divergence is a better choice.

Cichocki et al. (2006, 2008) developed update rules for the family of Amari's α divergences (Amari, 1985):

$$\begin{aligned} \Phi_\alpha(\boldsymbol{H}, \boldsymbol{W}) \quad = \quad & \sum_{t=1}^{N_T} \sum_{\omega=1}^{N_\omega} [\boldsymbol{V}]_{t,\omega} \frac{\left(\frac{[\boldsymbol{V}]_{t,\omega}}{[\boldsymbol{W}\boldsymbol{H}]_{t,\omega}} \right)^{\beta-1} - 1}{(\beta-1)\beta} \\ & + \frac{[\boldsymbol{W}\boldsymbol{H}]_{t,\omega} - [\boldsymbol{V}]_{t,\omega}}{\beta} \end{aligned} \tag{3.28}$$

The α-divergence depends on a given parameter β and covers a wide range of well known distances such as the Hellinger distance (for $\beta = 0.5$), Pearson's chi-square distance ($\beta = 2$), Neyman's chi-square distance ($\beta = -1$) and it converges to the Kullback-Leibler divergence ($\beta = 0.5$).

Dhillon and Sra (2006) introduced a technique called Nonnegative Matrix Approximation (NNMA) for dimensionality reduction and data analysis that yields a part-based, sparse, nonnegative representation for nonnegative input data. NNMA can be viewed as a generalization of the NMF algorithms from Lee and Seung (2001) to the family of Bregman divergences-based objective functions [2]. The authors derived multiplicative update rules for objective functions consisting of any Bregman divergence function:

$$\begin{aligned}\Phi_{\varphi}(\boldsymbol{H},\boldsymbol{W}) \quad = \sum_{t=1}^{N_T}\sum_{\omega=1}^{N_\omega} \quad & \varphi([\boldsymbol{V}]_{t,\omega}) \\ & -\varphi([\boldsymbol{HW}]_{t,\omega}) \\ & -\nabla\varphi([\boldsymbol{HW}]_{t,\omega})([\boldsymbol{V}]_{t,\omega}-[\boldsymbol{HW}]_{t,\omega})\end{aligned} \tag{3.29}$$

The solution to such objective functions becomes:

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \otimes \psi^{-1}\left(\frac{\boldsymbol{W}^\top \psi\left(\boldsymbol{V}\right)}{\boldsymbol{W}^\top \psi\left(\boldsymbol{WH}\right)}\right) \tag{3.30}$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \otimes \psi^{-1}\left(\frac{\psi\left(\boldsymbol{V}\right)\boldsymbol{H}^\top}{\psi\left(\boldsymbol{WH}\right)\boldsymbol{H}^\top}\right) \tag{3.31}$$

where $\psi(x) = \nabla\varphi(x)$.

Extending the work by Dhillon and Sra (2006), Raczyński et al. (2008) derived the solutions for a specific sub-family of the Bregman divergences. The proposed subset is the r-divergence family, which is generated by a function whose second derivative is of the form $\varphi''(x) = x^{-r}$. The simplest solution for the generating function becomes:

$$\varphi(x) = \begin{cases} xlog(x) - x & \text{if } r = 1 \\ -log(x) - 1 & \text{if } r = 2 \\ \frac{x^{2-r}}{(1-r)(2-r)} & \text{otherwise} \end{cases} \tag{3.32}$$

It's easy to show that for different values of r the divergence becomes the Euclidean distance ($r = 0$), the I-divergence ($r = 1$) and especially the Itakura-Saito divergence ($r = 2$) which had been proven useful for speech spectra (Itakura and Saito, 1968; Gray et al., 1980).

The update rules for this sub-family of divergences was derived from 3.29 as:

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \otimes \frac{\boldsymbol{W}^\top\left(\boldsymbol{V} \otimes \left(\boldsymbol{WH}\right)^{-r}\right)}{\boldsymbol{W}^\top\left(\left(\boldsymbol{WH}\right)^{1-r}\right)} \tag{3.33}$$

2. Bregman divergences are functions of the form: $D_\varphi(x,y) \triangleq \varphi(x) - \varphi(y) - \nabla\varphi(y)(x-y)$. They are nonnegative, convex in the first argument and zero if and only if $x = y$.

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \otimes \frac{\left(\boldsymbol{V} \otimes (\boldsymbol{W}\boldsymbol{H})^{-r}\right) \boldsymbol{H}^{\top}}{\left((\boldsymbol{W}\boldsymbol{H})^{1-r}\right) \boldsymbol{H}^{\top}} \tag{3.34}$$

Cichocki et al. (2006) and Kompass (2007) presented multiplicative update rules for another subset of Bregman divergences called the β-divergences. This subset also included the Euclidean distance, Kullback-Leibler and the Itakura-Saito:

$$\begin{aligned} \Phi_{\beta}(\boldsymbol{H}, \boldsymbol{W}) \quad = \quad & \sum_{t=1}^{N_T} \sum_{\omega=1}^{N_\omega} [\boldsymbol{V}]_{t,\omega} \frac{[\boldsymbol{V}]_{t,\omega}^{\beta-1} - [\boldsymbol{W}\boldsymbol{H}]_{t,\omega}^{\beta-1}}{\beta(\beta-1)} \\ & + [\boldsymbol{W}\boldsymbol{H}]_{t,\omega}^{\beta-1} \frac{[\boldsymbol{W}\boldsymbol{H}]_{t,\omega} - [\boldsymbol{V}]_{t,\omega}}{\beta} \end{aligned} \tag{3.35}$$

This family of divergences was later studied in more detail by FitzGerald et al. (2009) and Févotte et al. (2009) with special attention to music signals and the Itakura-Saito case:

$$\Phi_{is}(\boldsymbol{H}, \boldsymbol{W}) \quad = \quad \sum_{t=1}^{N_T} \sum_{\omega=1}^{N_\omega} \frac{[\boldsymbol{V}]_{t,\omega}}{[\boldsymbol{W}\boldsymbol{H}]_{t,\omega}} - log \frac{[\boldsymbol{V}]_{t,\omega}}{[\boldsymbol{W}\boldsymbol{H}]_{t,\omega}} - 1 \tag{3.36}$$

The proposed generalized update rules for the β-divergence, the basis $\boldsymbol{W}$ and the gains $\boldsymbol{H}$ are expressed as:

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \otimes \frac{\boldsymbol{W}^{\top} \left((\boldsymbol{W}\boldsymbol{H})^{[\beta-2]} \otimes \boldsymbol{V}\right)}{\boldsymbol{W}^{\top} (\boldsymbol{W}\boldsymbol{H})^{[\beta-1]}} \tag{3.37}$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \otimes \frac{\boldsymbol{H}^{\top} \left((\boldsymbol{H}\boldsymbol{W})^{[\beta-2]} \otimes \boldsymbol{V}\right)}{(\boldsymbol{W}\boldsymbol{H})^{[\beta-1]} \boldsymbol{H}^{\top}} \tag{3.38}$$

where all powers are elementwise and $0 \leq \beta \leq 2$ is the coefficient that will define the objection function that is being minimized. $\beta = 2$ for the Euclidean distance (NMF_{euc}), $\beta = 1$ for the Kullback-Leibler divergence (NMF_{kl}) and $\beta = 0$ for the Itakura-Saito divergence (NMF_{is}).

The special case of the Itakura-Saito divergence results in the following update rules:

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \otimes \frac{\boldsymbol{W}^{\top} \left(\frac{\boldsymbol{V}}{(\boldsymbol{W}\boldsymbol{H})^2}\right)}{\boldsymbol{W}^{\top} (\boldsymbol{W}\boldsymbol{H})^{-1}} \tag{3.39}$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \otimes \frac{\boldsymbol{H}^\top \left(\frac{\boldsymbol{V}}{(\boldsymbol{H}\boldsymbol{W})^2}\right)}{(\boldsymbol{W}\boldsymbol{H})^{-1}\boldsymbol{H}^\top} \tag{3.40}$$

The Itakura-Saito divergence is widely used in audio separation tasks due to the way it quantifies differences between spectra. As described in Section 2.4, human audio perception is characterized by perceiving audio differences in a wide loudness range. We may assume that an appropriate divergence should quantify similarly the differences in both high and low magnitude ranges. From this point of view, Itakura-Saito shows better performance than other well known divergences such as the Euclidean distance or the Kullback-Leibler.

NMF Regularizations

Févotte et al. (2009) proposed temporal smoothness using priors on the gains. However these were only proposed for an algorithm (SAGE) based on Expectation Maximization (EM), different from NMF multiplicative updates. Bertin et al. (2010) used harmonicity and temporal smoothness regularizations.

Although NMF had proven useful for certain decomposition tasks, the non-negativity constraint was not sufficient to always achieve a parts-based representation of the data. Li et al. (2001) and Feng et al. (2002) introduced into the standard NMF cost function a sparseness regularization term on the gains and a locality term on the basis. The goal of such cost terms was to learn spatially localized parts-based representation of visual patterns. This formulation lead to the Local Nonnegative Matrix Factorization algorithm (LNMF) whose cost function is defined as:

$$\begin{aligned} \Phi_{lnmf}(\boldsymbol{H},\boldsymbol{W}) &= \Phi_{kl}(\boldsymbol{H},\boldsymbol{W}) + \alpha \sum_{i,j=1}^{N_W} \left[\boldsymbol{W}^\top \boldsymbol{W}\right]_{ij} \\ &\quad -\beta \sum_{i,j=1}^{N_W} \left[\boldsymbol{H}\boldsymbol{H}^\top\right]_{ij} \end{aligned} \tag{3.41}$$

The multiplicative update rules that the authors proposed remain unchanged for the basis, and become the following for the gains:

$$\boldsymbol{H} \leftarrow \sqrt{\boldsymbol{H} \otimes \boldsymbol{W}^\top \frac{\boldsymbol{V}}{(\boldsymbol{W}\boldsymbol{H})}} \tag{3.42}$$

It is worth noting that even though the update rule is not strictly multiplicative, given the square root, the convergence is still fast.

Hoyer and Dayan (2004) considered adding limits to the sparsity of the basis $\boldsymbol{W}$ and/or of the gains $\boldsymbol{H}$ and derived the algorithms to enforce such limits during the factorization. Hoyer (2002) also considered adding a sparseness regularization term to the gains in the Euclidean distance objective function. This method was called Non-negative Sparse Coding (NNSC):

$$\Phi_{nnsc}(\boldsymbol{H},\boldsymbol{W}) \quad = \quad \Phi_{euc}(\boldsymbol{H},\boldsymbol{W}) + \lambda \sum_{t=1}^{N_T}\sum_{w=1}^{N_W} \boldsymbol{H}_{w,t} \tag{3.43}$$

However in this case the update rule of the basis is not multiplicative and the non-negativity constraint is implemented by setting the negative values of $\boldsymbol{W}$ to 0 after each update. For this objective function the author proposed the following update rules:

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \otimes \frac{\boldsymbol{W}^\top \boldsymbol{V}}{\boldsymbol{W}^\top (\boldsymbol{W}\boldsymbol{H} + \lambda)} \tag{3.44}$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} - \mu \left(\boldsymbol{W}\boldsymbol{H} - \boldsymbol{V}\right) \boldsymbol{H}^\top \tag{3.45}$$

Eggert and Korner (2004) also added the sparseness regularization term to the gains in the Euclidean distance objective function. The authors also took into consideration the normalization of the basis by inserting it into the objective function: The following equation does not fit properly.

$$\begin{aligned} \Phi_{eggert}(\boldsymbol{H},\boldsymbol{W}) \quad = \quad & \sum_{t=1}^{N_T}\sum_{\omega=1}^{N_\omega} \frac{1}{2} \left(\left[\frac{\boldsymbol{W}}{\|\boldsymbol{W}\|} \boldsymbol{H} \right]_{t,\omega} - [\boldsymbol{V}]_{t,\omega} \right)^2 \\ & + \lambda \sum_{t=1}^{N_T}\sum_{w=1}^{N_W} g\left(\boldsymbol{H}_{w,t}\right) \end{aligned} \tag{3.46}$$

where $g(x)$ is a $\Re \rightarrow \Re$ differentiable function that acts as a regularization term on the gains and $\|\boldsymbol{X}\|$ stands for any differentiable norm. This formulation allowed the authors to derive simple and easy to implement update rules similar to those presented by Lee and Seung (2001):

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \otimes \frac{\overline{\boldsymbol{W}}^\top \boldsymbol{V}}{\overline{\boldsymbol{W}}^\top \left(\overline{\boldsymbol{W}}\boldsymbol{H}\right) + \lambda g'\left(\boldsymbol{H}\right)} \tag{3.47}$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \otimes \frac{\boldsymbol{H}^{\top}\left[\boldsymbol{V} + \left(\overline{\boldsymbol{W}}\boldsymbol{H}\right)\nabla_{\boldsymbol{W}}\|\boldsymbol{W}\|\right]}{\boldsymbol{H}^{\top}\left[\boldsymbol{W}\boldsymbol{H} + \left(\boldsymbol{V}\overline{\boldsymbol{W}}\right)\nabla_{\boldsymbol{W}}\|\boldsymbol{W}\|\right]} \tag{3.48}$$

The authors also provided update rules for the specific case where $g(x) = x$ and where $\|\boldsymbol{X}\|$ is the Euclidean norm.

Cichocki et al. (2006, 2008) generalized the addition of sparseness constraints and regularization to the Euclidean distance and to the generalized Kullback-Leibler divergence (also known as the I-divergence) leading to the following objective functions:

$$\Phi_{eucreg}(\boldsymbol{H},\boldsymbol{W}) = \Phi_{euc}(\boldsymbol{H},\boldsymbol{W}) + \alpha_{\boldsymbol{W}}\boldsymbol{J}_{\boldsymbol{W}}(\boldsymbol{W}) + \alpha_{\boldsymbol{H}}\boldsymbol{J}_{\boldsymbol{H}}(\boldsymbol{H}) \tag{3.49}$$

$$\Phi_{klreg}(\boldsymbol{H},\boldsymbol{W}) = \Phi_{kl}(\boldsymbol{H},\boldsymbol{W}) + \alpha_{\boldsymbol{W}}\boldsymbol{J}_{\boldsymbol{W}}(\boldsymbol{W}) + \alpha_{\boldsymbol{H}}\boldsymbol{J}_{\boldsymbol{H}}(\boldsymbol{H}) \tag{3.50}$$

where $\boldsymbol{J}_{\boldsymbol{W}}(\boldsymbol{W})$ and $\boldsymbol{J}_{\boldsymbol{H}}(\boldsymbol{H})$ are custom cost functions (or regularization terms) for the basis and the gains respectively. $\alpha_{\boldsymbol{W}}$ and $\alpha_{\boldsymbol{H}}$ are the regularization parameters.

The proposed update rules for $\Phi_{eucreg}(\boldsymbol{H},\boldsymbol{W})$ are:

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \otimes \frac{\lfloor \boldsymbol{W}^{\top}\boldsymbol{V} - \alpha_{\boldsymbol{H}}\varphi_{\boldsymbol{H}}\rfloor_{\varepsilon}}{\boldsymbol{W}^{\top}(\boldsymbol{W}\boldsymbol{H}) + \varepsilon} \tag{3.51}$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \otimes \frac{\lfloor \boldsymbol{V}\boldsymbol{H}^{\top} - \alpha_{\boldsymbol{W}}\varphi_{\boldsymbol{W}}\rfloor_{\varepsilon}}{(\boldsymbol{W}\boldsymbol{H})\boldsymbol{H}^{\top} + \varepsilon} \tag{3.52}$$

where $\varepsilon = 10^{-9}$ is introduced to ensure non-negativity and avoid divisions by zero. $\lfloor x \rfloor_{\varepsilon} = max\{x, \varepsilon\}$ and the matrices $\varphi_{\boldsymbol{W}}$ and $\varphi_{\boldsymbol{H}}$ are defined as:

$$[\varphi_{\boldsymbol{W}}]_{i,j} = \frac{\delta \boldsymbol{J}_{\boldsymbol{W}}(\boldsymbol{W})}{\delta \boldsymbol{W}_{i,j}}, \qquad [\varphi_{\boldsymbol{H}}]_{i,j} = \frac{\delta \boldsymbol{J}_{\boldsymbol{H}}(\boldsymbol{H})}{\delta \boldsymbol{H}_{i,j}} \tag{3.53}$$

And for the $\Phi_{klreg}(\boldsymbol{H},\boldsymbol{W})$ objective function the authors proposed:

$$\boldsymbol{H} \leftarrow \left(\boldsymbol{H} \otimes \frac{\boldsymbol{W}^{\top}\frac{\boldsymbol{V}}{(\boldsymbol{W}\boldsymbol{H})}}{\lfloor \boldsymbol{W}^{\top} + \alpha_{\boldsymbol{H}}\varphi_{\boldsymbol{H}}\rfloor_{\varepsilon}}\right)^{1+\alpha_{s}\boldsymbol{H}} \tag{3.54}$$

$$\boldsymbol{W} \leftarrow \left(\boldsymbol{W} \otimes \frac{\boldsymbol{H}^{\top}\frac{\boldsymbol{V}}{(\boldsymbol{W}\boldsymbol{H})}}{\lfloor \boldsymbol{H}^{\top} + \alpha_{\boldsymbol{W}}\varphi_{\boldsymbol{W}}\rfloor_{\varepsilon}}\right)^{1+\alpha_{s}\boldsymbol{W}} \tag{3.55}$$

where additional small positive regularization terms $\alpha_{s\boldsymbol{H}} \geq 0$ and $\alpha_{s\boldsymbol{W}} \geq 0$ are introduced in order to enforce sparseness of the solution, if necessary. Typical values of $\alpha_{s\boldsymbol{H}}, \alpha_{s\boldsymbol{W}} \in [0.001, 0.005]$.

Chen and Cichocki (2005) and Chen et al. (2006) proposed a temporal smoothness regularization based on the ratio between the mean value and the exponentially weighted local average of the gain of the component. This regularization is based on a Toeplitz matrix that performs the local average filter. The authors also proposed a component decorrelation regularization. The proposed method is evaluated on EEG data. On the other hand, Pascual-Montano et al. (2006) proposed a temporally non-smooth version of the Nonnegative Matrix Factorization algorithm. The goal is to decompose the signal in factors that are localized in time. The technique is then applied to both synthetic and experimental data with a focus on brain imaging.

Virtanen (2007) proposed the addition of two regularization terms based on the gains matrix to the I-divergence cost function in the specific case of music audio source separation. This leads to the cost function 3.50 where the $\boldsymbol{J}_{\boldsymbol{W}}(\boldsymbol{W}) = 0$ and $\boldsymbol{J}_{\boldsymbol{H}}(\boldsymbol{H})$ is defined as:

$$\boldsymbol{J}_{\boldsymbol{H}}(\boldsymbol{H}) = \alpha_{tc}\boldsymbol{J}^{tc}_{\boldsymbol{H}}(\boldsymbol{H}) + \alpha_{s}\boldsymbol{J}^{s}_{\boldsymbol{H}}(\boldsymbol{H}) \tag{3.56}$$

One regularization term consists in penalizing temporal discontinuities on the gains of each component. The idea behind this term is that many musical sources have a slowly varying spectrum, such as the sustain phases of the notes. This consists in adding a cost term of the form:

$$\boldsymbol{J}^{tc}_{\boldsymbol{H}}(\boldsymbol{H}) = \sum_{w}^{N_W} \frac{1}{{\sigma^t_w}^2} \sum_{t}^{N_T} \Big([\boldsymbol{H}]_{w,t} - [\boldsymbol{H}]_{w,t-1} \Big)^2 \tag{3.57}$$

where the standard deviation of the components is estimated as $\sigma^t_w = \sqrt{(1/N_T)\sum_t^{N_T}([\boldsymbol{H}]^2_{w,t})}$.

The other term promotes the sparseness of the gains in the component dimension. This means that the solution with fewer gains active is preferred. The rationale for this term is to try to explain the majority of a given source with the fewest number of bases possible.

$$\boldsymbol{J}^{s}_{\boldsymbol{H}}(\boldsymbol{H}) = \sum_{w}^{N_W} \sum_{t}^{N_T} g([\boldsymbol{H}]_{w,t} / \sigma_w) \tag{3.58}$$

where $g(\cdot)$ is a function that penalizes non-zero gains. The update rules for the function $g(x) = |x|$ were derived by the authors.

In order to derive the update rules in this situation, the author calculated the gradient of the terms:

$$\begin{aligned}\left[\varphi^{tc}_{\boldsymbol{H}}(\boldsymbol{H})\right]_{w,t} &= 2N_T \frac{2\,[\boldsymbol{H}]_{w,t} - [\boldsymbol{H}]_{w,t-1} - [\boldsymbol{H}]_{w,t+1}}{\sum_i^{N_T} [\boldsymbol{H}]^2_{w,i}} \\ &\quad - N_T \frac{2\,[\boldsymbol{H}]_{w,t} \sum_{i=2}^{N_T} \left([\boldsymbol{H}]_{w,i} - [\boldsymbol{H}]_{w,i-1}\right)^2}{\left(\sum_i^{N_T} [\boldsymbol{H}]^2_{w,i}\right)^2} \end{aligned} \tag{3.59}$$

$$\begin{aligned}[\varphi^{s}_{\boldsymbol{H}}(\boldsymbol{H})]_{w,t} &= \frac{1}{\sqrt{\frac{1}{N_T}\sum_i^{N_T} [\boldsymbol{H}]^2_{w,t}}} \\ &\quad - \sqrt{N_T} \frac{[\boldsymbol{H}]_{w,t} \sum_i^{N_T} [\boldsymbol{H}]_{w,i}}{\left(\sum_i^{N_T} [\boldsymbol{H}]^2_{w,i}\right)^{3/2}} \end{aligned} \tag{3.60}$$

The gradients of the regularization terms proposed by Virtanen (2007) can take negative values and therefore cannot be directly used in equation 3.50. Instead, the author derives new update rules by separating the gradients into positive and negative contributions. This is also done with the Kullback-Leibler cost function and the new update rule factor is defined as the negative contribution divided by the positive contribution:

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \otimes \frac{\boldsymbol{W}^\top \frac{\boldsymbol{V}}{(\boldsymbol{W}\boldsymbol{H})} + \alpha_{\boldsymbol{H}} \varphi^{-}_{\boldsymbol{H}}}{\lfloor \boldsymbol{W}^\top + \alpha_{\boldsymbol{H}} \varphi^{+}_{\boldsymbol{H}} \rfloor_\varepsilon} \tag{3.61}$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \otimes \frac{\boldsymbol{H}^\top \frac{\boldsymbol{V}}{(\boldsymbol{W}\boldsymbol{H})} + \alpha_{\boldsymbol{W}} \varphi^{-}_{\boldsymbol{W}}}{\lfloor \boldsymbol{H}^\top + \alpha_{\boldsymbol{W}} \varphi^{+}_{\boldsymbol{W}} \rfloor_\varepsilon} \tag{3.62}$$

where $\varphi^{+}_{\boldsymbol{H}}$ and $\varphi^{-}_{\boldsymbol{H}}$ are the positive and negative contributions to $\varphi_{\boldsymbol{H}}$ respectively. We can observe that if the gradients of the regularization terms are positive ($\varphi^{-}_{\boldsymbol{H}} = 0$) the update rules are equivalent to Equations 3.54 and 3.55 with their regularization terms $\alpha_{s\boldsymbol{H}}$ and $\alpha_{s\boldsymbol{W}}$ set to 0.

Wilson et al. (2008b,a) applied the objective function 3.50 and the update rules 3.54 to the task of speech denoising and created the following regularization term:

$$\boldsymbol{J}(\boldsymbol{H}) = \boldsymbol{J}^{w}_{\boldsymbol{H}}(\boldsymbol{H}) + \boldsymbol{J}^{t}_{\boldsymbol{H}}(\boldsymbol{H}) \tag{3.63}$$

where $\boldsymbol{J}_{\boldsymbol{H}}^{w}$ accounts for the relations of the gains between all the bases at a given frame in time and $\boldsymbol{J}_{\boldsymbol{H}}^{t}$ accounts for the covariance of the gains between consecutive frames in time. They are defined as:

$$\boldsymbol{J}_{\boldsymbol{H}}^{w}(\boldsymbol{H}) = \alpha_w \frac{1}{2} \sum_{t=1}^{N_T} \left(log\boldsymbol{H}_{:,t} - \mu\right)^{\top} \boldsymbol{\Lambda}_w^{-1} \left(log\boldsymbol{H}_{:,t} - \mu\right) - log\left[(2\pi)^{N_W} |\boldsymbol{\Lambda}_w|\right] \tag{3.64}$$

$$\boldsymbol{J}_{\boldsymbol{H}}^{t}(\boldsymbol{H}) = \alpha_t \frac{1}{2} \sum_{w=1}^{N_W} \left(log\boldsymbol{H}_{w,:} - \mu_w \mathbf{1}^{\top}\right)_w^{\top} \boldsymbol{\Lambda}_t^{-1} \left(log\boldsymbol{H}_{w,:} - \mu_w \mathbf{1}^{\top}\right)^{\top} - log\left[(2\pi)^{N_W} |\boldsymbol{\Lambda}_t|\right] \tag{3.65}$$

where μ and μ_w are the means of the gains $\boldsymbol{H}$ in the basis and time dimensions respectively. Similarly $\boldsymbol{\Lambda}_w$ and $\boldsymbol{\Lambda}_t$ represent the covariance matrices of the gains $\boldsymbol{H}$. These statistics have been found previously by training a standard NMF on speech and noise signals separately and then concatenating their statistics by assuming independence. $\boldsymbol{X}_{i,:}$ represents the i^{th} row of $\boldsymbol{X}$ and $\boldsymbol{X}_{:,j}$ is the j^{th} column. In this work only the gains are learned, since the bases are fixed to those learned during the training phase. Therefore the regularization terms are used to penalize solutions of the $\boldsymbol{H}$ whose statistics deviate from those found during the training phase. Finally, by applying 3.54, the update rule of the gains becomes:

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \otimes \frac{\boldsymbol{W}^{\top} \frac{\boldsymbol{V}}{(\boldsymbol{W}\boldsymbol{H})}}{\lfloor \boldsymbol{W}^{\top} + \varphi_{\boldsymbol{H}}^{w}(\boldsymbol{H}) + \varphi_{\boldsymbol{H}}^{t}(\boldsymbol{H}) \rfloor_{\varepsilon}} \tag{3.66}$$

where:

$$\left[\varphi_{\boldsymbol{H}}^{w}(\boldsymbol{H})\right]_{w,t} = -\alpha_w \frac{\left(\Lambda_w^{-1} \left(log\boldsymbol{H}_{:,t} - \mu\right)\right)_w}{\boldsymbol{H}_{w,t}} \tag{3.67}$$

$$\left[\varphi_{\boldsymbol{H}}^{t}(\boldsymbol{H})\right]_{w,t} = -\alpha_t \frac{\left({}_w\Lambda_t^{-1} \left(log\boldsymbol{H}_{w,:} - \mu_w \mathbf{1}^{\top}\right)^{\top}\right)_t}{\boldsymbol{H}_{w,t}} \tag{3.68}$$

The authors showed that the use of these regularization terms in the NMF objective function improves the speech denoising task.

Raczyński et al. (2008) also studied the introduction of regularization terms in the NMF formulation in the context of music transcription tasks. By

extending the r-divergence objective function 3.29 and 3.32 with two extra regularization terms, they derived general update rules:

$$\Phi_{\varphi_{reg}}(\boldsymbol{H},\boldsymbol{W}) = \Phi_{\varphi}(\boldsymbol{H},\boldsymbol{W}) + \boldsymbol{J}_{\boldsymbol{W}}(\boldsymbol{W}) + \boldsymbol{J}_{\boldsymbol{H}}(\boldsymbol{H}) \tag{3.69}$$

where $\boldsymbol{J}_{\boldsymbol{W}}(\boldsymbol{W})$ and $\boldsymbol{J}_{\boldsymbol{H}}(\boldsymbol{H})$ are the regularization terms on the basis $\boldsymbol{W}$ and on the gains $\boldsymbol{H}$ respectively. The proposed update rules become:

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \otimes \frac{\boldsymbol{W}^\top \left(\boldsymbol{V} \otimes (\boldsymbol{W}\boldsymbol{H})^{-r}\right)}{\boldsymbol{W}^\top \left((\boldsymbol{W}\boldsymbol{H})^{1-r}\right) + \varphi_{\boldsymbol{H}}(\boldsymbol{H})} \tag{3.70}$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \otimes \frac{\left(\boldsymbol{V} \otimes (\boldsymbol{W}\boldsymbol{H})^{-r}\right)\boldsymbol{H}^\top}{\left((\boldsymbol{W}\boldsymbol{H})^{1-r}\right)\boldsymbol{H}^\top + \varphi_{\boldsymbol{W}}(\boldsymbol{W})} \tag{3.71}$$

Another interesting development in the work of Raczyński et al. (2008) is the proposal of three specific regularization terms on the gains $J_{\boldsymbol{H}}(\boldsymbol{H})$ for the context of polyphonic music transcription and separation. The first regularization proposed was a sparseness regularization that would lead to a preference for sparse gains:

$$J^{sp}_{\boldsymbol{H}}(\boldsymbol{H}) = \alpha_{sp}|\boldsymbol{H}^p|, \tag{3.72}$$

$$\varphi^{sp}_{\boldsymbol{H}}(\boldsymbol{H}) = \alpha_{sp} p \boldsymbol{H}^{p-1} \tag{3.73}$$

This regularization is simply the l_p-norm of the gains. The p must be kept under 2.

The next regularization term proposal is a penalty for cross-correlation of the gains. This permits decreasing the cross-talk between activities of different notes:

$$J^{cr}_{\boldsymbol{H}}(\boldsymbol{H}) = \alpha_{cr}|\boldsymbol{C} \otimes \left(\boldsymbol{H}\boldsymbol{H}^\top\right)|, \tag{3.74}$$

$$\varphi^{cr}_{\boldsymbol{H}}(\boldsymbol{H}) = 2\alpha_{cr}\boldsymbol{C}\boldsymbol{H} \tag{3.75}$$

where $\boldsymbol{C}$ is a weight matrix that selects what cross-terms to penalize and by how much. The weights are set so that the non-cross terms (elements on the diagonal) are not penalized $\boldsymbol{C}_{ii} = 0$ and so that the penalties only depend on the intervals between the notes, therefore the matrix $\boldsymbol{C}$ is circulant. These conditions allow the simple derivation that leads to $\varphi^{cr}_{\boldsymbol{H}}(\boldsymbol{H})$. Usually the $\boldsymbol{C}$ is used to especially penalize octave, fifths and thirds intervals.

Together the sparseness and cross-correlation terms help in the reduction of the typical errors encountered in the transcription of polyphonic music such as octave and fifths errors.

The last proposed regularization term is used to encourage temporal smoothness of the gains. It is done like the cross-correlation penalty:

$$J_{\boldsymbol{H}}^{sm}(\boldsymbol{H}) = -\alpha_{sm}|\boldsymbol{D} \otimes \left(\boldsymbol{H}^{\top}\boldsymbol{H}\right)|, \tag{3.76}$$

$$\varphi_{\boldsymbol{H}}^{sm}(\boldsymbol{H}) = -2\alpha_{sm}\boldsymbol{H}\boldsymbol{D} \tag{3.77}$$

where $\boldsymbol{D}$ is used to penalize discontinuities between consecutive frames. As in 3.74 $\boldsymbol{D}$ is forced to be circulant and have a nullified diagonal ($\boldsymbol{D}_{ii} = 0$) in order to achieve a simple derivative that will lead to the proposed $\varphi_{\boldsymbol{H}}^{sm}(\boldsymbol{H})$. The specific proposed $\boldsymbol{D}$ has an exponentially decaying profile, so that elements close to the diagonal have high values and those that are far tend to zero. However the authors noted that negative elements in $\varphi_{\boldsymbol{H}}^{sm}(\boldsymbol{H})$ tend to lead to instabilities and negative solutions in the algorithm. To avoid this $\varphi_{\boldsymbol{H}}^{sm'}(\boldsymbol{H}) = e^{\varphi_{\boldsymbol{H}}^{sm}(\boldsymbol{H})}$ is used instead. The authors showed that in a specific multipitch transcription task, the use of a combination of these regularization terms improves accuracy.

NMF Constraints

The use of regularizations in the objective function can greatly help to improve the physical interpretation of the spectrum templates in the basis matrix $\boldsymbol{W}$ or of the gains $\boldsymbol{H}$. For instance in an audio mixture factorization, promoting a set of gains to be sparse will likely be used to reconstruct impulsive sounds.

Another method to improve the physical interpretation of the components and to guide the factorization process is fixing some of the matrices' elements. This is equivalent to reducing the number of free parameters.

We must note that the NMF algorithms discussed do not guarantee convergence towards a global minimum of the objective function. These multiplicative update rules only guarantee convergence towards a local minimum.

We present a series of algorithms where the basis matrix $\boldsymbol{W}$ or a subset of it is fixed. Even though the idea of fixing some or all components of the basis matrix $\boldsymbol{W}$ is simple, there has been a significant amount of research into what values to use for the fixed components. Two main trends have developed. First, supervised or semi-supervised methods rely on using

ground truth data and/or sources in isolation to estimate a set of useful bases. This family of methods is often called Supervised or Semi-supervised NMF. Second, the basis can be set or constrained using knowledge of the type of sources we want to separate. This second family of methods do not need training and therefore are often called unsupervised NMF.

Abdallah and Plumbley (2004) presented a semi-supervised method for polyphonic music transcription based on Non-negative Sparse Coding (NNSC) (Hoyer, 2002). In this method a basis of piano harmonic spectra is learned using isolated piano notes as training data. In the second stage of the method, the basis is fixed and used to learn the gains or activations of the piano notes in the mixture. In order to enforce a harmonic structure on the learned piano basis, the basis components are initialized to harmonic combs before the training stage.

Schmidt and Olsson (2006) presented a supervised NMF method based on the Sparse NMF (SNMF) objective function for use on speech separation. The bases were fixed to a set of components that were previously trained on audio of individual speakers. The authors presented two ways to estimate the fix basis. The first approach consisted of applying a non-constrained SNMF and also learning the basis matrix $\boldsymbol{W}$, on a large amount of individual and isolated speaker data. The learned basis would then be used fixed in the separation stage. The second approach consisted in reducing the amount of preliminary data necessary by first segmenting the isolated speaker data into phonemes. Then the non-constrained SNMF would be run independently on all the segments of each phoneme, generating a set of bases per phoneme. All these bases would finally be concatenated to create the learned basis used in the separation phase.

Smaragdis et al. (2007) proposed the use of a method called Probabilistic Latent Component Analysis (PLCA) in order to perform supervised or semi-supervised single channel source separation. This method is closely related to NMF, in fact, the two procedures can be proven to be numerically identical [3]. The authors proposed a learning stage where a set of sounds are learned using PLCA, which leads to a set of basis spectra. The learning procedure uses an entropic prior that permits controlling the amount of sparsity in the solutions. The learned basis spectra may be used in the separation procedure by fixing them as the full set of bases with which to reconstruct the data. This method is useful in the case where we have been able to train our models for all the sources present in the mixture

3. This relation is further described in the section 3.4

and is called supervised PLCA or NMF. On the other hand, when only some of the sources are known, the authors allowed a set of spectra in the bases to be learned during the separation phase, and that is known as the semi-supervised method.

Instead of performing an NMF on isolated sources in order to extract the set of learned bases, Smaragdis et al. (2009) used as basis components actual spectrum frames from training data of the sources in isolation. This is known as exemplar-based NMF and the study showed that it can improve the performance of the separation in comparison to similar systems based on statistical models. Raj et al. (2011) proposed using an exemplar-based NMF approach in the context of speech enhancement. In order to build the basis of the NMF the authors concatenated a set of randomly selected spectra magnitudes of the desired and other competing sources. These were used as bases for the factorization. For the estimation of the desired sources the gains of the competing sources were set to 0.

An alternative way to restrict the solution space of the NMF is to constrain some of the elements of the factors to constant values based on *a priori* knowledge about the components and their activations. The large amount of *a priori* information in musical mixtures makes this method attractive for such situations. One simple way of adding constraints is by initializing some of the elements to 0, given the multiplicative nature of the updates, these elements will remain at 0. Another way is to set them to some constant that respects a given structure and not update them when computing the update rules.

As we will see in Section 3.4, one of the first uses of constraints in the context of music mixtures NMF was Heittola et al. (2009), who used the source-filter model in order to constrain the excitation components to harmonically structured spectra and the filter components to smooth spectra. Other examples of the use of such constraints are found in Durrieu et al. (2009b), Hennequin et al. (2010) and Durrieu et al. (2011).

These types of constraints are not used only in harmonic sources. Transients can also be modeled using temporal localized spectra. Wu et al. (2011) and Ewert and Müller (2011) proposed spectrally constrained and temporally localized Gaussians as constant components during the factorization. In Ewert and Müller (2012) these constraints were taken a step further by using external information specific to the mixture. In their method, there are two types of basis components. One type represents harmonic spectra and is enforced by setting to zero all the bins but those around the partial

positions. The other type represents the transients, in which all the elements are non-zero and therefore are free to take any positive value. The gains of the transient components are only set nonzero around the frames where the notes start, given the score. On the other hand, the gains of the harmonic components are set nonzero during the sustain regions of the notes.

NMF Spectrum Models

The standard NMF algorithm proposed by Lee and Seung (2001) performed the factorization of a given method into two matrices $\boldsymbol{V} = \boldsymbol{W}\boldsymbol{H}$. In the context of audio, traditionally the NMF decomposition would be applied to the magnitude or power spectrogram of the mixture signal, so that each component in our basis matrix $\boldsymbol{W}$ would represent a single spectral frame. This method fits well with sources that have constant or linearly evolving spectral profiles, and the decomposition has proven very useful in many tasks and applications. However in some situations we may have *a priori* information about the components that can help us restrict even further the decomposition. This is frequently the case with audio and musical signals that present a highly structured spectrum (see 2.4 and 2.4). In the music context, NMF has been applied to more complex spectrum models in which the spectrogram of a signal is not simply the multiplication of two unknown non-negative matrices.

The use of more complex spectrum models permits a better interpretation of the components. For instance, if we want to model the source filter voice production phenomena we can decompose the basis matrix into two other matrices, one with the sources or excitation spectra and another one with the filterbanks. Another reason to introduce more than two factors into our spectrum models is to constrain the components of our basis. If we know that our basis components can only be generated by the filtering of some given excitation spectra, then using such a factorization will force the basis components to contain that structure. Finally one more reason to further decompose the factors of the standard NMF is to reduce the number of free parameters. For example, if we have a basis matrix $\boldsymbol{W}$ of size $N_\omega \times N_W$ and we know that its components are structured in a specific way, then we may break the basis up into two other matrices so that $\boldsymbol{W} = \boldsymbol{W}_s\boldsymbol{W}_f$ where $\boldsymbol{W}_s \in N_\omega \times k$ and $\boldsymbol{W}_f \in k \times N_W$. Given a small enough k $N_\omega N_W \geq N_\omega k + kN_W$.

Smaragdis (2004) created a set of basis components that represent sound objects and that have a support in time longer than that of one STFT

frame. In this way the basis components are thus not rows of $\boldsymbol{W}$ but rather matrices composed of several consecutive spectrum frames. The spectrum model can be expressed in the following way:

$$\boldsymbol{V} \approx \sum_{h=1}^{H} \boldsymbol{W}_h \overset{h\rightarrow}{\boldsymbol{H}} \tag{3.78}$$

where the $\overset{i\rightarrow}{(\cdot)}$ operator shifts the columns of its argument by i positions to the right, filling the new columns with 0. This model implies that the spectrogram is the result of a linear mixture of convolutions between each non-negative basis component and its gains. This also implies that the basis components have one more dimension h representing the temporal dimension. Smaragdis (2004) also proposed a set of multiplicative update rules to minimize the Kullback-Leibler divergence objective function. This solution is called convolutive NMF and the proposed update rules are:

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \otimes \frac{\boldsymbol{W}_t^\top \left[\frac{\boldsymbol{V}}{\boldsymbol{\Lambda}}\right]^{\overset{\leftarrow t}{}}}{\boldsymbol{W}^\top \boldsymbol{1}} \tag{3.79}$$

$$\boldsymbol{W}_t \leftarrow \boldsymbol{W}_t \otimes \frac{\frac{\boldsymbol{V}}{\boldsymbol{\Lambda}} \overset{t\rightarrow}{\boldsymbol{H}}^\top}{\boldsymbol{1}\boldsymbol{H}^\top} \tag{3.80}$$

where $\boldsymbol{1}$ is a matrix of ones the same size as $\boldsymbol{V}$, $\boldsymbol{\Lambda} = \sum_{h=1}^{H} \boldsymbol{W}_h \overset{h\rightarrow}{\boldsymbol{H}}$ and the overset operator $\overset{\leftarrow i}{(\cdot)}$ shifts the columns of its argument by i positions to the left. These rules have to be applied for each basis $t \in [0 \dots T-1]$. Following the convolutive NMF work, O'Grady and Pearlmutter (2006) and Mørup and Schmidt (2006) proposed a solution that adds sparsity constraints on the $\boldsymbol{H}$. Schmidt and Mørup (2006a,b) also proposed convolutive NMF spectrum models where the convolution is performed on both the frequency and time domains. They also proposed the use of sparsity constraints. However for sparse constraints no multiplicative update rules were proposed.

Sra and Dhillon (2006) and Dhillon and Sra (2006) studied the derivation of different NMF problems mainly using the Karush-Karger-Tucker (KKT) conditions. This work led to the easy derivation of complex factorization models with more than two factors. The authors generalized the factorization of a matrix into multiple non-negative matrices $\boldsymbol{V} \approx \boldsymbol{W}_1 \boldsymbol{W}_2 \dots \boldsymbol{W}_r \dots \boldsymbol{W}_R$. This was called Multifactor Non-negative Matrix

Approximation:

$$\boldsymbol{V} \approx \prod_{r=1}^{R} \boldsymbol{W}_r \tag{3.81}$$

The multiplicative rule for such factorization using a Bregman divergence Φ_φ cost function 3.29 becomes:

$$\boldsymbol{W}_r \leftarrow \boldsymbol{W}_r \otimes \frac{\hat{\boldsymbol{B}}^\top(\psi(\boldsymbol{R}) \otimes \boldsymbol{V})\hat{\boldsymbol{C}}^\top}{\hat{\boldsymbol{B}}^\top(\psi(\boldsymbol{R}) \otimes \boldsymbol{R})\hat{\boldsymbol{C}}^\top} \tag{3.82}$$

where $\hat{\boldsymbol{B}} = \boldsymbol{W}_1\boldsymbol{W}_2 \ldots \boldsymbol{W}_{r-1}$, $\hat{\boldsymbol{C}} = \boldsymbol{W}_{r+1}\boldsymbol{W}_{r+2} \ldots \boldsymbol{W}_R$, $\boldsymbol{R} = \boldsymbol{W}_1\boldsymbol{W}_2 \ldots \boldsymbol{W}_r \ldots \boldsymbol{W}_R$ and $\psi(x) = \nabla\varphi(x)$.

Another of the applications of the NNMA is the use of weights in non-negative factorization problems. Weights can be applied in different ways during the factorization process. The objective function may be weighted elementwise, we may also weight the approximant ($\boldsymbol{M} \otimes \boldsymbol{W}\boldsymbol{H}$) or we may weight each of the approximant factors by multiplying them by weight matrices $\boldsymbol{V} \approx \boldsymbol{M}_1\boldsymbol{W}\boldsymbol{H}\boldsymbol{M}_2$.

FitzGerald et al. (2008) proposed a spectrum model specifically designed for stereo signals. The sources are panned independently and therefore contribute with different gains to each channel. This model approximates the stereo spectrograms of the signal as the outer product of several tensors. The tensors represent the gains of the instruments per channel, the basis components of the instruments' spectra and the gains of these components. The standard NMF cannot be used to perform the factorization since it deals with 2-dimensional matrices. Instead the authors derived a method called Nonnegative Tensor Factorization (NTF). This method is derived similar to the NMF algorithm and leads to multiplicative update rules to estimate the different factors. FitzGerald et al. (2008) extended the work by proposing and exploring signal tensor frameworks that incorporate several restrictions and constraints. Among these models we find the incorporation of shift-invariance as in Shifted 2D Nonnegative Tensor Factorization (SNTF), fixed harmonic basis as in Sinusoidal Shifted 2D Nonnegative Tensor Factorization (SSNTF), source-filter models (SF-SSNTF) and models containing wideband noise components (SF-SSNTF+N).

Another spectrum model specifically designed for musical instruments was proposed by Virtanen and Klapuri (2006) and Vincent et al. (2008). The

proposal consisted of a spectrum model defined as:

$$\boldsymbol{V} \approx \hat{\boldsymbol{V}} = \boldsymbol{W}_{\Gamma}\boldsymbol{W}_{f_0}\boldsymbol{H}_{f_0} \tag{3.83}$$

The authors derived multiplicative rules for the Kullback-Leibler objective function:

$$\boldsymbol{H}_{f_0} \leftarrow \boldsymbol{H}_{f_0} \otimes \frac{\left(\boldsymbol{W}_{\Gamma}\boldsymbol{W}_{f_0}\right)^{\top}\frac{\boldsymbol{V}}{\hat{\boldsymbol{V}}}}{\left(\boldsymbol{W}_{\Gamma}\boldsymbol{W}_{f_0}\right)^{\top}} \tag{3.84}$$

$$\boldsymbol{W}_{f_0} \leftarrow \boldsymbol{W}_{f_0} \otimes \frac{\boldsymbol{W}_{\Gamma}^{\top}\frac{\boldsymbol{V}\boldsymbol{H}_{f_0}^{\top}}{\hat{\boldsymbol{V}}}^{\top}}{\boldsymbol{W}_{\Gamma}^{\top}\boldsymbol{H}_{f_0}^{\top}} \tag{3.85}$$

and:

$$\boldsymbol{W}_{\Gamma} \leftarrow \boldsymbol{W}_{\Gamma} \otimes \frac{\left(\boldsymbol{W}_{f_0}\boldsymbol{H}_{f_0}\right)^{\top}\frac{\boldsymbol{V}}{\hat{\boldsymbol{V}}}^{\top}}{\left(\boldsymbol{W}_{f_0}\boldsymbol{H}_{f_0}\right)^{\top}} \tag{3.86}$$

The authors also proposed another version of the model that reduces the number of free parameters by allowing frequency shifting in the logarithmic domain of the excitation filters.

Heittola et al. (2009) also proposed specific spectrum models for musical instruments source separation. However they set some of the factors to constant values and added an extra factor in order to limit the possible filter shapes. The method consisted of a source-filter model with an augmented NMF algorithm for instrument recognition and separation. The model can be expressed as:

$$\boldsymbol{V} \approx \hat{\boldsymbol{V}} = \left(\boldsymbol{W}_{\Gamma}\boldsymbol{H}_{\Gamma}\right) \otimes \left(\boldsymbol{W}_{f_0}\boldsymbol{H}_{f_0}\right) \tag{3.87}$$

The factor $\boldsymbol{W}_{f_0}$ is fixed for each frame given the results of a multipitch estimator. These are constructed by computing the spectra of a set of sinusoidal components of unity amplitude. The frequencies of the sinusoids are the harmonics of the detected fundamental frequency of a given frame f_0. The components of $\boldsymbol{W}_{\Gamma}$ are also fixed and consist of the spectra of a filterbank of triangular filters uniformly spaced on a Mel frequency scale. The gains $\boldsymbol{H}_{\Gamma}$ and $\boldsymbol{H}_{f_0}$ are learned using the derived update rules that minimize the Kullback-Leibler divergence:

$$\boldsymbol{H}_{f_0} \leftarrow \boldsymbol{H}_{f_0} \otimes \frac{\boldsymbol{W}_{f_0}^{\top}\frac{\boldsymbol{V}}{\hat{\boldsymbol{V}}}}{\boldsymbol{W}_{f_0}^{\top}\boldsymbol{1}} \tag{3.88}$$

and:

$$\boldsymbol{H}_{\Gamma} \leftarrow \boldsymbol{H}_{\Gamma} \otimes \frac{\boldsymbol{W}_{\Gamma}^{\top}\frac{\boldsymbol{V}}{\hat{\boldsymbol{V}}}^{\top}}{\boldsymbol{W}_{\Gamma}^{\top}\boldsymbol{1}} \tag{3.89}$$

where $\mathbf{1}$ is a matrix of ones the same size as $\boldsymbol{V}$. The authors also proposed a streaming method where the gains of previous frames are used in combination with a random matrix to initialize the gains of new frames. Finally instrument recognition is performed by applying MFCC feature extraction and Gaussian Mixture Models (GMM) classifications to the filter estimates.

Durrieu et al. (2009b, 2010, 2011) approximated the mixture spectrum as a combination of the spectrum of the monophonic lead instrument $\hat{\boldsymbol{X}}_v$ and the spectrum of the accompaniment $\hat{\boldsymbol{X}}_m$. The accompaniment was modeled as a standard NMF spectrum model $\hat{\boldsymbol{X}}_m = \boldsymbol{W}_m\boldsymbol{H}_m$. However for the lead instrument a source-filter model was used $\hat{\boldsymbol{X}}_v = \boldsymbol{X}_\Phi \otimes \boldsymbol{X}_{f_0}$, where both the source and the filter components are further decomposed into basis and gains matrices as $\hat{\boldsymbol{X}}_\Phi = \boldsymbol{W}_\Gamma\boldsymbol{H}_\Gamma\boldsymbol{H}_\Phi$ and $\hat{\boldsymbol{X}}_{f_0} = \boldsymbol{W}_{f_0}\boldsymbol{H}_{f_0}$. By performing this decoupling between the excitation and the filter of the lead instrument, the model reduces significantly the degrees of freedom and the number of free parameters. This model is called the Smoothed Instantaneous Mixing Model (SIMM), in which the observed mixture spectrum $\boldsymbol{V}$ is approximated by the spectrum model $\hat{\boldsymbol{V}}$ in the following way:

$$\boldsymbol{V} \approx \hat{\boldsymbol{V}} = (\boldsymbol{W}_\Gamma\boldsymbol{H}_\Gamma\boldsymbol{H}_\Phi) \otimes (\boldsymbol{W}_{f_0}\boldsymbol{H}_{f_0}) + \boldsymbol{W}_m\boldsymbol{H}_m \tag{3.90}$$

where $\boldsymbol{W}_\Gamma$ and $\boldsymbol{W}_{f_0}$ are fixed matrices and the rest are learned from the data. $\boldsymbol{W}_\Gamma$ is fixed to a set of filters of a smooth filterbank that represents the resonant body of the lead instrument. $\boldsymbol{H}_\Gamma$ represents the different possible weight combinations of the resonant body filter. $\boldsymbol{H}_\Phi$ represents the evolution of the filter of the lead instrument, this evolution is modeled as a linear combination of the filters generated by $\boldsymbol{W}_\Gamma\boldsymbol{H}_\Gamma$. The $\boldsymbol{W}_{f_0}$ is fixed to a set of harmonic combs representing the excitation spectra. When targeting voice signals as the lead instrument, the envelope of the excitation spectra can be set using the glottal source model by Klatt and Klatt (1990) (see 2.7). $\boldsymbol{H}_{f_0}$ represents the pitch evolution of the lead instrument. Finally $\boldsymbol{W}_m$ and $\boldsymbol{H}_m$ represent the accompaniment music without any specific model of the spectrum other than the standard NMF. The authors also proposed a version of this method where the smoothness constraint on the filterbank is not enforced. This is done avoiding the factorization $\boldsymbol{W}_\Phi = \boldsymbol{W}_\Gamma\boldsymbol{H}_\Gamma$ and leaving these filters to adopt any shape. The resulting method receives the name of Instantaneous Mixing Model (IMM). The authors used a two step procedure to estimate the factors presented. The first step is a rough estimation of the pitch where the matrix $\boldsymbol{H}_{f_0}$ is completely unconstrained and learned. In a second step the maximum pitch is selected and tracked in $\boldsymbol{H}_{f_0}$. Then a monophonicity constraint is applied to $\boldsymbol{H}_{f_0}$, that sets to 0 the

values for all the pitches other than those around the tracked pitch. Finally the other free parameters of the model are learned. In both stages the β -divergence objective function is used (3.29), and the update rules become:

$$\boldsymbol{H}_{f_0} \leftarrow \boldsymbol{H}_{f_0} \otimes \frac{\boldsymbol{W}_{f_0}^{\top}\left(\hat{\boldsymbol{X}}_{\Phi} \otimes \hat{\boldsymbol{V}}^{(\beta-2)} \otimes \boldsymbol{V}\right)}{\boldsymbol{W}_{f_0}^{\top}\left(\hat{\boldsymbol{X}}_{\Phi} \otimes \hat{\boldsymbol{V}}^{(\beta-1)}\right)} \tag{3.91}$$

$$\boldsymbol{H}_{\Phi} \leftarrow \boldsymbol{H}_{\Phi} \otimes \frac{\left(\boldsymbol{W}_{\Gamma}\boldsymbol{H}_{\Gamma}\right)^{\top}\left(\hat{\boldsymbol{X}}_{f_0} \otimes \hat{\boldsymbol{V}}^{(\beta-2)} \otimes \boldsymbol{V}\right)}{\left(\boldsymbol{W}_{\Gamma}\boldsymbol{H}_{\Gamma}\right)^{\top}\left(\hat{\boldsymbol{X}}_{f_0} \otimes \hat{\boldsymbol{V}}^{(\beta-1)}\right)} \tag{3.92}$$

$$\boldsymbol{H}_{m} \leftarrow \boldsymbol{H}_{m} \otimes \frac{\boldsymbol{W}_{m}^{\top}\left(\hat{\boldsymbol{V}}^{(\beta-2)} \otimes \boldsymbol{V}\right)}{\boldsymbol{W}_{m}^{\top}\hat{\boldsymbol{V}}^{(\beta-1)}} \tag{3.93}$$

$$\boldsymbol{H}_{\Gamma} \leftarrow \boldsymbol{H}_{\Gamma} \otimes \frac{\boldsymbol{W}_{\Gamma}^{\top}\left(\hat{\boldsymbol{X}}_{f_0} \otimes \hat{\boldsymbol{V}}^{(\beta-2)} \otimes \boldsymbol{V}\right)\boldsymbol{H}_{\Phi}^{\top}}{\boldsymbol{W}_{\Gamma}^{\top}\left(\hat{\boldsymbol{X}_{f_0}} \otimes \hat{\boldsymbol{V}}^{(\beta-1)}\right)\boldsymbol{H}_{\Phi}^{\top}} \tag{3.94}$$

$$\boldsymbol{W}_{m} \leftarrow \boldsymbol{W}_{m} \otimes \frac{\left(\hat{\boldsymbol{V}}^{(\beta-2)} \otimes \boldsymbol{V}\right)\boldsymbol{H}_{m}^{\top}}{\hat{\boldsymbol{V}}^{(\beta-1)}\boldsymbol{H}_{m}^{\top}} \tag{3.95}$$

Ozerov et al. (2010) proposed a general framework allowing flexible inclusion of the different aspects of most spectrum models that have been presented. The framework, named Flexible Audio Source Separation Toolbox (FASST), consists of a spectrum model that can be hierarchically decomposed into factors that reconstruct the different components in musical signals. It contains a set of constant factors for wideband or harmonic sustained sources that target sustained notes. Additionally they proposed a set of factors for wideband or harmonic fast decaying elements, which models transient sounds such as attacks or note onsets. It also further decomposes the sustained harmonic components following a source-filter model. The spectrum model can be resumed as:

$$\boldsymbol{V} \approx \hat{\boldsymbol{V}} = \sum_{m}^{N_M}(\boldsymbol{W}_{f_0}^{m}\boldsymbol{H}_{f_0}^{m}\boldsymbol{H}_{t_{env}}^{m}\boldsymbol{W}_{t_{env}}^{m}) \otimes \boldsymbol{W}_{\Gamma}^{m} \tag{3.96}$$

where $\boldsymbol{W}_{f_0}^{m}$ is a dictionary of narrowband spectral patterns, containing a set of bases with harmonic structures and another set of bandpass filters

bases. $\boldsymbol{H}^m_{f_0}$ is a matrix of spectral pattern weights that are used to create linear combinations of the bases in $\boldsymbol{W}^m_{f_0}$ that will represent the characteristic spectral patterns present of the m^{th} source in the mixture. The spectral patterns are then modulated in time using the matrices $\boldsymbol{H}^m_{t_{env}}$ and $\boldsymbol{W}^m_{t_{env}}$. $\boldsymbol{H}^m_{t_{env}}$ contains the temporal pattern weights and $\boldsymbol{W}^m_{t_{env}}$ is composed of time localized patterns. Finally, $\boldsymbol{W}^m_{\Gamma}$ contains the filters applied to the spectral patterns at each frame. Despite the large number of components, many of them are constrained to predetermined constant values ($\boldsymbol{W}^m_{f_0}$ and $\boldsymbol{W}^m_{t_{env}}$) and therefore the number of free parameters is much lower. Several generic iterative solvers are proposed allowing an automatic solving of the factorization problem.

NMF Derivations and Interpretations

Various different derivations of the NMF update rules have been proposed in the literature. The differences between them are in the perspective taken or the framework used. It is worth explaining here some of the most widely used derivations.

The problem of non-negative matrix factorization was initially viewed as a gradient descent problem with a specially selected step. Lee and Seung (2001) were the first to derive a set of multiplicative update rules that ensured that the cost function would not increase. The nonincreasing of the cost function was enforced by using an auxiliary function that serves as an upper bound.

While considering the addition of other regularization terms to the cost function Virtanen (2007) used a gradient descent method. In this derivation the gradient of the cost function is calculated and it is partitioned into additive and substractive elementwise nonnegative terms. This can be done because the elements of the gradients are only formed of elements of the data $\boldsymbol{V}$, of the basis $\boldsymbol{W}$ and of the gains $\boldsymbol{H}$ that are all non-negative. In order to descend the gradient, multiplicative rules are defined where the multplicative term is the ratio between the substractive term and the additive term.

Another approach to deriving the multiplicative update rules is to use the Karush-Kuhn-Tucker (KKT) conditions (Sra and Dhillon, 2006; Dhillon and Sra, 2006) to generalize the method of Lagrange multipliers to inequality constraints. The authors viewed the NMF problem, called Non-negative Matrix Approximation (NNMA), as a minimization of the objective function given an inequality constraint for the non-negativity of the factors. Using

such a derivation the authors easily found multiplicative update rules for the Bregman divergences including generic regularization terms on the gains and the basis. Additionally this derivation technique allowed the authors to solve more complex factorization models, with more than two factors.

Finally it has become widely popular to derive the solutions to the NMF problem under a probabilistic formulation. As presented by Winther and Petersen (2007) and Schmidt and Laurberg (2008), the problem of factorizing $\boldsymbol{V} = \boldsymbol{W}\boldsymbol{H} + \boldsymbol{B}$ can be formulated as a Maximum Likelihood estimation. Where $\boldsymbol{V}$ are the observations, $\boldsymbol{W}$ and $\boldsymbol{H}$ the parameters and $\boldsymbol{B}$ is the residual. By assuming the elements of the residual to be independent, identically distributed and Gaussian random variables with variance σ_N^2 we can express the likelihood in the following way:

$$\mathcal{P}(\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H})_{euc} = \frac{1}{\sqrt{2\pi}\sigma_N} e^{\left(-\frac{\|\boldsymbol{V}-\boldsymbol{W}\boldsymbol{H}\|^2}{2\sigma_N^2}\right)} \tag{3.97}$$

It is trivial to show that maximizing the likelihood function corresponds to minimizing the negative log likelihood, which serves as a cost function and is equal to:

$$\mathcal{L}^{euc}_{\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H}}(\boldsymbol{W},\boldsymbol{H}) = \frac{1}{2\sigma_N^2}\|\boldsymbol{V} - \boldsymbol{W}\boldsymbol{H}\|^2 + \gamma \tag{3.98}$$

where γ is a constant with respect to $\boldsymbol{W}$ and $\boldsymbol{H}$. Therefore the update rules corresponding to the Euclidean distance cost function 3.25 and 3.24, can be interpreted as a Maximum Likelihood Estimator (MLE).

The update rules of the other cost functions can be interpreted in a probabilistic framework as Maximum Likelihood estimators of different random processes (Févotte et al., 2009). For the Itakura-Saito divergence, a frame of the spectrogram is modeled as a sum of random variable components:

$$\widetilde{\boldsymbol{V}}_t = \sum_{w=1}^{N_W} \boldsymbol{c}_{w,t} \tag{3.99}$$

where each component follows a proper multivariate complex Gaussian distribution $\boldsymbol{c}_{w,t} \sim \mathcal{N}_c(0, \boldsymbol{H}_{w,t} diag(\boldsymbol{W}_w))$ and the proper complex Gaussian distribution is $\mathcal{N}_c(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = |\pi\Sigma|^{-1} e^{-(\boldsymbol{x}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$. Due to the additivity of the Gaussian distributions, the likelihood can be expressed as:

$$\mathcal{P}(\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H})_{is} = \sum_{t=1}^{N_T}\sum_{\omega=1}^{N_\omega} \mathcal{N}_c\left(\boldsymbol{V}_{\omega,t}|0, \sum_{w=1}^{N_W} \boldsymbol{W}_{\omega,w}\boldsymbol{H}_{w,t}\right) \tag{3.100}$$

and the negative log likelihood can be shown to be:

$$\mathcal{L}^{is}_{\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H}}(\boldsymbol{W},\boldsymbol{H}) = \sum_{t=1}^{N_T}\sum_{\omega=1}^{N_\omega} log\left(\sum_{w=1}^{N_W}\boldsymbol{W}_{\omega,w}\boldsymbol{H}_{w,t}\right) + \frac{|\widetilde{\boldsymbol{V}}_{\omega,t}|^2}{\sum_{w=1}^{N_W}\boldsymbol{W}_{\omega,w}\boldsymbol{H}_{w,t}} + \gamma \tag{3.101}$$

where γ is a constant with respect to $\boldsymbol{W}$ and $\boldsymbol{H}$. Therefore the update rules corresponding to the Itakura-Saito distance cost function 3.40 and 3.39, can be interpreted as a Maximum Likelihood Estimator (MLE). Finally the Kullback-Leibler divergence can be related (Févotte et al. (2009)) to the Maximum Likelihood Estimator where the magnitude of the spectral bins are modeled as a sum of random variables following the Poisson distribution:

$$|\widetilde{\boldsymbol{V}}_t| = \sum_{w=1}^{N_W}|\boldsymbol{c}_{w,t}| \tag{3.102}$$

where $|c_{w,\omega,t}| \sim \mathcal{P}(\boldsymbol{W}_{\omega,w}\boldsymbol{H}_{w,t})$, where $\mathcal{P}(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}$ is the Poisson distribution. However one must bear in mind that the Poisson distribution is defined for integer random variables and the magnitude of a spectrum bin is a continuous value. In order to adapt the data one can assume an appropriate scaling and a fine grain quantization, this draws interesting parallelisms between a spectrum and a histogram (Shashanka et al. (2008)).

Shashanka et al. (2008) proposed interpreting a normalized version of the spectrogram as a probability distribution $\mathcal{P}(X_\omega, X_t)$ of two multinomial random variables $X_\omega \in \{1 \dots N_\omega\}$ and $X_t \in \{1 \dots N_T\}$. In other words, it can be viewed as the probability distribution of randomly drawing a given time-frequency bin. The authors then propose two ways to model $\mathcal{P}(X_t, X_\omega)$. The symmetric factorization model is characterized by the independence of the conditional distributions of the temporal and frequency dimensions given the index of a latent class (or hidden component) Z_w:

$$\mathcal{P}(X_\omega, X_t) = \sum_{w}^{N_W}\mathcal{P}(Z_w)\mathcal{P}(X_\omega|Z_w)\mathcal{P}(X_t|Z_w) \tag{3.103}$$

This model is equivalent to the Probabilistic Latent Component Analysis (PLCA) method (Smaragdis and Raj (2007)). By representing the probability distributions $\mathcal{P}(X_\omega, X_t)$, $\mathcal{P}(Z_w)$, $\mathcal{P}(X_\omega|Z_w)$, $\mathcal{P}(X_t|Z_w)$ as the matrices

$\boldsymbol{P}$, $\boldsymbol{S}$, $\boldsymbol{W}$ and $\boldsymbol{G}$ we arrive at $\boldsymbol{P} = \boldsymbol{WSG}$ which is equivalent to NMF's well known $\boldsymbol{V} = \boldsymbol{WH}$. The parameter can be estimated using the EM algorithm which leads to the NMF multiplicative rules with a scaling factor difference, since $\boldsymbol{P}$ is a normalized version of $\boldsymbol{V}$.

The other model presented by the authors for the probability distribution $\mathcal{P}(X_\omega, X_t)$ is called asymmetric factorization:

$$\mathcal{P}(X_\omega, X_t) = \mathcal{P}(X_i)\mathcal{P}(X_j|X_i) \tag{3.104}$$

$$\mathcal{P}(X_j, X_i) = \sum_{Z_w}^{N_W} \mathcal{P}(X_j|Z_w)\mathcal{P}(Z_w|X_i) \tag{3.105}$$

where $i, j \in 1, 2$, $i \neq j$. This method is popularly known as Probabilistic Latent Semantic Analysis (PLSA). As in the previous case, this approach leads to the NMF multiplicative update rules when using a matrix representation and an EM algorithm. This relation was previously shown by Gaussier and Goutte (2005).

This way to derive the NMF updates allowed the authors to derive update rules for other more complex spectrum models. The main benefit of such a perspective is that it provides an easy way of generalizing the derivation of update rules to data in more than 2 dimensions. The authors derived multiplicative update rules for convolutive NMF and for Non-negative Tensor Factorization (NTF).

Another probabilistic approach to derive the NMF update rules consists in modeling the error $\boldsymbol{E}$ between the observed $\boldsymbol{V}$ and estimated spectra $\hat{\boldsymbol{V}} = \boldsymbol{WH}$ as random variables. This is equivalent to assuming a certain probability distribution for $p(\boldsymbol{V}_{\omega,t}|\hat{\boldsymbol{V}}_{\omega,t})$. For instance by modeling the residual as an additive Gaussian noise, the Maximum Likelihood estimation results in the Euclidean Distance cost function update rules. If $p(\boldsymbol{V}_{\omega,t}|\hat{\boldsymbol{V}}_{\omega,t})$ is assumed to follow a Poisson distribution, the ML estimation corresponds to the Kullback-Leibler divergence update rules. Finally if we assume multiplicative noise following a Gamma distribution, the ML estimation leads to the Itakura-Saito divergence update rules.

Schmidt and Laurberg (2008) derived the NMF decomposition with regularizations by computing the Maximum A Posteriori (MAP) estimator. This allows adding *a priori* knowledge to the NMF factorization. This derivation is inspired by the Maximum Likelihood estimator previously presented.

Following the Bayes rule:

$$\mathcal{P}(\boldsymbol{W},\boldsymbol{H}|\boldsymbol{V}) \quad = \quad \frac{\mathcal{P}(\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H})\mathcal{P}(\boldsymbol{W},\boldsymbol{H})}{\mathcal{P}(\boldsymbol{V})} \tag{3.106}$$

and assuming the denominator $\mathcal{P}(\boldsymbol{V})$ as constant. The log likelihood of the basis and gains is defined as follows:

$$\mathcal{L}_{\boldsymbol{W},\boldsymbol{H}|\boldsymbol{V}}(\boldsymbol{W},\boldsymbol{H}) \quad \propto \quad \mathcal{L}_{\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H}}(\boldsymbol{W},\boldsymbol{H}) + \mathcal{L}_{\boldsymbol{W},\boldsymbol{H}}(\boldsymbol{W},\boldsymbol{H}) \tag{3.107}$$

Non-negative Sparse Coding (NNSC) update rules (see Equation 3.45) correspond to a MAP estimator when the probability is set as follows:

$$\mathcal{P}^{NNSC}(\boldsymbol{W},\boldsymbol{H}) \quad = \quad \prod_{i,j} \lambda e^{-\lambda \boldsymbol{H}_{i,j}} \tag{3.108}$$

The authors also proposed a general family of priors that use a linking function and a set of underlying Gaussian Processes, specified by their covariances. This is referred to as Gaussian Process Priors NMF (GPP-NMF). It is derived by defining the priors on the basis and gains as independent and each of them belonging to a distribution of the following form:

$$\begin{aligned}\mathcal{P}^{GPPNMF}(\boldsymbol{H}) \quad \propto \quad & \exp\left(-\frac{1}{2} f_{\boldsymbol{H}}(vec(\boldsymbol{H}))^{\top} \Sigma_{\boldsymbol{H}}^{-1} f_{\boldsymbol{H}}(vec(\boldsymbol{H}))\right) \\ & \prod_{i} |f'_{\boldsymbol{H}}(vec(\boldsymbol{H}))|_i \end{aligned} \tag{3.109}$$

where $vec()$ is an operator that rearranges the elements of a matrix into a vector. $f_{\boldsymbol{H}} : \Re^{+} \rightarrow \Re$ is a strictly increasing and derivable linking function that maps the non-negative elements of $\boldsymbol{H}$ to elements drawn from a multivariate Gaussian distribution. This derivation leads to update rules such as those used by Wilson et al. (2008a) (see Equation 3.63).

Another probabilistic interpretation of the NMF regularizations was developed by Virtanen et al. (2008a) who proposed a regularized NMF where temporal discontinuities are penalized, by imposing priors on the gains. To derive the prior probabilities the gains are modeled as Gamma chains.

NMF has been linked to previous well-established methods from other fields. Gaussier and Goutte (2005) showed that the NMF with the Kullback-Leibler cost function is equivalent to the Probabilistic Latent Semantic Indexing (PLSI) proposed by Hofmann (1999) and often used in text analysis works.

Smaragdis et al. (2007) proposed a straightforward extension of PLSI named Probabilistic Latent Component Analysis (PLCA), where the spectra are interpreted as probabilistic distributions and where the estimation of the components' probabilities are found using EM. A similar interpretation is used in the work of Durrieu et al. (2009b) to derive the proposed update rules.

3.5 Evaluation

As in all research problems a proper evaluation must be defined to assess the performance of the proposed solutions. For blind source separation there have been many proposals of evaluation frameworks as well as data sets. In this section we will review the most popular and established evaluation methodologies for BSS problems.

Performance Measures

Lambert (1999) and Schobben and Torkkola (1999) proposed the first set of generic BSS evaluation measures and data in order to compare different speech enhancement and separation methods. Before their proposal, most evaluation work was performed specifically for each solution. Additionally evaluations were often based on indirect performance measurements or subjective tests such as speech recognition rates, plots of separated signals, plots of cascaded mixing/unmixing impulse responses and signal to noise ratios. Finally, since most evaluation measures were different for each algorithm, comparing results between algorithms was difficult or impossible. Schobben and Torkkola (1999) provided a unified methodology of evaluating BSS algorithms, by providing measures and data. The methodology targets a wide range of BSS applications focusing on the context of acoustic speech data. The authors propose two categories of test cases. First, a number of controllable synthetic separation problems, useful for testing the limits of the solutions and allowing a straightforward performance measure since the sources are available. Second, a set of real world recordings of mixtures and of their individual sources which are available online [4] for future research in the area. The authors proposed a set of parameters defining the difficulty of the problem as well as a protocol to perform the mixture and sources recordings. It is worth noting that the mixing process is also applied during the recording of the individual sources, which leads to multiple mixture

4. http://web.archive.org/web/19991004145501/http://www.ele.tue.nl/ica99/

signals per source recording. In the methodology, $v_{o,x_m}[n]$ refers to the mixture v_o when the source x_m is recorded (or synthetically mixed) in isolation. Finally the authors propose two objective measures given the recording of the individual sources $v_{o,x_m}[n]$ and the separated sources $\hat{x}_m[n]$. The first measure is **distortion**, a measure of how distorted the estimated source is with respect to the actually recorded source. This distortion measure is defined as:

$$D_m^{distortion} = 10log_{10}\left(\frac{E\left[(v_{m,x_m} - \alpha_m \hat{x}_m)^2\right]}{E\left[v_{m,x_m}^2\right]}\right) \tag{3.110}$$

with $\alpha_m = E\left[v_{m,x_m}^2\right] / E\left[\hat{x}_m^2\right]$ and where the separated sources' indices have been chosen such that $\hat{x}_m$ corresponds to the m^{th} source. This measure is robust to scaling and permutation indetermination in source separation tasks. The second measure proposed assesses the amount of separation performed. And is defined as:

$$D_m^{separation} = 10log_{10}\left(\frac{E\left[\hat{x}_{m,x_m}^2\right]}{E\left[\sum_{i\neq j}\hat{x}_{j,x_i}^2\right]}\right) \tag{3.111}$$

where $\hat{x}_{j,x_i}$ is the j^{th} output of the cascaded mixing/unmixing system when only x_i is active.

With a closer focus on audio and as a preliminary step to define a global BSS evaluation framework, Vincent et al. (2003) proposed a series of tasks and applications of Blind Audio Source Separation (BASS). The authors defined two main categories of BASS applications, a set of Audio Quality Oriented (AQO) applications and another set of Significance Oriented (SO) applications. The main objective of the first group of applications is to achieve a good Signal to Noise Ratio (SNR) and reduce artifacts in the separated sources or the remixed signals. While the second group focuses on retrieving features and/or descriptions of the audio scene by estimating the sources and/or parameters of the mixing process.

Following this line of work, Gribonval et al. (2003) and Vincent et al. (2006) proposed several evaluation measures that also take into account the gain indeterminacies of the BSS algorithms. In their new approach the total error term e_m^{total} between each real source x_m and the source separated by the BSS method $\hat{x}_m$ is decomposed into three different type of errors ($e^{total} = e^{interf} + e^{noise} + e^{artif}$): the interferences from other sources, the

distortion from the noise and the artifacts from the algorithm. The authors proposed a measure for each type of error.

The authors also showed the existence of an upper bound of Source to Interference Ratio in most common source separation problems. They proposed the use of such an upper bound to assess performance and compare different BSS methods.

The total distortion between a separated source and the true source is defined as:

$$D_m^{total} = \frac{\|\hat{x}_m\|^2 - |\langle \hat{x}_m, x_m \rangle|^2}{|\langle \hat{x}_m, x_m \rangle|^2} \tag{3.112}$$

where in the general case one can safely assume that the different sources may be correlated but are still linearly independent. $\boldsymbol{P}_x$ is considered the orthogonal projector onto the subspace spanned by the set of source signals $\{x_m[n]\}_{m=1}^{N_M}$ and $\boldsymbol{P}_{x,\boldsymbol{B}}$ the orthogonal projector onto the subspace spanned by the source signals together with the noise signals $\{\boldsymbol{B}_o[n]\}_{o=1}^{N_O}$. Where $\boldsymbol{B}_o[n]$ is the additive noise signal on the o^{th} mixture.

Under such assumptions the error terms can be expressed as:

$$e_m^{interf} = \boldsymbol{P}_x \hat{x}_m - \langle \hat{x}_m, x_m \rangle x_m \tag{3.113}$$
$$e_m^{noise} = \boldsymbol{P}_{x,\boldsymbol{B}} \hat{x}_m - \boldsymbol{P}_x \hat{x}_m \tag{3.114}$$
$$e_m^{artif} = \hat{x}_m - \boldsymbol{P}_{x,\boldsymbol{B}} \hat{x}_m \tag{3.115}$$

$\boldsymbol{P}_x \hat{x}_m$ can be computed as:

$$\boldsymbol{P}_x \hat{x}_m = \sum_{l=1}^{N_M} c_l x_l = \boldsymbol{c}^\top \boldsymbol{x} \tag{3.116}$$
$$\boldsymbol{c} = conj(\boldsymbol{G})^{-1} \boldsymbol{d} \tag{3.117}$$
$$\boldsymbol{G} = \boldsymbol{x}\boldsymbol{x}^H \tag{3.118}$$
$$[\boldsymbol{d}]_k = \langle \hat{x}_m, x_k \rangle \tag{3.119}$$

and $\boldsymbol{P}_{x,\boldsymbol{B}} \hat{x}_m$ can be calculated as:

$$\boldsymbol{P}_{x,\boldsymbol{B}} \hat{x}_m = \boldsymbol{P}_x \hat{x}_m + \sum_{o=1}^{N_O} \langle \hat{x}_m, \boldsymbol{B}_o \rangle \boldsymbol{B}_o / \|\boldsymbol{B}_o\|^2 \tag{3.120}$$

Using such error terms we can then define the relative distortions. The interference distortion is the error term due to interference from the other

sources:

$$D_m^{interf} = \frac{\|e_m^{interf}\|^2}{|\langle \hat{x}_m, x_m \rangle|^2} \tag{3.121}$$

The relative distortion due to additive noise considered linearly independent from all the sources can be defined as:

$$D_m^{noise} = \frac{\|e_m^{noise}\|^2}{\|\langle \hat{x}_m, x_m \rangle x_m + e_m^{interf}\|^2} \tag{3.122}$$

The relative distortion due to artifacts added by the algorithm is expressed in the following form:

$$D_m^{artif} = \frac{\|e_m^{artif}\|^2}{\|\langle \hat{x}_m, x_m \rangle x_m + e_m^{interf} + e_m^{noise}\|^2} \tag{3.123}$$

Finally the widely used Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), Signal to Noise Ratio (SNR) and Signal to Artifact Ratio (SAR) are defined as:

$$SDR_m = 10log_{10}\left(\frac{1}{D_m^{total}}\right) \tag{3.124}$$

$$SIR_m = 10log_{10}\left(\frac{1}{D_m^{interf}}\right) \tag{3.125}$$

$$SNR_m = 10log_{10}\left(\frac{1}{D_m^{noise}}\right) \tag{3.126}$$

$$SAR_m = 10log_{10}\left(\frac{1}{D_m^{artif}}\right) \tag{3.127}$$

Vincent et al. (2007c) created a new set of quantitative measures specific to stereo signals based on the previously presented measures. In this case the error of the signal is decomposed into the following set of components:

$$\hat{x}_{o,m} = x_{o,m} + e_{o,m}^{spat} + e_{o,m}^{interf} + e_{o,m}^{artif} \tag{3.128}$$

$$e_{o,m}^{spat} = \boldsymbol{P}_{m,L}\hat{x}_{o,m} - x_{o,m} \tag{3.129}$$

$$e_{o,m}^{interf} = \boldsymbol{P}_{all,L}\hat{x}_{o,m} - \boldsymbol{P}_{m,L}\hat{x}_{o,m} \tag{3.130}$$

$$e_{o,m}^{artif} = \hat{x}_{o,m} - \boldsymbol{P}_{all,L}\hat{x}_{o,m} \tag{3.131}$$

where $\hat{x}_{o,m}$ and $x_{o,m}$ are the estimated and true signals for the o^{th} channel of the m^{th} source. $\boldsymbol{P}_{x,L}$ is the orthogonal projector onto the subspace spanned

by all the channels of delayed versions of the m^{th} source, $x_{k,m}[t-\tau]$ for $1 \leq k \leq N_O$, $0 \leq \tau \leq L-1$. $\boldsymbol{P}_{all,L}$ is the orthogonal projector onto the subspace spanned by all the channels of delayed versions of all the sources. The filter length L is set to the largest tractable value (512 samples or 32 ms). The energy ratios are derived as in the previous approach:

$$ISR_m = 10log_{10}\left(\frac{\sum_o \sum_t x_{o,m}^2}{\sum_o \sum_t e_{o,m}^{spat\,2}}\right) \quad (3.132)$$

$$SIR_m = 10log_{10}\left(\frac{\sum_o \sum_t (x_{o,m} + e_{o,m}^{spat})^2}{\sum_o \sum_t e_{o,m}^{interf\,2}}\right) \quad (3.133)$$

$$SAR_m = 10log_{10}\left(\frac{\sum_o \sum_t (x_{o,m} + e_{o,m}^{spat} + e_{o,m}^{interf})^2}{\sum_o \sum_t e_{o,m}^{artif\,2}}\right) \quad (3.134)$$

$$SDR_m = 10log_{10}\left(\frac{\sum_o \sum_t x_{o,m}^2}{\sum_o \sum_t (e_{o,m}^{spat} + e_{o,m}^{interf} + e_{o,m}^{artif})^2}\right) \quad (3.135)$$

These measures were used in the first stereo audio source separation evaluation campaign (Vincent et al., 2009), and several underdetermined source separation algorithms were tested.

The main objection to such a quantitative evaluation framework is that it may not correctly represent the perceptual quality of the separation. The proposed measures do not take into account important auditory phenomena such as loudness perception and spectral masking. Emiya et al. (2011) tried to overcome this limitation by proposing a subjective test protocol to assess the perceived quality of an audio source separation task. Additionally the authors propose a set of objective measures aiming to predict the resulting subjective scores. These new measures are based on a multiband processing of the error components. An auditory filterbank based on ERB Gammatone filters is used for decomposition. Additionally a Perceptual Similarity Metric (PSM), provided by the PEMO-Q auditory model, is computed for each of the error components. Finally these metrics are fed to a nonlinear mapping to compute the new measures named Overall Perceptual Score (OPS), Target-related Perceptual Score (TPS), Interference-related Perceptual Score (IPS) and Artifacts-related Perceptual Score (APS). This evaluation framework is freely and publicly distributed online [5] as the Perceptual Evaluation of Audio Source Separation (PEASS).

5. http://bass-db.gforge.inria.fr/peass/

Oracle Estimators

The performance of a source separation technique is highly dependent on the data. It would not be sensible to directly aggregate this measure for many data instances. Likewise we cannot draw conclusions from a small number of data examples. In order to overcome this problem, Vincent et al. (2007a) derived oracle estimators for several types of methods. These estimators find the optimal separation of a method given the true sources. In order to use the oracle BSS estimators we need a controlled scenario in which the true sources are known. These oracle separations can then be used as an upper bound of the performance measures. We are then able to get a performance relative to the best case scenario. This strategy leads to a smaller dependence of the performance on the data.

Vincent et al. (2007a) proposed oracle estimators for three classes of methods: multichannel time-invariant filtering, single-channel time-frequency masking and multichannel time-frequency masking. In order to derive the oracle estimators, the authors defined a separation method as:

$$\hat{\boldsymbol{x}} \quad = \quad f(\boldsymbol{v}, \theta) \text{ with } \theta \in \Theta \tag{3.136}$$

where f is a fixed function, which given a set of mixtures $\boldsymbol{v}$ and a set of parameters θ returns a set of estimated separated sources $\hat{\boldsymbol{x}}$. Θ is the set of possible parameters sets, that can be defined by the constraints that delimit such set. Given a function f the oracle estimator is defined as the parameter set $\widetilde{\theta}(\boldsymbol{v}, \boldsymbol{x}, \Theta)$ that minimizes a distortion measure $d(\boldsymbol{x}, \hat{\boldsymbol{x}})$ for a set of sources $\boldsymbol{x}$ and mixtures $\boldsymbol{v}$:

$$\widetilde{\theta}(\boldsymbol{v}, \boldsymbol{x}, \Theta) \quad = \quad \arg\min_{\theta \in \Theta} d(\boldsymbol{v}, f(\boldsymbol{x}, \theta)) \tag{3.137}$$

Note that in a controlled situation where the true sources $\boldsymbol{x}$ are known, the oracle estimator $\widetilde{\theta}$ may be found using an exhaustive search over the parameter space Θ. The authors propose three different functions f and methods to compute the optimal parameters $\widetilde{\theta}(\boldsymbol{v}, \boldsymbol{x}, \Theta)$ for the distortion measure D_m^{total} as defined in 3.112. The proposed oracle estimators were made available online (see Vincent et al. (2007b)).

Datasets

Several datasets of multitrack audio recordings have been made publicly available to assess the separation methods in different scenarios and applications. Vinyes (2008) and Vinyes et al. (2006) prepared a database

(MASS [6]) to help evaluate Musical Audio Signal Separation algorithms and statements on a representative set of professionally produced music (i.e. real music recordings).

Vincent et al. (2007a,b) also included a set of of multitrack audio examples in their toolbox (BSS Oracle [7]) to create oracle estimators for musical audio source separation methods.

As part of the SiSEC [8] source separation evaluation campaign, the organizers published a set of multitrack audio excerpts to prepare submissions to the professionally produced music recordings task.

Recently Hsu and Jang (2010b) published a large dataset (MIR-1K [9]) of 1000 song clips containing singing voice and accompaniment in separate tracks. The dataset is designed to evaluate singing voice separation methods. The dataset also contains manual annotations of the pitch contours, unvoiced frames, lyrics and voiced-unvoiced segments. Speech recordings of the lyrics performed by the singer of each song are included in the dataset.

3.6 Summary of Part I

In this first part we presented the basic theoretical framework to describe and study the source separation problem, and reviewed currently available techniques to perform audio and music source separation. We paid special attention to signal decomposition methods, and more specifically NMF, due to their flexibility and widespread use in the context of audio. We also presented the methods and datasets that are used nowadays to evaluate performance. Our main objectives consist in developing methods that are more adapted to real world and practical applications. To achieve this we take two separate paths. In Part II we focus on reducing the computational complexity and latency of current existing methods, sacrificing separation quality if necessary. In Part III we concentrate on the quality of the separation and propose methods that target the separation and isolation of certain components that current techniques do not take into account. In the latter chapter we do not impose latency constraints, thereby allowing batch processing and offline user guidance.

6. http://www.mtg.upf.edu/static/mass
7. http://bass-db.gforge.inria.fr/bss_oracle/
8. http://sisec.wiki.irisa.fr/
9. https://sites.google.com/site/unvoicedsoundseparation/mir-1k

Part II

Chapter 4

Low Latency Pitch Estimation and Tracking

4.1 Introduction

Pitch estimation and tracking is often used as a first step in audio source separation of mixes. Acoustic sources very often concentrate most of their energy in a harmonic structure. Pitch estimation consists in finding the fundamental frequency of such a structure, which can then be used to estimate the time-frequency location of its energy. Estimating the pitch of a particular target source is crucial in music source separation, where frequently many pitched sources are simultaneously present. This task is similar to melody track estimation, which has been widely studied by the Music Information Retrieval (MIR) community. However in some cases melody may be interpreted by multiple sources and in such cases a different approach should be taken. Furthermore most melody estimation techniques presented until now do not enforce latency-limited constraints. Here we focus on lowering the computational cost and latency of existing source separation methods. Since pitch estimation and tracking is often a required step of many monaural music source separation methods, these processes must also work under low-latency and low computational cost conditions.

In this chapter we introduce the Tikhonov regularization approach to matrix factorization. This method presents interesting qualities for computations under low-latency, low computational cost and realtime constraints. We propose using this method as an alternative to Non-negative Matrix Factorization (NMF) in predominant and multiple pitch estimation scenarios.

The proposed method is tailored to target specific sources by means of timbre models. Additionally the method is designed to operate within the latency specified by the user.

Our proposed technique is compared to other well-established techniques in the context of predominant pitch estimation. Finally, we propose modifications to improve the use of Tikhonov regularization in tasks of multiple pitch estimation, specifically targeting issues related to glissandos and vibratos that are common in singing voice scenarios. We also target problems related to octave errors which are quite common when working with joint pitch likelihood estimation, such as NMF and Tikhonov regularization.

4.2 Tikhonov Regularization

Over the past decade Non-negative Matrix Factorization (NMF) has been gaining a lot of attention in tasks of pitch estimation and source separation in music signals. The interpretability of the results due to the non-negativity constraint, the flexibility of the factorization model that allows incorporating multiple factors with different constraints or regularizations, and the convenient multiplicative update rules have made this method very attractive in many applications. However new types of applications are posing challenges to NMF-based methods. In real-time and low-latency scenarios, the high computational cost and its iterative nature are the main problems with NMF techniques. These problems are often dealt with by reducing the number of free parameters of the NMF, using constant basis or initializing to 0 certain gains. However, the update of the free parameters remains iterative and requires a large number of computations.

We propose the use of Tikhonov Regularization as an alternative factorization method for certain specific situations and show its viability in some common scenarios. We target those tasks in which the problem can be stated or approximated in the linear form $\boldsymbol{V} = \boldsymbol{W}\boldsymbol{H} + \boldsymbol{B}$ where $\boldsymbol{V}$ is the magnitude spectra of the mixture, $\boldsymbol{W}$ is a constant set of previously known fixed basis components, $\boldsymbol{H}$ are the gains over time of the basis components and $\boldsymbol{B}$ is random noise. This is often the case in pitch likelihood estimation and in unsupervised or semi-supervised source separation (Virtanen and Klapuri, 2006; Vincent et al., 2008; Heittola et al., 2009; Durrieu et al., 2009b, 2011). Even though most of these methods use more complex spectrum models, they can often be reformulated or approximated using the targeted linear model. In most cases the matrix $\boldsymbol{W}$ is composed of spec-

tral harmonic combs or other similar correlated patterns. This leads to an ill-posed problem where the condition number of $\boldsymbol{W}$ is very high, implying that the solution may be very sensitive to model/data noise or non-unique solutions (Hansen, 2010). In these situations the naive Least Square Means solution $\hat{\boldsymbol{H}} = (\boldsymbol{W}^\top \boldsymbol{W})^{-1}\boldsymbol{W}^\top \boldsymbol{V}$ often gives unsatisfactory results.

Tikhonov regularization (Tikhonov, 1963) is a well-studied method that has seen several incarnations and has received many interpretations over the years (Riley, 1955; Phillips, 1962; Twomey, 1963; Golub, 1965; Foster, 1961). In the field of statistics, Tikhonov regularization is known as ridge regression and has also been widely studied (Hoerl and Kennard, 1970a,b; Marquardt, 1970).

The Tikhonov solution $\hat{\boldsymbol{H}}^{\lambda}$ in its simple form is defined as the solution to:

$$\min_{\boldsymbol{H}} \left\{ |\boldsymbol{W}\boldsymbol{H} - \boldsymbol{V}|^2 + \lambda^2 |\boldsymbol{H}|^2 \right\} \tag{4.1}$$

where the regularization parameter λ is a positive parameter that controls the weighting between the two terms of the objective function.

There are two ways to compute the solution. The most commonly used is by transforming it into a Least Squares Means (LSM) problem, and solving it using common LSM approaches. The other way of computing the solution is by using the so-called normal equations. In this case, we obtain a closed form solution $\hat{\boldsymbol{H}}^{\lambda}$, where the resolution matrix $\hat{\boldsymbol{H}}^{\lambda} = \boldsymbol{R}^{\lambda}\boldsymbol{V}$ is defined in the following way:

$$\boldsymbol{R} = \boldsymbol{W}^t[\boldsymbol{W}\boldsymbol{W}^t + \lambda I_{N_\omega}]^+ \tag{4.2}$$

where $[\boldsymbol{Z}]^+$ denotes the Moore–Penrose pseudoinverse of $\boldsymbol{Z}$.

There exist other forms of Tikhonov regularization that allow including *a priori* knowledge about the solution. These will be explored for certain specific tasks.

From a linear algebra point of view the Tikhonov regularization solution can be expressed as a filtered Singular Value Decomposition (SVD) expansion of the form:

$$\hat{\boldsymbol{H}}^{\lambda} = \sum_{w}^{N_W} \varphi_w^{\lambda} \frac{u_w^\top \boldsymbol{V}}{\sigma_w} v_w \tag{4.3}$$

where σ_w, u_w and v_w are the w^{th} singular value, left and right singular vectors of $\boldsymbol{W}$ respectively. φ_w are the filtering factors, that in the case of

Tikhonov regularization are:

$$\varphi_w^{\lambda} = \frac{\sigma_w^2}{\sigma_w^2+\lambda^2} \approx \begin{cases} 1, & , \sigma_w \gg \lambda \\ \sigma_w^2/\lambda^2, & , \sigma_w \ll \lambda \end{cases} \tag{4.4}$$

This has the effect of gradually filtering out from the contribution to the solution the singular vectors with singular values smaller that λ. It is similar to the effect of applying a Truncated SVD, however the transition between retained and filtered SVD components is smoother.

Calvetti and Somersalo (2008) show that Tikhonov regularization can also be derived from a Bayesian perspective if the gains $\boldsymbol{H}$ and the noise $\boldsymbol{B}$ are considered random variables, independent and Gaussian. The Tikhonov regularization solution is equivalent to the Maximum A Posteriori (MAP) estimation.

Regularization parameter

The regularization parameter λ plays an important role in the Tikhonov regularization method. It controls the tradeoff between the smoothness of the solution and its fit to the data. There has been extensive work in finding ways to automatically estimate the optimal value of regularization. Many methods have been proposed to automatically determine an optimal regularization parameter. The discrepancy principle technique makes use of an estimation of the noise's $\boldsymbol{B}$ standard deviation (Hansen, 1987). The L-curve method exploits the curvature of the plot that relates the two terms of the objective function (Castellanos et al., 2002; Hansen et al., 2007). The Generalized Cross Validation method (GCV) uses statistical tools to find the regularization parameter that maximizes the ability of the model to predict missing data (Wahba, 1990). However, to date there is no method that achieves an efficient, robust and reliable way to compute the optimal regularization parameter.

Computational Cost

One of the main reasons we propose the use of Tikhonov regularization to perform spectral decomposition is that it accepts a closed form solution with a low computational cost in comparison with the well established NMF. It is trivial to show that the computational complexity, in big 'Ω' notation, of both methods is the same. In the scenario that we set, where each spectral

frame is independently decomposed into a fixed set of constant basis components, both algorithms are dominated by a matrix-vector multiplication operation.

In the case of the the Tikhonov regularization solution the single matrix-vector multiplication performed is $\boldsymbol{RV}$. Since $\boldsymbol{R} \in \Re^{N_\omega \times N_W}$ is the previously computed resolution matrix, the complexity of the Tikhonov regularization spectral decomposition method is $\mathrm{O}(N_\omega N_W)$. For the Euclidean distance NMF (presented in Section 3.4) the computational complexity is dominated by three matrix-vector multiplications: $\boldsymbol{WH}$, $\boldsymbol{W}^\top \boldsymbol{V}$ and $\boldsymbol{W}^\top(\boldsymbol{WH})$. Since $\boldsymbol{W} \in \Re^{N_\omega \times N_W}$ the complexity also becomes $\mathrm{O}(N_\omega N_W)$. However the actual computational cost in terms of number of multiplications, divisions and additions is much larger for the case of NMF. The total number of operations in the Tikhonov regularization method is exactly $N_\omega N_W$ since the matrix-vector multiplication is only performed once. For the NMF method the total number of arithmetic operations is $k(3N_\omega N_W + 2N_W)$ where k is the number of NMF update iterations and usually $k \in [1, 60]$. Similar results can be derived for the NMF update rules of the other objective functions.

Therefore using the Tikhonov regularization method translates into a significant reduction in computational cost for the spectral decomposition task $N_\omega N_W << k(3N_\omega N_W + 2N_W)$.

Advantages

In our context, one of the main advantages of the Tikhonov regularization method is the possibility of obtaining a closed form solution. In the closed form solution we see that the resolution matrix can be computed independently from the data. Furthermore the only computation that must be performed on the arrival of the data is a matrix-vector multiplication, for which many optimized implementations are available. Another interesting characteristic of this method is that, unlike other methods such as NMF, Tikhonov regularization results in a unique solution.

Limitations

Tikhonov regularization has certain limitations compared to NMF or other current approaches. The basis matrix $\boldsymbol{W}$ must be fixed and constant, and cannot be estimated from the data. The estimated solution can contain negative values and this may pose interpretability problems. The use of the closed form solution requires the Euclidean distance to be used for fitting the

data, however for certain type of signals this limitation can be solved using pre-whiting. Finally, in the case of wanting to compute the resolution matrix before receiving and independently from the data, the regularization cannot depend on the analyzed data. This last issue can be addressed by previously computing multiple resolution matrices with different regularization values.

Throughout this work we evaluate the use of Tikhonov regularization in pitch estimation and source separation tasks, under low-latency and low computational cost constraints. We explore the use of different basis matrices $\boldsymbol{W}$ depending on the use case and compare the results to other state-of-the-art methods such as NMF.

4.3 Target Instrument Pitch Estimation and Tracking

Predominant pitch estimation and tracking is often the first step in source separation methods that target harmonic sources and specially for removing or isolating singing vocals. In this Section we propose a pitch tracking method that estimates the predominant pitch of a specific instrument. This method of pitch tracking is useful for both low-latency and high-latency source separation scenarios, producing better results than other current methods in both cases.

Harmonic Summation Likelihood Estimation

Harmonic summation (Klapuri, 2006) is a well known and conceptually simple frequency domain technique to estimate the pitch salience of an audio frame. We use this method in this work as the reference for a computationally inexpensive, low-latency pitch likelihood estimation approach.

The salience of each pitch candidate is computed by summing the energy of the spectrum bins contributing to that pitch weighted by the strength of their contribution. The strength of their contribution can be computed in many ways. Klapuri (2006) proposes training the weighting function using a database of samples of individual instruments with annotated pitch.

For a given spectrum frame $\boldsymbol{V}$, the salience is defined as:

$$s[j] \quad = \quad \sum_{r=1}^{N_R} g(j,r) \max_{\omega \in \kappa_{j,r}} |\widetilde{\boldsymbol{V}}[\omega]| \tag{4.5}$$

where N_R is the number of hypothetical harmonics to take into account, $g(j,r)$ is the weighting function, $\widetilde{\boldsymbol{V}}$ is the whitened and noise-suppressed spectrum, and $\kappa_{j,r}$ defines the frequency neighborhood of a partial position given the pitch candidate frequency j and the harmonic index r. It can be defined as:

$$\kappa_{j,r} = [\langle rN_\omega/(j+\Delta j/2)\rangle, \ldots, \langle rN_\omega/(j-\Delta j/2)\rangle] \tag{4.6}$$

$\langle\cdot\rangle$ denotes rounding to the nearest integer. Δj controls the width of the harmonic neighborhood when computing the salience, in order to take into account widening of the partials and inharmonicity of the pitch.

Klapuri (2006) proposes the following weighting function:

$$g(j,r) = \frac{f_s/j+\alpha}{rf_s/j+\beta} \tag{4.7}$$

where f_s is the sample rate of the analyzed signal, and the parameters $\alpha = 27$Hz and $\beta = 320$Hz are chosen by training the model on existing harmonic envelopes.

Finally the pitch likelihood that we use in the following sections is a normalization of the salience function $\boldsymbol{L}[j] = s[j]/\sum s[j]$.

The harmonic summation method is useful for predominant pitch estimation tasks, however it presents problems in multiple pitch estimation due to the independent estimation of the candidates and does not provide a generative model of the spectrum which is especially interesting in spectral decomposition and source separation scenarios.

Tikhonov Regularization Likelihood Estimation

The pitch likelihood estimation method that we propose is a linear signal decomposition model. Similar to NMF, this method allows us to perform a joint pitch likelihood estimation. The main strengths of the presented method are low latency, implementation simplicity and robustness in multiple pitch scenarios with overlapping partials. This technique performed better than a simple harmonic summation method in our preliminary tests.

The main assumption is that the spectrum $\boldsymbol{V}_t \in \mathbb{R}^{N_\omega \times 1}$ at a given frame t, is a linear combination of N_W elementary spectra, also named basis components. This can be expressed as $\boldsymbol{V}_t = \boldsymbol{W}\boldsymbol{H}_t$, N_ω being the size of the spectrum. $\boldsymbol{W} \in \mathbb{R}^{N_\omega \times N_W}$ is the basis matrix, whose columns are the basis components. $\boldsymbol{H}_t \in \mathbb{R}^{N_W \times 1}$ is a vector of component gains for frame t.

We set the spectra components as filter combs in the following way:

$$\begin{aligned} \varphi[m,n] &= 2\pi f_l H N_P \frac{2^{\frac{iH-F/2+n}{HN_P}}-1}{S_r \ln(2)} \\ \boldsymbol{W}_m[k] &= \sum_{n=0}^{F} w_a[n] \left(\sum_{h=1}^{N_h} sin\left(h\varphi[m,n]\right) \right) e^{-j2\pi nk/N_\omega} \end{aligned} \tag{4.8}$$

with $H = (1-\alpha)F$. Where α is a coefficient to control the frequency overlap between the components, F is the frame size, S_r the sample rate, $w_a[n]$ is the analysis window, N_h is the number of harmonics of our components, $\boldsymbol{W}_m$ is the spectrum of size N_ω of the component of m^{th} pitch. Flat harmonic combs have been used in order to estimate the pitch likelihoods of different types of sources.

The condition number of the basis matrix $\boldsymbol{W}$ defined in Equation 4.8 is very high ($\kappa(\boldsymbol{W}) \approx 3.3 \cdot 10^{16}$), possibly due to the harmonic structure and correlation between the components in our basis matrix. For this ill-posed problem we propose using the well-known Tikhonov regularization method to find an estimate of the components gains vector $\hat{\boldsymbol{H}}_t$ given the spectrum $\boldsymbol{V}_t$. This consists in the minimization of the following objective function:

$$\Phi(\boldsymbol{H}_t) = |\boldsymbol{W}\boldsymbol{H}_t - \boldsymbol{V}_t|^2 + \lambda |\boldsymbol{H}_t|^2 \tag{4.9}$$

where λ is a positive scalar parameter that controls the effect of the regularization on the solution. Under the assumption of Gaussian errors, the problem has the closed-form solution $\hat{\boldsymbol{H}}_t = \boldsymbol{R}\boldsymbol{V}_t$ where $\boldsymbol{R}$ is defined as:

$$\boldsymbol{R} = \boldsymbol{W}^t[\boldsymbol{W}\boldsymbol{W}^t + \lambda I_{N_\omega}]^+ \tag{4.10}$$

and $[\boldsymbol{Z}]^+$ denotes the Moore–Penrose pseudoinverse of $\boldsymbol{Z}$. The calculation of $\boldsymbol{R}$ is computationally costly, however $\boldsymbol{R}$ only depends on $\boldsymbol{W}$, which is defined by the parameters of the analysis process, therefore the only operation that is performed at each frame is $\hat{\boldsymbol{H}}_t = \boldsymbol{R}\boldsymbol{V}_t$.

We must note that in contrast to NMF, our gains $\hat{\boldsymbol{H}}_t$ can take negative values. In order to have a proper likelihood we define the pitch likelihood as:

$$\boldsymbol{L}_t = [\hat{\boldsymbol{H}}_t]_+ / sum([\hat{\boldsymbol{H}}_t]_+) \tag{4.11}$$

where $[\hat{\boldsymbol{H}}_t]_+$ denotes the operation of setting to 0 all the negative values of the vector $\hat{\boldsymbol{H}}_t$.

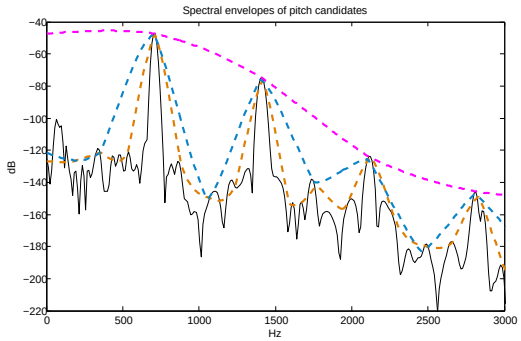

Figure 4.1: Spectrum magnitude (solid black line) and the harmonic spectral envelopes (colored dashed lines) of three pitch candidates.

Timbre Classification

Estimating the pitch track of the target instrument requires determining when the instrument is not active or not producing a harmonic signal (e.g. in fricative phonemes).

We select a limited number of pitch candidates N_J by finding the largest local maxima of the pitch likelihood function $\boldsymbol{L}_t$ 4.11. For each candidate a feature vector c is calculated from its harmonic spectral envelope $e_h(\omega)$ and a classification algorithm predicts the probability of it being a *voiced* envelope of the target instrument. The feature vector c of each of the candidates is classified using Support Vector Machines (SVM). The envelope computation $e_h(\omega)$ comes from the Akima interpolation (Akima, 1970) between the magnitude at harmonic frequencies bins. The timbre features c are a variant of the Mel-Frequency Cepstrum Coefficients (MFCC), where the input spectrum is replaced by an interpolated harmonic spectral envelope $e_h(\omega)$. This way the spectrum values between the harmonics, where the target instrument is often not predominant, have no influence on the classification task. Figure 4.1 shows an example of a spectrum $\boldsymbol{V}_t[\omega]$ (in black) of a singing voice signal, and the interpolated harmonic spectral envelopes $e_{h,1}(\omega)$, $e_{h,2}(\omega)$ and $e_{h,3}(\omega)$ (in magenta, blue and orange respectively), of three different pitch candidates.

The features vector c contains the first 13 coefficients of the Discrete Cosine Transform (DCT), which are computed from the interpolated envelope

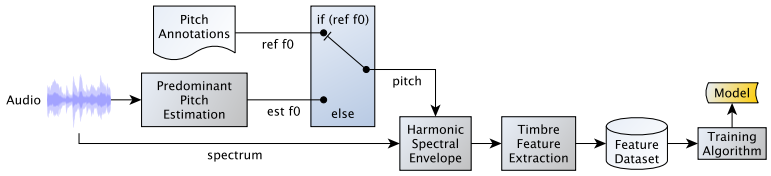

Figure 4.2: In the training stage, the $e_h(\omega)$ is based on the annotated pitch if it exists *if (ref. f0)*, and on the estimated pitch otherwise.

$e_h(\omega)$ as:

$$c = DCT\left(10 \cdot \log\left(E[l]\right)\right) \tag{4.12}$$

where $E[l] = \sum_{f_{l,low}}^{f_{l,high}} e_h(\omega)^2$, and $f_{l,low}$ and $f_{l,high}$ are the low and high frequencies of the l^{th} band in the Mel scale. We consider 25 Mel bands in a range $[0...5kHz]$. Given an audio signal sampled at $44.1kHz$, we use a window size of 4096 and a hop size of 512 samples. The workflow of our supervised training method is shown in Figure 4.2. Two classes are defined: *voiced* and *unvoiced* in a frame-based process [1]. *Voiced* frames contain pitched frames from monophonic singing voice recordings (i.e. only a vocal source). Pitched frames have been manually annotated. In order to generalize well to real audio mixtures, we also include audio examples composed of an annotated vocal track mixed artificially with background music. *Unvoiced* frames come from three different sources: *a)* non-pitched frames from monophonic singing voice recordings (e.g. fricatives, plosives, aspirations, silences, etc.); *b)* other monophonic instrument recordings (sax, violin, bass, drums); and *c)* polyphonic instrumental recordings not containing vocals. We employ a radial basis function (RBF) kernel for the SVM algorithm (Chang and Lin, 2001). As a pre-process step, we apply standardization to the dataset by subtracting the mean and dividing by the standard deviation. We also perform a random subsampling to reduce model complexity. We obtain an accuracy of 83.54%, when evaluating the model against the test dataset.

1. The original training and test datasets consist of $384,152$ $(160,779/223,373)$ and $100,047$ $(46,779/53,268)$ instances respectively. Sub-sampled datasets contain $50,000$ and $10,000$ respectively. Values in brackets are given for the voiced and unvoiced instances respectively.

Instrument Pitch Tracking

The timbre classification described in the previous section is integrated into the instrument pitch tracking step. We describe first an offline dynamic programming algorithm for estimating the sequence of fundamental frequencies f_0 belonging to the singing voice. The algorithm is divided into two steps, as shown in Figure 4.3.

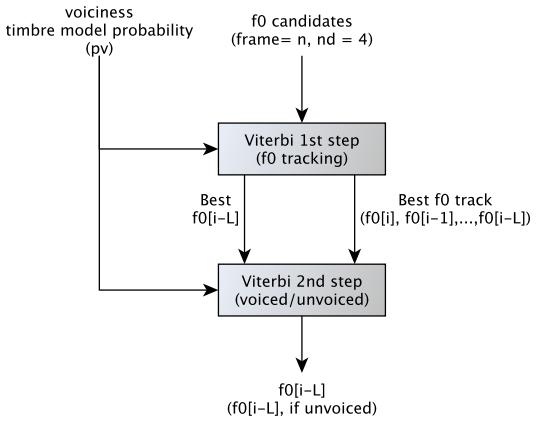

Figure 4.3: Block diagram of the predominant pitch estimation.

The output of the algorithm is the estimated predominant pitch of frame i, with a fixed latency of L frames. If the frame is unvoiced, i.e. no predominant pitch found, the estimated predominant pitch is set to 0.

The first step is selecting the best sequence of f_0 candidates. The number of candidates N_J per frame is an arbitrary value. For each frame, the N_J largest peaks of the f_0 likelihood function within the target frequency band are selected as f_0 candidates. We used four candidates per frame in our experiments ($N_J = 4$). For each node in the matrix, the probabilities are computed according to the following criteria [2]:

f_0-estimation likelihood : the higher the f_0-estimation likelihood value, the higher the probability. For each frame, the f_0-likelihood values of the candidates are normalized with respect to the highest one, and

2. Gaussian function parameters μ and σ are set empirically for each probability.

the probability is computed as:

$$\mathcal{P}_{1,f_0} = e^{-\frac{(x-\mu_{f_0})^2}{2\sigma_{f_0}^2}} \tag{4.13}$$

where x is the normalized f_0-likelihood value, $\mu_{f_0} = 1$, and $\sigma_{f_0} = 0.4$.

voiciness : the higher the voiciness value predicted by SVM, the higher the probability. Voiciness values go from 0=unvoiced to 1=voiced. This offers a more controlled mapping function on the SVM output prediction. The probability is computed as:

$$\mathcal{P}_{1,v} = e^{-\frac{(p_v-\mu_v)^2}{2\sigma_v^2}} \tag{4.14}$$

where p_v is the voiciness value, $\mu_v = 1$, and $\sigma_v = 1$.

frequency continuity : the shorter the frequency distance in semitones, the higher the probability. The probability is computed as:

$$\mathcal{P}_{1,f} = e^{-\frac{(x-\mu_f)^2}{2\sigma_f^2}} \tag{4.15}$$

where $x = min(6, max(0, |\Delta f| - 0.5))$, Δf is the frequency difference in semitones, $\mu_f = 0$, and $\sigma_f = 4$.

Probabilities $\mathcal{P}_{1,f_0}$ and $\mathcal{P}_{1,v}$ are combined into the state probability $\mathcal{P}_{1,S} = \mathcal{P}_{1,f_0}\mathcal{P}_{1,v}$, and $\mathcal{P}_{1,f}$ is used as transition probability $\mathcal{P}_{1,T} = \mathcal{P}_{1,f}$. Note that one fundamental frequency value is found per frame, therefore there are no unvoiced segments in the estimated path.

The second step is selecting the best path in a matrix with two rows (states) and several columns corresponding to the frames. The first row corresponds to the voiced state (e.g. the estimated f_0 sequence found in the first step of the algorithm), and the second row corresponds to the unvoiced state. In this case, node probability is computed from the voiciness measure as

$$\mathcal{P}_{2,v} = \begin{cases} e^{-\frac{(0.5-\mu_v)^2}{2\sigma_v^2}} & \text{for the } \textit{unvoiced} \text{ row,} \\ e^{-\frac{(p_v-\mu_v)^2}{2\sigma_v^2}} & \text{for the } \textit{voiced} \text{ row.} \end{cases}$$

where p_v is the voiciness value, $\mu_v = 1$, and $\sigma_v = 1$. Node jump probability is computed as:

$$\mathcal{P}_{2,f} = \begin{cases} e^{-\frac{(x-\mu_f)^2}{2\sigma_f^2}} & \text{for the } \textit{voiced-unvoiced} \text{ jump,} \\ 1 & \text{otherwise.} \end{cases}$$

where $x = min(6, max(0, |\Delta f| - 0.5))$, Δf is the frequency difference in semitones, $\mu_f = 0$, and $\sigma_f = 4$. Note that jumping from voiced to unvoiced or vice versa is not penalized, thus irregular f_0 segments or low-voiciness segments will become unvoiced segments.

We have adapted the dynamic programming described to work with a fixed latency that can be manually set. The adaptation consists in updating the matrices at each new frame and performing an incremental forward pass and a normal backtracking pass. At each new frame, a new column is added to the matrix of the first step, we calculate transition and state costs only for the new nodes and the first column of the matrix is then removed. Then a partial backtracking is performed for a number of columns equal to the latency parameter. The same process is performed on the the matrix of the second step. It is updated with the found f_0 trajectory, and the node and jump probabilities are computed as previously reported. A partial backtracking is performed in this second matrix, and we choose the best node in the column corresponding to the frame at the desired latency.

Evaluation

First, we compute several measures for the estimated lead vocals pitch, including a voiced/unvoiced frame classification when the lead voice is not present.

With regard to the evaluation measures, we used those of the MIREX evaluation campaign (Downie et al., 2005): *Voiced recall*, *Voiced precision*, *Voiced false alarm* and *Overall accuracy*. For the pitch frequency estimation *Raw pitch* is the percentage of *voiced* frames in the reference with a difference below 50 semitone cents (1/4 tone) between estimated and reference pitch. *Raw pitch chroma* does not consider octave errors

From the MIREX'2005 evaluation campaign's training dataset (MIREX, 2005), we took nine examples containing vocals as lead instrument by female and male singers. The results are comparable to the best participants in

Latency (frames)	1	3	10	20	100
voiced-recall	100.00	92.50	93.84	**95.01**	93.80
voiced-precision	65.41	72.16	71.63	70.19	71.66
voiced-false-alarm	100.00	67.47	70.27	76.30	70.13
overall-accuracy	52.99	62.37	62.35	59.25	62.43
raw-pitch	81.01	78.15	79.60	78.06	79.64
raw-pitch-chroma	83.28	80.24	81.51	79.62	81.56

Table 4.1: Latency influence on the pitch accuracy for the LLIS-SVM method. Latency is expressed in number of frames (frame time is 11.6 ms).

the MIREX'2005 evaluation campaign for the task of melody extraction [3], which obtain a voiced recall of 81.8% and an overall accuracy of 71.4%, with a voiced false alarm rate of 17.3% (MIREX, 2005; Dressler, 2005). Our approach has a much higher false alarm rate (76%), although in the use-case of lead vocals removal, obtaining a good voiced recall is more important.

Table 4.1 shows the effect of latency in the pitch tracking step. Latency is expressed in number of frames (frame duration is 11.6 ms) and corresponds to the size of the Viterbi backtracking as detailed in section 4.3. For lead instrument removal, we need a high value for voice recall since false negatives (FN) will result in by-passing the lead instrument. Therefore we choose a latency of 20 frames for the pitch tracking step, taking the highest voiced recall values while maintaining good voice precision.

Table 4.2 contains the results on pitch accuracy, comparing our algorithm with an implementation of the Instantaneous Mixture Model (IMM) approach by Durrieu et al. (2010) which is an NMF-based approach presented in Section 3.4. The implementation code is available online [4].

Additionally, we can observe the contribution of the voice timbre classification of section 4.3. The LLIS-noSVM method forces the algorithm to output an estimated predominant pitch (i.e. all frames are considered as voiced). The accuracy results, raw pitch and raw chroma, are computed taking into account only frames that are voiced in the reference pitch files. By adding the timbre classifier (LLIS-SVM), we accomplish an improvement of 5% in voiced precision v_p, with a decrease of 3% in raw pitch accuracy.

3. Although these results were computed with a test dataset from the MIREX collection that is not publicly available, the results should be comparable to some extent.

4. `http://durrieu.ch/phd/software.html` (last accessed on January 3, 2011)

Method	IMM	LLIS-noSVM	LLIS-SVM	HarmSum-SVM
voiced-recall	-	-	95.01	92.90
voiced-precision	65.44	65.41	70.19	70.34
voiced-false-alm	-	-	76.30	74.08
overall-accuracy	-	-	59.26	61.39
raw-pitch	71.77	81.06	78.06	80.14
raw-pitch-chroma	76.19	84.07	79.62	81.50

Table 4.2: Pitch accuracy evaluation. Note that the some measures are not applicable since the algorithm does not provide voiciness detection.

However the proposed method is not yet capable of achieving the results obtained by the Harmonic Summation approach combined with the timbre model (HarmSum-SVM). Overall, the LLIS-SVM method seems to offer an adequate raw pitch detection but still needs improvement in the detection of false positives.

4.4 Extension to Multiple Pitch Estimation and Tracking

Given the proposed predominant pitch estimation and tracking method, and motivated by addressing several shortcomings of it, we developed an extension capable of tracking multiple pitches under a controllable latency constraint. This work shows another use of the Tikhonov spectrum decomposition method that we propose in Section 4.2, and can serve as a base for further studies.

This section demonstrates the use of Tikhonov regularization in multiple fundamental frequency likelihood estimation, and how it would fit into a full latency-controlled multiple pitch tracking system.

In Chapter 6 Section 6.6 we show that the estimation of multiple pitches present in the mixture signal can improve even more the separation quality.

Introduction

This method addresses two main problems with the technique presented in Section 4.3. The previously proposed method fails with regions where the pitch is highly modulated in frequency. Singing voice often presents these

types of regions, notably glissandos and vibratos. Another shortcoming with the proposed predominant pitch tracking method comes from using the results of timbre modeling and classification in the Hidden-Markov Model of the tracking step. The classifier does not always output correct results and this leads to breaks in the pitch contours. This problem is very noticeable in source separation tasks.

The method developed in this section lessens these two problems, while at the same time extending the method to estimate multiple pitches simultaneously. This is done by splitting the system into three stages: pitch likelihood estimation, pitch tracking and pitch contour selection (see Figure 4.4).

The pitch likelihood estimation stage consists in finding the presence probability of all possible pitches. This stage is done in the same way as in the predominant pitch method (see Section 4.3), and it is not developed any further here.

The pitch tracking step will not use information from the timbre. Only pitch likelihood and transition information is needed to create the pitch contours independently of the timbre. The timbre information could be very useful when computing pitch transition probabilities. It could be done by creating a measure of timbre similarity between consecutive frames, however this falls out of the scope of this study.

The pitch contour selection is based on the features of a short history of frames of the contours. These features extract information from the contour rather than from the instantaneous spectrum frame. In this work we propose a set of simple features based on the timbre classification information and on the probability of there being an octave error. In the future, other features such as timbre evolution over time or frequency statistics (Salamon and Gomez, 2012) could be extracted and could increase the flexibility and accuracy of the selection.

Pitch Tracking

The pitch tracking stage in a multipitch scenario consists in assigning certain frequencies of the pitch likelihood function to a set of pitch tracks in time. For a single pitch, this task is often modeled as a Hidden-Markov Model as done in Section 4.3. In the case of single pitch we have a HMM with as many states as possible pitches, and in order to reduce complexity we often reduce the candidates to a selection of peaks in the pitch likelihood function. However for multiple pitches the number of states increases significantly,

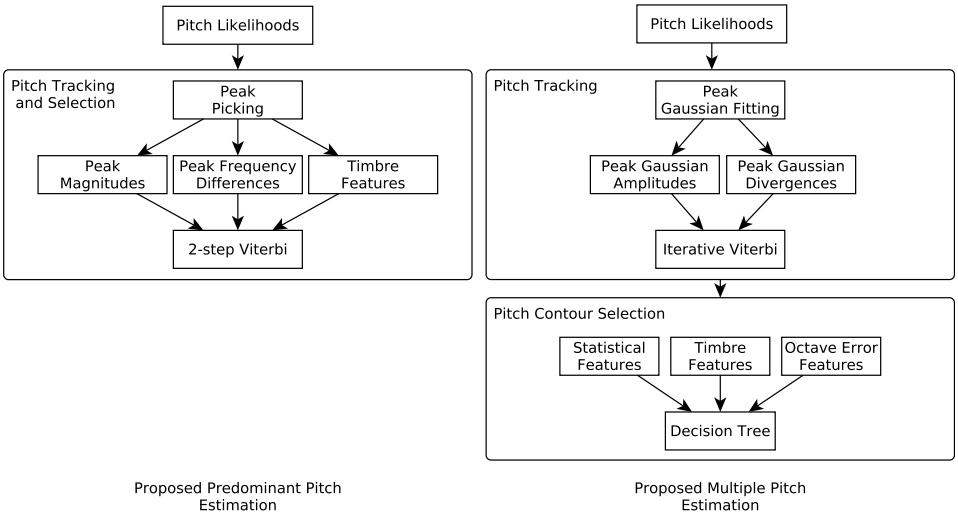

Figure 4.4: Block diagram comparing the predominant pitch estimation method (from Section 4.3) and the multipitch estimation extension.

since there are many combinations of paths and possible pitches. In order to reduce the complexity we perform iteratively a single pitch tracking and removal process. This iterative method is especially disadvantageous in situations where pitch contours cross, however it serves as proof of concept here.

In this work we mainly focus on the use of the pitch likelihood function resulting from decomposition methods, such as Tikhonov regularization and NMF, in multipitch tracking without going in depth into the details of the tracking itself.

With relation to the pitch likelihood function, there is an important difference between harmonic summation methods and decomposition methods. In the former the computation of the likelihoods is performed independently for each pitch, while in the latter it is performed jointly. Furthermore decomposition methods are generative models in the sense that they can reconstruct the spectrum given the pitch likelihood or an intermediate representation of it. In decomposition methods the likelihood function not only depends on the position and energy of the partials but also on their shape.

This property becomes apparent when dealing with vibratos and glissandos,

as can be seen in Figure 4.5. In these cases the partials of the spectrum become wider and of lower magnitude. This happens because the energy of the signal in a given analysis window is not concentrated on one single specific frequency, it is distributed in a region of frequencies. In decomposition methods, this translates to a spreading of the peaks in the pitch likelihood function, which become wider and of lower magnitude when the pitch is frequency modulated. This property may also affect some harmonic summation methods, if these take the shape of the partials into account. However most often these techniques sum all the energy of the partials, independently of their shape, to the likelihood of the corresponding pitch. While this property could be seen as a problem, it can be used in the tracking stage as a cue to ensure continuity of the tracks.

The first proposal to address this issue is to use the energy of the pitch likelihood peaks as the probability of belonging to a pitch contour. However since the peaks in the pitch likelihood function may partially overlap, it is not straightforward to compute the energy from the pitch likelihood function.

In order to compute the energy of a given pitch likelihood peak, we first fit the pitch likelihood function with a set of Gaussian functions (see Figure 4.5). Lets assume w_j are the positions of the N_J largest peaks in the pitch likelihood function. The Gaussian fitting is performed by using a mixture of Gaussian distributions to model the likelihood function around these peaks. We minimize to the objective function:

$$\Phi(\{a_j, \sigma_j, \mu_j : \forall j\}) \quad = \quad \sum_{w \in N_{W\,sel}} \left\| \boldsymbol{L}(w) - \sum_{j=1}^{N_J} a_j \mathcal{G}(w; \sigma_j, \mu_j) \right\|^2 \quad (4.16)$$

where $\mathcal{G}(x; \sigma, \mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ is the Gaussian distribution. $N_{W\,sel}$ is a subset of the bins of the pitch likelihood function in order to only take into account in the fitting the regions around the selected candidate pitches $N_{W\,sel} = \{w : |w - w_j| < d, \quad \forall w, j\}$. Where d is a threshold distance around the peak positions. The peak positions w_j and magnitudes $\boldsymbol{L}(w_j)$ are used for the initial values of the μ_j and a_j respectively, during the optimization of Equation 4.4.

The parameters resulting from the minimization of Φ are then used as the characteristics of the pitch candidates. a_j represents the energy of the pitch candidate and can be related to the likelihood or predominance in

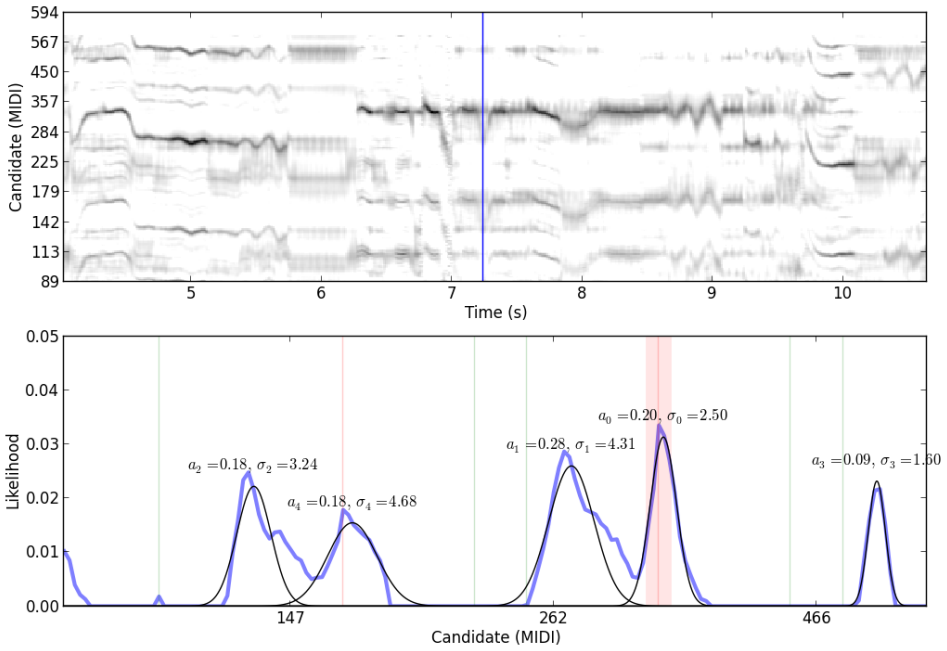

Figure 4.5: Example of the Gaussian fitting procedure to model the pitch candidates on a mixture with singing voice and acoustic guitar. Pitch likelihood (often referred to as pitchgram or chromogram) over time (*top*). Pitch likelihood slice corresponding to the vertical line on the top plot (*bottom*). The thick curve is the pitch likelihood of the given frame. The fitted Gaussians are plotted as thin curves. The red thick vertical line corresponds to the global maximum of the likelihood function. Note that the Gaussian of the pitch likelihood peak of the singing voice vibrato has a larger amplitude than that of the peak corresponding to the less predominant acoustic guitar ($a_1 > a_0$).

the current time frame. μ_j represents the frequency of the pitch. Finally the frequency modulation rate of the pitch track affects the pitch likelihood peak width and therefore the parameter σ_j.

The parameters can then be used to perform the tracking in a way similar to that shown in Section 4.3. We create an HMM with $t \in [0, N_T]$ time indices and $j \in [0, N_J]$ states for each time index. The extra state $j = 0$ corresponds to a track set to non-pitched, the other states correspond to the peaks of the likelihood function.

We may use a_j and μ_j in the computation of the state probability $\mathcal{P}_{1,S}$ and

the transition probability $\mathcal{P}_{1,T}$ respectively:

$$\mathcal{P}_{1,S}(j_t) = \begin{cases} e^{-\frac{(\mu_{j_t}-\mu_S)^2}{2\sigma_S^2}} & \text{for } j_t > 0, \\ \mathcal{P}^u_{1,S} & \text{for } j_t = 0. \end{cases}$$

$$\mathcal{P}_{1,T}(j_t, j_{t-1}) = \begin{cases} 0 & \text{if } j_{t-1} = 0, j_t = 0 \\ e^{-\frac{(x-\mu_T)^2}{2\sigma_T^2}} & \text{if } j_{t-1} \neq 0, j_t \neq 0 \\ \mathcal{P}^{p-u}_{1,T} & \text{else} \end{cases}$$

where $\mu_S = 0.15$, $\sigma_S = 0.1$, $x = min(6, max(0, |\Delta\mu^s| - 0.5))$, $\Delta\mu^s$ is the difference between μ_j in semitones, $\mu_T = 1$ and $\sigma_T = 1$. $\mathcal{P}^u_{1,S}$ is the probability of a pitch track being unpitched and $\mathcal{P}^{p-u}_{1,T}$ is the transition probability between a pitched and unpitched state of the track. Note that in this extension to multipitch tracking the voiciness probability $\mathcal{P}_{1,v}$ is not used.

However we propose a different method to compute the transition probability based on the divergence between two Gaussians. We take into account not only the difference between frequencies of consecutive frames, but also the differences in energies and frequency modulation rates. As in Zouari and Chollet (2006) we have chosen the symmetric Kullback-Leibler as a divergence measure between weighted Gaussians. In our case the Gaussians are 1-D resulting in:

$$\begin{aligned} kl(a_1, \sigma_1, \mu_1; a_2, \sigma_2, \mu_2) &= \frac{1}{2}\left(a_1\frac{\sigma_1}{\sigma_2} + a_2\frac{\sigma_2}{\sigma_1}\right) \\ &\quad + \frac{1}{2}(\mu_1 - \mu_2)^2\left(\frac{a_1}{\sigma_1} + \frac{a_2}{\sigma_2}\right) \\ &\quad - (a_1 + a_2) \end{aligned} \tag{4.17}$$

In this case the transition probability between two consecutive pitch candidates is defined as:

$$\begin{aligned} gd(j_t, j_{t-1}) &= \log\left(\left|kl(a_{j_t}, \sigma_{j_t}, \mu_{j_t}; a_{j_{t-1}}, \sigma_{j_{t-1}}, \mu_{j_{t-1}})\right|\right) \\ \mathcal{P}_{1,T}(j_t, j_{t-1}) &= e^{-\frac{(gd(j_t, j_{t-1}) - \mu_{T_{gd}})^2}{2\sigma^2_{T_{gd}}}}, \text{if } j_t > 0, j_{t-1} > 0 \end{aligned} \tag{4.18}$$

where $\mu_{T_{gd}} = 0$, and $\sigma_{T_{gd}} = 0.15$.

Finding the best path in the HMM is performed using a Viterbi algorithm as in Section 4.3. However in this case we want to be able to find the first

N_{cont} best paths. Therefore we apply the Viterbi algorithm iteratively on multiple copies of the HMM using the following algorithm:

$n = 1$
Compute state costs of HMM$[n], \forall n$
while $n < N_{cont}$ **do**
  Compute Viterbi's forward step of HMM$[n]$
  Estimate best path $\vec{p} = [j_0, j_1 ... j_{N_T}]$ using Viterbi's backward step on HMM$[n]$
  Set to 0 state probability $\mathcal{P}_{1,S}(\vec{p}[t]) = 0, \forall t$ of HMM$[n], \forall n$
  $n = n + 1$
end while

In order to limit the latency of the system the Viterbi is only performed on a limited history of the signal in the same manner as in Section 4.3. The forward step, computing the state costs and accumulated costs, can be done one frame at a time since it only depends on the past. However the best path resulting from the backward step might change with the arrival of each new frame. If the best path changes, it may cause jumps in the resulting pitch contours. This may be attenuated by raising the history and therefore the latency of the system. However this effect can also be reduced by raising the number of HMM (N_{cont}), such that the state probability of the candidates that could lead to best path changes are set to 0, reducing the chances of obtaining a different path for that HMM.

Pitch Contour Selection

The pitch contour selection consists in choosing the tracks resulting from the previous step that are of interest for a given application. The basic idea is to use a limited history of the pitch contours, compute some features on them and select the ones that better match a given criterion. In this way we can capture mid-term and long-term characteristics that are not obtained using the HMM.

Salamon and Gomez (2012) compute first and second order pitch contour statistics to select melody lines in a mixture. In Marxer et al. (2011) we compute timbre classification features that are integrated in the HMM to select the pitch contours of a specific instrument. Here we propose three features that are useful for some general use cases such as selecting a specific instrument and/or avoiding common errors in pitch trackers.

Statistical Features

The statistical features of the contours are useful for most tasks. The contours are time series of two dimensions: frequency and likelihood. The likelihood is often related to the intensity, salience and/or loudness of the pitch. There are many useful statistical features that can be extracted from the pitch contours. Salamon and Gomez (2012) propose a set of characteristics that can be used for melody selection. Among the features proposed we find the means and standard deviations of the salience and frequency. They also propose other global characteristics including the sum of the salience, the length of the contour or the presence of vibrato.

In this work we focus on systems with a controllable latency and therefore we limit ourselves to features that can be computed incrementally. Of the characteristics proposed by Salamon and Gomez (2012), the means and standard deviations of the salience and frequency allow for incremental estimates to be performed. The sums of the salience and duration require the full pitch contour to be calculated making incremental estimates impossible.

Timbre Features

The timbre feature is based on the work presented in Section 4.3. This feature consists of the incremental mean of the result from the SVM classifier applied to the c feature vector from Equation 4.12. In contrast to the method proposed in Marxer et al. (2011), here the timbre information is used as a feature of the pitch contour instead of as a feature of each instantaneous frame. This feature is useful for selecting specific instruments for which we have trained the timbre model.

Octave Error Features

Octave errors are some of the most common problems in multiple fundamental frequency estimation methods. This is a consequence of the octave ambiguity: a pitched sound of fundamental frequency $2f_0$ can be spectrally very similar to a pitched sound of fundamental frequency f_0 if all the odd partials (1st, 3rd, 5th,...) are set to 0 by the filter (the timbre of the sound) or if they are masked by other sounds present in the mixture. This case is a limit situation and quite rare, however intermediate scenarios where pitches and their octaves occur simultaneously or the odd partials are simply lowered are not uncommon, and the same ambiguity applies.

We propose a feature for finding the pitch based on its timbre to alleviate octave errors. As in Section 4.3 we characterize the timbre of a pitch by using the Harmonic Spectral Envelope (HSE) $e_h(f)$. For the development of this feature we propose several measures for a given fundamental frequency f based on the divergence between its HSE ($hse_r(f) = e_h(f)$) and the HSE of its relative lower octave $hse_l(f) = e_h(f/2)$ and higher octave $hse_h(f) = e_h(2f)$. These measures are defined as:

$$\begin{aligned} dhse_l(f) &= D_{SKL}(hse_l||hse_r) \\ dhse_h(f) &= D_{SKL}(hse_h||hse_r) \\ dhse_{hl}(f) &= D_{SKL}(hse_h||hse_l) \end{aligned} \tag{4.19}$$

where $D_{sKL}(x||y) = D_{KL}(x||y) + D_{KL}(y||x)$ is the symmetrised Kullback-Leibler divergence between the distributions x and y, with $D_{KL} = x \log(x/y)$ The main hypothesis we propose is that for the correct pitch the ratio between $dhse_l(f_0)$ and $dhse_h(f_0)$ is larger than for the higher and lower octaves. In other words assuming f_0 is the correct fundamental frequency of a given pitch:

$$\frac{dhse_l(f_0)}{dhse_h(f_0)} > \frac{dhse_l(2f_0)}{dhse_h(2f_0)} > \frac{dhse_l(f_0/2)}{dhse_h(f_0/2)} \tag{4.20}$$

The motivation to propose this hypothesis comes from two main assumptions. The first assumption is that the HSE of the true fundamental frequency can be modeled as a spectrally smooth filter. This assumption has been widely used in the literature (Klapuri, 2003; Every, 2006; Durrieu et al., 2010; Benetos and Dixon, 2011a). The second assumption is that the magnitude of the spectrum between the true pitch partials is significantly lower than at the partial positions. In other words the magnitude of the background spectrum is low with respect to the HSE of the true fundamental frequency.

The theory behind this method can be explained in terms of downsampling and upsampling an imaginary envelope. If we consider any octave compared to the true pitch, going down or up an octave is equivalent to upsampling and downsampling the HSE respectively. Since the HSE is assumed smooth (it has low energy in the high frequency components) the aliasing effect will be small and the resulting interpolated envelope will be similar. Therefore the ratio $dhse_l/dhse_h$ should be close to one. Let's consider a different situation where we are computing the feature for the frequency an octave lower than the true pitch. The HSE of the given fundamental frequency

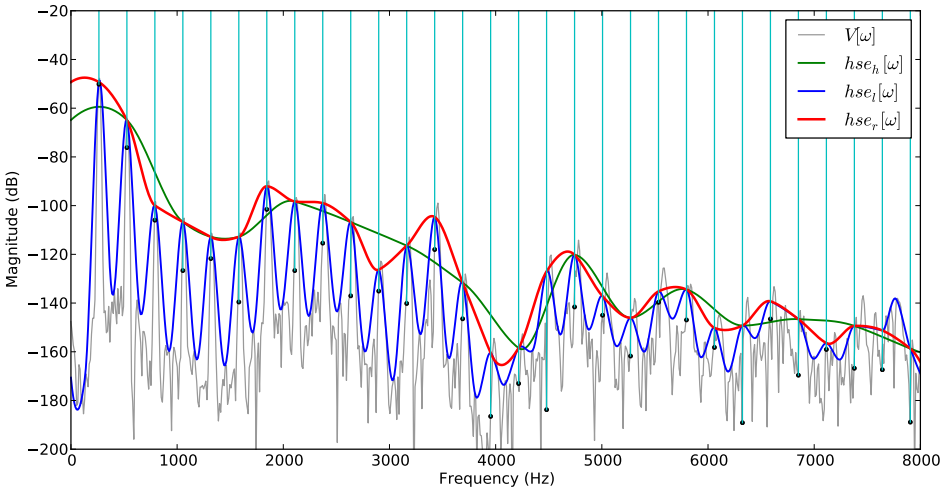

Figure 4.6: Harmonic envelopes for the reference, lower octave and higher octave pitches for a singing voice in isolation. Notice the harmonic envelope of the octave lower is significantly different from the envelope of the correct pitch and the octave higher.

is not smooth, because it corresponds to the multiplex of the true smooth HSE with a low magnitude background spectrum. If we go up an octave we will be computing the HSE of the true pitch and therefore a sampling of a smooth spectrum. However if we compute the HSE of a lower octave we will again end up with a multiplex of the background spectrum and the true pitch HSE. This leads to a ratio of $dhse_l/dhse_h$ lower than one. Finally, if we compute the feature for the true pitch the result is a ratio of $dhse_l/dhse_h$ larger than one.

Taking into account normalization terms in order to maintain the value of the feature in a specific range, the octave error feature is defined as:

$$oerr(f) \quad = \quad \log\left(\frac{dhse_l(f)}{dhse_h(f)}\right) \tag{4.21}$$

Evaluation

To evaluate the octave error feature we use monophonic audio excerpts for which the pitch is previously estimated. The estimated pitch is considered as reference f_{ref}. The excerpts are mixed together at different polyphonies.

The octave error feature is computed at every frame on the mixtures for the pitches $f_{low} = f_{ref}/2$, $f_{high} = 2f_{ref}$ and f_{ref} of each of the sources present.

We consider two different multi-track datasets: *wind* (a wind instruments quintet) and *choir* (a vocal quartet).

The first consists of the wind instruments database for the Multiple Fundamental Estimation task of the Third Music Information Retrieval Evaluation Exchange (MIREX2007). This dataset is composed of a woodwind quintet recording of Beethoven's Variations for String Quartet Op.18 No. 5. Each instrument (flute, oboe, clarinet, horn, or bassoon) was recorded separately while the performer listened to the other parts (recorded previously) through headphones. The mixtures are generated by mixing the recordings of the individual instruments, with polyphonies ranging from 2 to 5. This combinatorial process results in a total of 26 individual mixtures.

The second dataset consists of recordings of four voices (bass, tenor, alto and soprano) of the choir composition "Water Night", composed by Eric Whitacre. Isolated solo recordings were downloaded from the Virtual Choir site [5]. The goal is to observe the effect of having sources with similar timbre (singing voice) and with pitch contours having overlapping harmonic partials. With polyphonies ranging from 2 to 4, the dataset has a total of 11 mixtures.

The pitch annotation of each track is carried out automatically using a monophonic pitch estimation method (de Cheveigné and Kawahara, 2002) on the individual recordings. The same pitch range ($30-1800$ Hz) and voiciness threshold are used to process all recordings. Pitch data is computed using a frame rate of 86 fps.

In order to test the hypothesis of Equation 4.4 presented in Section 4.4, we computed the distributions of the octave error features for the reference pitch (solid red), the lower octave (dashed blue) and the higher octave (dotted green). Figure 4.7 shows the result for the different polyphony rates. The first observation we make is that, as in our hypothesis, the mean of the octave error feature for the reference pitch is highest, followed by those of the higher octave pitch distribution and finally the lower octave pitch presents the lowest mean. We may also notice that the individual distributions have a large spread.

5. These recordings are copyrighted and available online: http://ericwhitacre.com/the-virtual-choir/resources

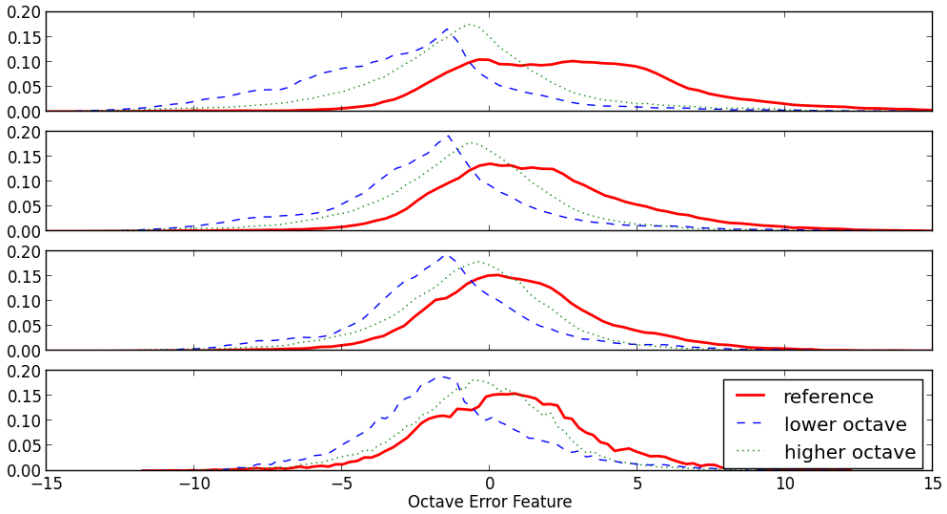

Figure 4.7: Histogram of octave error feature by polyphony.

Another important observation is that the overlap between the distributions rises with the polyphony rate. This can be explained by the fact that the background spectrum is higher when the polyphony increases. This happens, for example, when partials of other instruments appear between the partials of the reference. Another source of this effect could be the overlap of the partials of other instruments with the ones from the reference. Finally, this could also occur because the harmonic spectral envelope (HSE) of the reference pitch cannot be assumed to be smooth enough to avoid aliasing.

The overlap of the distributions is an issue if using the octave error feature as an absolute value for simple thresholding classification. This is the case when we must decide if a given individual pitch contour is a false positive or a true positive. However this feature can be used in conjunction with other features of the frame to perform a higher dimensional classification. We may consider two other frame features such as the pitch value and salience that are linked to this feature and could aid in obtaining a larger disjointness. Another consideration to keep in mind is that these are frame-based features, aggregation of multiple frames could also decrease the overlap.

When performing multipitch estimation and tracking we may identify candidates that are separated by an octave and share a very close contour evolution. In these cases we may assume one of them can be an octave

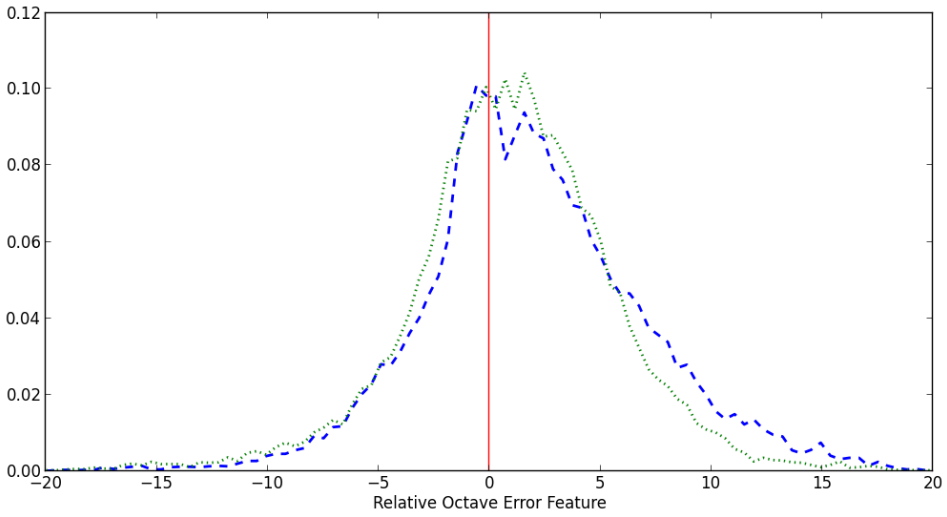

Figure 4.8: Histogram of octave error relative feature for the lower octave (*dashed*) and higher octave (*dotted*).

error and a selection should be performed. In this situation we do not need to perform a classification based on an absolute value of the octave error feature and we may use the relative value. In Figure 4.8 we show the relative difference between the reference and higher octave (dashed blue) and between the reference and lower octave (dotted green) for all polyphony rates aggregated. As expected the means of both distributions are over zero, with the lower octave presenting a smaller overlap with the negative plane than the higher octave.

Overall the results show that on average the proposed features could help improve octave error decisions. However we also observe that the assumptions are not always valid leading to ambiguous feature values in certain cases. There are other sources present in the mixture and the background spectrum is not always lower than the partials. This issue is more visible as the polyphony rate of the mixture increases. Furthermore the timbre of the true pitch is not always necessarily smooth.

We have also conducted quantitative tests, by performing T-tests to show the effect of the polyphony and dataset on this feature. Table 4.3 shows the T-test results for the different polyphonies.

The T-tests reveal that there is a significant difference between the feature

	df	reference	lower octave	higher octave
2	293784	M=2.6;SD=3.9	M=-2.9;SD=3.9; t(df)=-386.0	M=-1.0;SD=3.6; t(df)=-261.1
3	373276	M=1.6;SD=3.4	M=-1.9;SD=3.6; t(df)=-301.9	M=-0.9;SD=3.2; t(df)=-222.7
4	203916	M=0.9;SD=3.1	M=-1.5;SD=3.2; t(df)=-171.6	M=-0.7;SD=2.9; t(df)=-120.5
5	39744	M=0.6;SD=2.9	M=-1.3;SD=2.9; t(df)=-67.7	M=-0.4;SD=2.6; t(df)=-36.0

Table 4.3: T-tests of the higher and lower octave pitches with relation to the reference pitch. The first column indicates the polyphony number of the tested dataset. In all T-tests $p \ll 0.01$.

values of the reference pitches and those of the octave errors. This difference decreases as the polyphony of the mixture increases. These results show that the octave error feature can be useful to differentiate between the true pitches and the octave errors.

Figure 4.9 illustrates the advantages of using the Gaussian fitting pitch tracking method and the octave error feature for pitch selection. In this example we analyze an excerpt of an in-house multitrack professional music recording containing singing voice, electric guitar, bass and drums. The groundtruth pitch track is created by performing a single pitch analysis using the well-known YIN pitch estimation method (de Cheveigné and Kawahara, 2002) on the isolated singing voice track. The proposed extensions show an improvement in the predominant pitch estimation, with fewer octave errors and pitch discontinuities.

Furthermore Figure 4.10 shows the use of the proposed techniques in a task of multiple fundamental frequency estimation and tracking. In this figure we plot the estimated pitches of our proposed method on a mixture of bassoon, saxophone and violin in the excerpt *Fur Deinen Thron* which can be found in the *Bach10* (Duan et al., 2010) pitch-annotated dataset [6]. We can observe how the proposed method recovers most of the pitch contours, even in sections where the sounds produced by the instruments have a low energy. This is the case at the boundaries of the pitch contours, such as at the end of the vibratos at seconds 16 and 24. These regions are marked as non-pitched on the groundtruth annotations, however listening to the excerpt we notice that the instruments are present with low energy and with only one or two partials over the background spectral noise. We also note that there are still many false positives which would lead to low accuracy in subjective evaluations. This is due to the fact that currently we do

6. http://www.cs.northwestern.edu/~zdu459/multipitch/multipitch.html

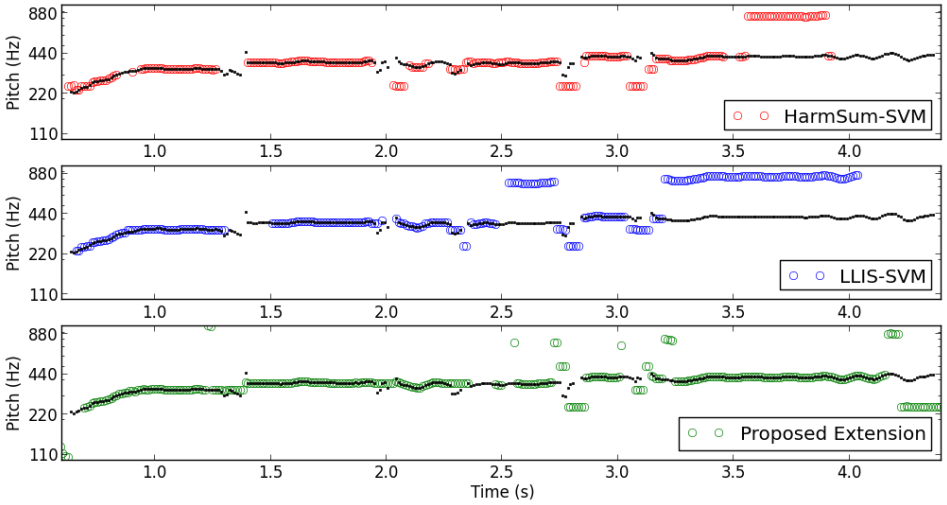

Figure 4.9: Comparison of predominant pitch estimation and tracking methods.

not perform an estimation of the polyphony in the piece and a constant number of pitch tracks are assumed present at all times. In the future this method of multipitch estimation and tracking should be extended to perform polyphony estimation and be objectively evaluated and compared with existing techniques (Pertusa and Inesta, 2008; Duan et al., 2010).

Conclusion

In this section we have shown that the Tikhonov regularization method for spectrum factorization can be useful in tasks of multipitch estimation under low-latency constraints. We have proposed a multiple fundamental frequency estimation method based on a two stage approach, where tracking of pitch candidates and selection of pitch contours are performed independently.

We have proposed solutions for two main issues that arise when working with joint pitch likelihood estimations such as the ones produced using Tikhonov regularization:

– The spreading of the likelihood function due to frequency modulation is addressed by using Gaussian fitting.
– The octave error ambiguities are handled by computing an octave error feature based on the timbre smoothness assumption. We also presented

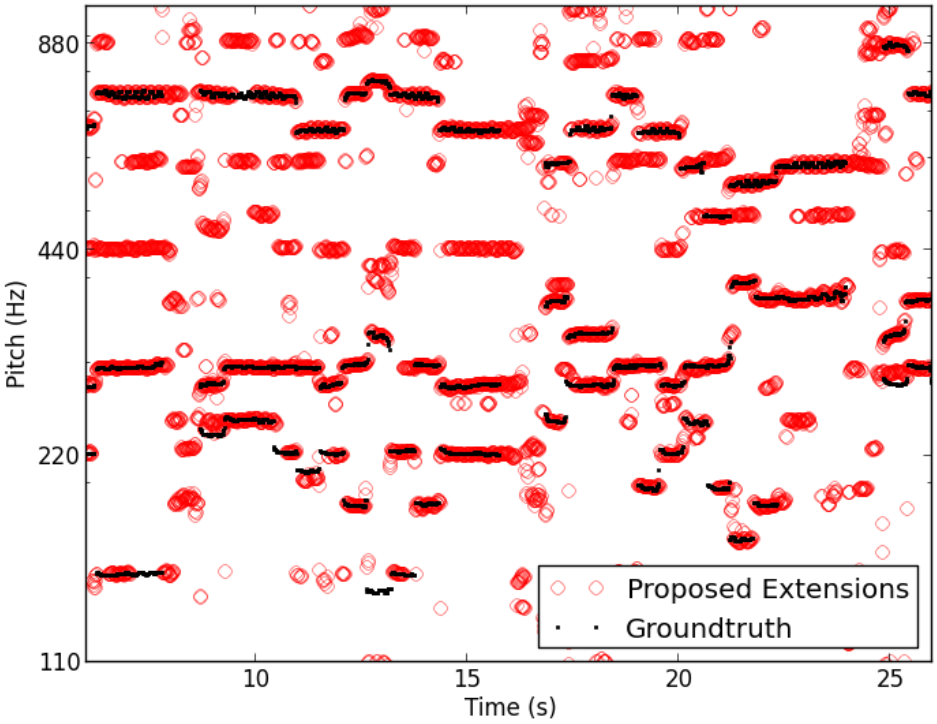

Figure 4.10: Example of multiple fundamental frequency estimation and tracking.

a preliminary evaluation of this solution to differentiate between octave errors and true pitches.

In future work the pitch tracking stage could be improved by adding an extra state to the HMM representing hidden or masked peaks. We could also consider adding a dynamic model for the pitch tracks that could account for the track direction. With respect to the selection stage, future research should focus on testing different classification techniques and adding other dimensions to the pitch track features such as the pitch and salience. Aggregations of multiple frames of a given pitch contour should also be considered. The Gaussian fitting, octave error and predominant pitch estimation and tracking extensions show promise and future work on them might prove useful. These suggestions should be tested in the context of pitch, melody and multipitch estimation by performing a quantitative evaluation and comparison using state-of-the-art techniques.

Chapter 5

Low Latency Audio Source Separation

5.1 Introduction

Audio source separation consists in retrieving one or more audio sources given a set of one or more mixture signals. Audio source separation in the field of music processing has received special attention in the past few decades. A number of methods have been proposed, most of them based on time-frequency masks. We differentiate between two main strategies in the creation of time-frequency masks depending on the constraints imposed on the separation process.

Realtime solutions are often based on binary masks, because of their simple and inexpensive computation. These solutions assume the target sources are orthogonal in the time-frequency domain. The most common binary mask used in stereo music recordings is based on panning information of the sources (Yilmaz and Rickard, 2004; Jourjine et al., 2000; Vinyes et al., 2006).

Non-realtime approaches do not make such an orthogonality assumption, and make use of a soft mask based on Wiener filtering (Benaroya et al., 2006) which requires estimating all spectrograms of the constitutive sources. For harmonic sources this estimation is often performed in two steps. First the pitch track of the target source is estimated and then the spectrum of that given pitch track is estimated. The first step often relies on melody extraction algorithms (Goto and Hayamizu, 1999; Fujihara et al., 2006). Some

methods estimate the pitch of the components independently (Ryynänen, 2006), while others perform a joint estimation of the pitches in the spectrum (Ryynänen, 2006; Yeh et al., 2010). Most joint pitch estimation methods are computationally expensive since they evaluate a large number of possible pitch combinations. NMF approaches to multipitch likelihood estimation (Sha and Saul, 2005; Févotte et al., 2009) address this limitation by factoring the spectrogram into a multiplication of two positive matrices, a set of spectral templates and a set of time-dependent gains. In Durrieu et al. (2010) and Ozerov et al. (2010) the spectral templates are fixed to a set of comb filters representing the spectra generated by each individual pitch spectrum.

Here we briefly introduce a state of the art method used in music source separation under low latency realtime constraints based on pan-derived binary masks. We generalize binary masks under the perspective of spectrum bin classification techniques, where several spectral bin features can be taken into account for the classification (e.g. the bin position with respect to the harmonic structure of the estimated pitch). As an alternative, we propose applying the Tikhonov Regularization spectrum decomposition technique. This technique leverages the low computational cost and realtime property of the Tikhonov regularization and the advantages of using spectrum decomposition to perform source estimation. In addition it allows using Wiener filtering, which results in a soft time-frequency mask separation technique, with the goal of decreasing the presence of artifacts and musical noise in the resulting signal. Tikhonov regularization spectral decomposition is applied to three common source separation scenarios in the context of Western commercial music:

– The separation of the singing voice
– The separation of the drums
– The separation of the bass line

Several modifications and extensions of the signal model are presented in order to adapt it to each scenario, however the focus of this work remains on the use of Tikhonov regularization as a spectrum decomposition technique. The study of how the signal model affects the decomposition and posterior separation is not the main research target and will only be considered in certain specific contexts.

5.2 Singing Voice Separation Using Binary Masks

We propose combining several sources of information for the creation of the binary mask in order to better the results of currently existing methods while maintaining low latency. We propose two main sources of information for the creation of the masks. Spectral bin classification based on measures such as lateralization (panning), phase difference between channels and absolute frequency is used to create a first mask. Information gathered through a pitch-tracking system is used to create a second mask for the harmonic part of the main melody instrument.

A similar system was proposed in Fujihara et al. (2006). The authors perform a pitch likelihood estimation, a pitch tracking process and finally a timbre classification of the detected pitch using Gaussian Mixture Models. In their work the main focus is on the vocal pitch detection and they demo the possibilities of using such output for audio source separation tasks. In our case we are focusing on the source separation capabilities of a similar system with realtime constraints. We also present the use and combination of other non-harmonic masks such as those based on lateralization and absolute frequency.

Introduction

We present the use of binary masks as a common method for separating musical sources under realtime and low-latency constraints. These methods are generalized as spectral bin classification separation techniques. Spectral bin classification is performed using simple thresholding or decision trees that can easily integrate user control information, especially interesting in interactive applications. The classification is based on features such as panning and frequency, as well as the distance to a hypothetical harmonic position.

This approach is taken as a starting point and a baseline for the methods we propose.

Spectral Bin Classification Masks

Panning information is one of the features that have been used successfully (Jourjine et al., 2000; Yilmaz and Rickard, 2004) to separate sources in real-time. Vinyes et al. (2006) used the pan and the IPD (inter-channel

phase difference) features to classify spectral bins. An interesting feature for source separation is the actual frequency of each spectrum bin, which can be a good complement when the panning information is insufficient. Using pan and frequency descriptors we define a filter in the frequency domain using a binary mask to mute a given source:

$$m_t^{pf}[\omega] = \begin{cases} 0 & \text{if } p_{low} < p_t[\omega] < p_{high} \text{ and } \omega_{low} < \omega < \omega_{high}, \\ 1 & \text{otherwise.} \end{cases}$$

where $p_t[\omega]$ is the pan value of the spectral bin ω at frame t. The parameters p_{low} and p_{high} are the pan boundaries and ω_{low} and ω_{high} are the frequency boundaries fixed at -0.25, 0.25 and 60Hz and 6000Hz respectively, to keep the method unsupervised.

Results from this method are acceptable in some situations. The most obvious limitation is that it is not capable of isolating sources that share the same pan/frequency region. This technique is also ineffective in the presence of strong reverberation or in mono recordings which have no pan information.

Harmonic Mask

Harmonic mask creation is based on two assumptions: that the vocal component is fully localized in the spectral bins around the position of the singing voice partials and that the singing voice is the only source present in these bins. Under such assumptions an optimal mask to remove the singing voice consists of zeros around the partials positions and ones elsewhere.

These assumptions are often violated. The singing voice is composed of other components than the harmonic components such as consonants, fricatives or breath. Additionally other sources may contribute significantly to the bins where the singing voice is located. This becomes clear in the results where signal decomposition methods such as Instantaneous Mixture Model (IMM) (Durrieu et al., 2010) that do not rely on such assumptions perform better than our binary mask proposal. However these assumptions allow us to greatly simplify the problem.

Under these assumptions we define the harmonic mask m^h to mute a given source as:

$$m_t^h[\omega] = \begin{cases} 0 & \text{for } (f0_t \cdot h) - L/2 < \omega < (f0_t \cdot h) + L/2, \forall h, \\ 1 & \text{otherwise.} \end{cases}$$

where $f0_t$ is the pitch of the t^{th} frame, and L is the width in bins to be removed around the partial position.We may also combine the harmonic and spectral bin classification masks using a logical operation by defining a new mask m_t^{pfh} as:

$$m_t^{pfh}[\omega] = m_t^{pf}[\omega] \vee m_t^{h}[\omega] \tag{5.1}$$

Finally, we are also able to produce a *soloing* mask $\bar{m}_t[\omega]$ by inverting any of the previously presented muting masks $\bar{m}_t[\omega] = \neg m_t[\omega]$.

In order to estimate the pitch contour $f0_t$ of the chosen instrument, we follow a three-step procedure: pitch likelihood estimation, timbre classification and pitch tracking presented in Section 4.3.

Evaluation

The material used in the evaluation of the source separation method consists of 15 multitrack recordings of song excerpts with vocals, compiled from publicly available resources (MASS [1], SiSEC [2], BSS Oracle [3])

Using the well known BSSEval toolkit (Vincent et al., 2007c), we compare the Signal to Distortion Ratio (SDR) error (difference from the ideal binary mask SDR) of several versions of our algorithm and the IMM approach (Durrieu et al., 2010). The evaluation is performed on the "all-minus-vocals" mix versions of the excerpts. Table 5.1 presents the SDR results averaged over 15 audio files in the dataset. We also plot the results of individual audio examples and the average in Figure 5.1. The *Pan-freq mask* method consists of applying the m^{pf} mask from Equation 5.1.

The quality of our low-latency approach to source separation is not as high as for off-line methods such as IMM, which shows an SDR almost 3 dBs higher. However, our LLIS-SVM method shows an increase of 2.2 dBs in the SDR compared to the LLIS-noSVM method. Moreover, adding azimuth information to the multiplicative mask (method *LLIS-SVM-pan*) increases the SDR by 0.7 dBs.

Conclusions

We present a source separation approach well suited to low-latency applications. The separation quality of the method is inferior to offline approaches,

1. http://www.mtg.upf.edu/static/mass
2. http://sisec.wiki.irisa.fr/
3. http://bass-db.gforge.inria.fr/bss_oracle/

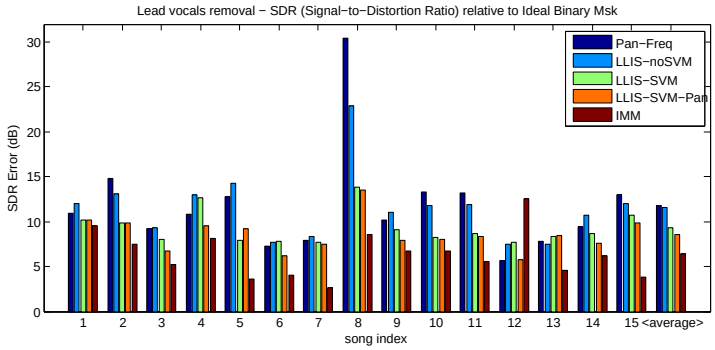

Figure 5.1: SDR error of various excerpts for four methods: pan-frequency mask, LLIS and IMM.

Method	SDR-vocals	SDR-accomp
pan-freq	0.21	4.79
LLIS-noSVM	0.47	5.05
LLIS-SVM	2.70	7.28
LLIS-SVM-pan	3.43	8.01
IMM	6.31	10.70
Ideal	12.00	16.58

Table 5.1: Signal-To-Distortion Ratio (in dB) for the evaluated methods. The Ideal column shows the results of applying an ideal binary mask with zeros in the bins where the target source is predominant and ones elsewhere.

such as NMF-based algorithms, but it performs significantly better than other existing real-time systems. Maintaining low-latency (232 ms), an implementation of the method runs in real-time on current, consumer-grade computers. The method only targets the harmonic component of a source and therefore does not remove other components such as the unvoiced consonants of the singing voice. Additionally it does not remove the reverberation component of sources. However these are limitations common to other state-of-the-art source separation techniques and are out of the scope of our study.

We propose a method with a simple implementation for low-latency pitch likelihood estimation. It performs joint multipitch estimation, making it well-adapted for polyphonic signals. We also introduce a technique for de-

tecting and tracking a pitched instrument of choice in an online manner by means of a classification algorithm. This study applies the method to the human singing voice, but is general enough to be extended to other instruments.

Finally, we show how the combination of several sources of information can enhance binary masks in source separation tasks. The results produced by the ideal binary mask show that there are still improvements to be made.

5.3 Singing Voice Source Estimation Using Wiener Filtering

We present a Tikhonov regularization-based method as an alternative to the Non-negative Matrix Factorization (NMF) approach for source separation in professional audio recordings. This method is a direct and computationally less expensive solution, which makes it useful in low-latency scenarios. The technique removes the non-negativity constraint which characterizes NMF in exchange for a closed-form solution to the problem of spectrum factorization. We quantitatively evaluated it in terms of reconstruction and separation quality on a dataset of excerpts of professionally recorded songs with singing voice. Results show that the the proposed approach achieves quality similar to that of NMF.

Introduction

Spectrum decomposition has often been used in audio transcription and source separation tasks. It consists in modeling the spectral representation of a signal as a combination of a set of spectral components.

Some techniques such as Harmonic Temporal Clustering (HTC) (Kameoka et al., 2007; Kameoka, 2007) propose spectrum components with parametrized frequency and temporal envelopes and with a fixed harmonic structure. Similarly Wu et al. (2011) consider components for the modeling of transients. In both cases the parameters are found using iterative Expectation Maximization update rules.

Non-negative Matrix Factorization (NMF) has received a lot of attention in the past few years. NMF was first introduced in the context of music transcription in Smaragdis and Brown (2003). The main strengths of such methods are the non-negativity constraints on the component gains, the ability to learn the components and its flexibility in adding additional cost

terms. Raczyński et al. (2007) use a harmonic initialization of the components and musically inspired penalties on the factorization. Durrieu et al. (2010) propose an NMF method to decompose a signal using a source-filter model and then performing NMF on the residual. Ozerov et al. (2010) present a source separation framework in which priors on the distributions of the spectral components can be introduced in a hierarchical way. In all cases the decomposition is performed by iterating over a set of multiplicative rules.

Existing spectrum decomposition methods have proven useful in audio source separation tasks, however their iterative nature carries a high computational cost. We present here an alternative method based on Tikhonov regularization that sacrifices the flexibility and the non-negativity constraints of NMF or the generality of other methods in exchange for a direct and rapid solution with a much lower computational cost.

Signal Decomposition Model

The main assumption of our spectrum decomposition method is that the short-term Fourier transform (STFT) of our audio signal, $\boldsymbol{V}$ is a linear combination of N_W elementary spectra, also called basis components. This can be expressed as $\boldsymbol{V} = \boldsymbol{W}\boldsymbol{H}$ where $\boldsymbol{V} \in \mathbb{R}^{N_\omega \times 1}$ is the spectrum at a given frame t, N_ω being the size of the spectrum. $\boldsymbol{W} \in \mathbb{R}^{N_\omega \times N_W}$ is the matrix whose columns are the basis components, it is also referred to as the basis matrix. $\boldsymbol{H} \in \mathbb{R}^{N_W \times 1}$ is a vector of component gains for the current frame.

Our focus is on low latency, unsupervised applications which require the decomposition of each spectrum frame to be done very quickly. Therefore, we will only consider solutions in which the basis components $\boldsymbol{W}$ are constant and fixed a priori.

It is obvious that the choice of the basis matrix has a large influence on the decomposition results. It is not in the scope of this experiment to study the effect of the basis matrix, but rather to propose a computationally cheap method to perform the decomposition given a suitable basis matrix.

As in other NMF-based approaches (Virtanen, 2007; Durrieu et al., 2010), we set the basis matrix to be a set of N_L single-pitch, multiple-harmonic spectra. We must allow different spectral envelopes in order to model harmonic sources of different timbres. Therefore we filter the single-pitch components with a filterbank of N_I filters. This results in a total of $N_L \cdot N_I$ harmonic basis components.

Modeling only harmonic sources is often not enough to explain all the possible observed spectra. Wu et al. (2011) propose modeling wideband components to reconstruct transient sounds or background noise. We take a similar approach by adding to our basis matrix the spectra of the filters in our filterbank as wideband components. This results in a total of $N_W = (N_L+1)\cdot N_I$ basis components.

The spectra components can be defined as:

$$\begin{aligned} \varphi[l,n] &= 2\pi f_l H N_L \frac{2^{\frac{iH-F/2+n}{HN_L}}-1}{S_r \ln(2)} \\ E_l[\omega] &= \sum_{n=0}^{F} w[n] \left(\sum_{h=1}^{N_h} sin\left(h\varphi[l,n]\right) \right) e^{-j\omega n} \\ \boldsymbol{W}_{l,i}[\omega] &= \begin{cases} U_i[\omega]E_l[\omega] & \text{if } l \leq N_L \\ U_i[\omega] & \text{if } l = N_L+1 \end{cases} \end{aligned} \tag{5.2}$$

with $H = (1-\alpha)F$. Where α is a coefficient to control the frequency overlap between the components, F is the frame size, S_r the sample rate, $w[n]$ is the analysis window, N_h is the number of harmonics of our components, $\boldsymbol{W}_{l,i}$ is the spectrum of the component of l^{th} pitch filtered by i^{th} filter. U_i is the spectrum of the i^{th} filter in our filterbank. U_i is constructed as a sequence of N_I Hann windows, linearly distributed in the Mel scale and with a 50% overlap.

The column vectors $\boldsymbol{W}_{l,i}$ are stacked horizontally to form the matrix $\boldsymbol{W}$. This results in the spectrum $\boldsymbol{W}_{l,i}$ of the component of l^{th} pitch and i^{th} filter being the column vector $\boldsymbol{W}_{lN_I+i}$.

Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) has been widely used in audio source separation tasks (Durrieu et al., 2010; Ozerov et al., 2010; Févotte et al., 2009). The NMF-based approach to solving our spectrum decomposition problem $\boldsymbol{V} = \boldsymbol{W}\boldsymbol{H}$ consists in finding the best non-negative estimate of the component gains $\hat{\boldsymbol{H}}$ that minimizes a given objective function. We

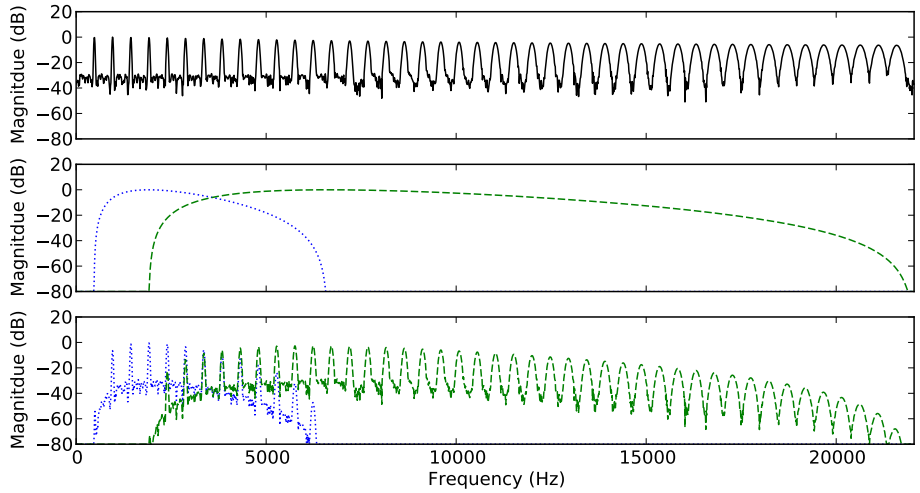

Figure 5.2: Two components of our basis matrix $\boldsymbol{W}$. Top shows $E_l[\omega]$ for a frequency of 480Hz. Middle shows $U_i[\omega]$ for two consecutive values of i. Bottom shows $\boldsymbol{W}_{l,i}[\omega]$ for the selected $E_l[\omega]$ and $U_i[\omega]$.

consider the following objective functions:

$$\Phi_{euc}(\boldsymbol{H}) = \sum_{k=1}^{N_\omega} \frac{1}{2} \left([\boldsymbol{W}\boldsymbol{H}]_k - [\boldsymbol{V}]_k\right)^2 \tag{5.3}$$

$$\Phi_{kl}(\boldsymbol{H}) = \sum_{k=1}^{N_\omega} [\boldsymbol{V}]_k log \frac{[\boldsymbol{V}]_k}{[\boldsymbol{W}\boldsymbol{H}]_k} - [\boldsymbol{V}]_k + [\boldsymbol{W}\boldsymbol{H}]_k \tag{5.4}$$

$$\Phi_{is}(\boldsymbol{H}) = \sum_{k=1}^{N_\omega} \frac{[\boldsymbol{V}]_k}{[\boldsymbol{W}\boldsymbol{H}]_k} - log \frac{[\boldsymbol{V}]_k}{[\boldsymbol{W}\boldsymbol{H}]_k} - 1 \tag{5.5}$$

where $[\boldsymbol{X}]_k$ is the k^{th} element of vector $\boldsymbol{X}$. It is well known (Févotte et al., 2009) that the solution to the non-negative factorization problem given these objective functions results in the following multiplicative update rule:

$$\hat{\boldsymbol{H}}_n^{NMF} = \hat{\boldsymbol{H}}_{n-1}^{NMF} \otimes \frac{\boldsymbol{W}^t \left(\left(\boldsymbol{W}\hat{\boldsymbol{H}}_{n-1}^{NMF} \right)^{[\beta-2]} \otimes \boldsymbol{V} \right)}{\boldsymbol{W}^t \left(\boldsymbol{W}\hat{\boldsymbol{H}}_{n-1}^{NMF} \right)^{[\beta-1]}} \tag{5.6}$$

where $\otimes$ is the Hadamard product (an elementwise multiplication of the matrices), all divisions are elementwise and $0 \leq \beta \leq 2$ is the coefficient that

will define the objection function that is being minimized. $\beta = 2$ for the Euclidean distance (NMF_{euc}) $\beta = 1$ for the Kullback-Leibler divergence (NMF_{kl}) and $\beta = 0$ for the Itakura-Saito divergence (NMF_{is}). Finally n is the iteration of the solution and $\hat{\boldsymbol{H}}_0^{NMF}$ is a random positive vector.

Tikhonov Regularization

The condition number of the basis matrix $\boldsymbol{W}$ defined in Equation 5.2 is very high ($\kappa(\boldsymbol{W}) \approx 5.9 \cdot 10^{17}$), therefore we may assume that our problem is ill posed. This could be due to the harmonic structure and correlation between the components in our basis matrix.

We propose using the Tikhonov regularization (TR) method (Tikhonov, 1963) to find an estimate of the components gains vector $\hat{\boldsymbol{H}}$ given the spectrum $\boldsymbol{V}$. This consists in minimizing the following objective function:

$$\Phi_{TR}(\boldsymbol{H}) \quad = \quad \sum_{k=1}^{N_\omega} ([\boldsymbol{W}\boldsymbol{H}]_k - [\boldsymbol{V}]_k)^2 + ([\Gamma\boldsymbol{H}]_k)^2$$

where Γ is the Tikhonov matrix that defines the preference among all possible solutions. In this study we set $\Gamma = \lambda\boldsymbol{L}$ where $\boldsymbol{L} \in \mathbb{R}^{N_W \times N_W}$ is a singular matrix that allows weighting the *a priori* probabilities of the solutions. λ is a positive scalar hyperparameter. This parameter controls the effect of the regularization on the estimated solution.

We decided to give preference to solutions with low norms while compensating for biases due to energy differences between components of different pitch. This is known as the Weighted Minimum Norm Estimate (WMNE) and it can be achieved by defining $\boldsymbol{L}$ as a diagonal matrix such that:

$$diag(\boldsymbol{L})_{iN_I+k} = \sqrt{\sum_{\omega=1}^{N_\omega} \sum_{k=1}^{N_I} \boldsymbol{W}_{i,k}^2[\omega]}$$

where $i = 1...N_L + 1$ and $k = 1...N_I$. The main reason for such a choice is that we assume that the basis components correspond quite well to the sources in the audio signal and only a few sources are simultaneously present in the audio. Therefore the gains of the components should have few high values and many low values, leading to a small norm.

The TR method, results in the following closed-form solution $\hat{\boldsymbol{H}}^{TR} = \boldsymbol{R}\boldsymbol{V}$ where $\hat{\boldsymbol{H}}^{TR}$ is the estimated components gains, and $\boldsymbol{R}$ is defined as:

$$\boldsymbol{R} = (\boldsymbol{L}^t\boldsymbol{L})^{-1}\boldsymbol{W}^t[\boldsymbol{W}(\boldsymbol{L}^t\boldsymbol{L})^{-1}\boldsymbol{W}^t + \lambda I_{N_\omega}]^+$$

where $[\boldsymbol{Z}]^+$ denotes the Moore–Penrose pseudoinverse of $\boldsymbol{Z}$. The calculation of $\boldsymbol{R}$ is computationally costly, however this operation is independent of the input spectra and can be performed before the analysis of the audio signal. The $\boldsymbol{R}$ matrix only depends on $\boldsymbol{W}$ and Γ. As we saw in section 5.3 the $\boldsymbol{W}$ only needs the parameters of the analysis process, therefore the only operation that is performed at each frame is $\hat{\boldsymbol{H}}^{TR} = \boldsymbol{R}\boldsymbol{V}$.

Compared to the NMF method, the TR approach does not constrain the component gains to be non-negative. However, as we will show in the experiments, this assumption has little impact on the performance of the reconstruction and source separation tasks.

Evaluation

The main goal of the experiment is to compare the TR closed-form solution with the NMF solution in the general context of source separation. The comparison will be made on two main factors:
– How faithful is the factorization to the data?
– How well does the factorization separate the data?
In order to evaluate the factorization quantitatively, we simply compare the Signal to Noise Ratio (SNR) of the reconstruction without modifying the factors (components and gains). The reconstruction is computed as $\hat{\boldsymbol{V}} = \boldsymbol{W}\hat{\boldsymbol{H}}$. And the SNR calculation is performed in the frequency domain:

$$SNR = 10 \cdot \log_{10} \frac{\sum \boldsymbol{V}^2}{\sum \left| \hat{\boldsymbol{V}} - \boldsymbol{V} \right|^2} \tag{5.7}$$

To quantitatively evaluate how well the factorization separates the data, we perform a simple separation of the vocal track on a set of audio recordings. The separation produces two versions of the excerpt, one with only the voice track (foreground) and another with all but the voice track (background). We follow the same procedure as in Durrieu et al. (2009a) for the separation. We reconstruct the spectrum selecting the candidates in $\hat{\boldsymbol{H}}$ that correspond to the voice. We have run two different tests: a supervised test in which the pitch of the vocal track is estimated in a previous stage using the well known Yin method (de Cheveigné and Kawahara, 2002) on the vocal track in isolation, and an unsupervised test in which the pitch is estimated using $\hat{\boldsymbol{H}}$:

$$i^{f_0} = \arg \max_{i=1\ldots N_L} \left(\sum_{k=1\ldots N_I} \max \left(\hat{\boldsymbol{H}}_{i,k}, 0 \right) \right)$$

where i^{f_0} is the index corresponding to the f_0 at a given frame. Due to correlations between pitches with harmonic relations, we also remove pitches that are at intervals Θ_f (Θ_i in pitch index units) from the predominant pitch.

$$i_{sel} = \{i^{f_0} + o | o \in \Theta_i\}$$

Since the voice often presents pitch fluctuations a series of adjacent basis components will also be selected. In our experiments, we select Δf semitones (Δi in pitch index units) around the selected pitches. This results in the following set of selected indices:

$$C_{sel} = \{(i \pm j)N_I + k | i \in i_{sel},\ j \leq \frac{\Delta i}{2},\ k \leq N_I\}$$

where $j \geq 0$ and $k \geq 1$. The estimate of the foreground and background spectra are computed using a binary mask $M \in \mathbb{R}^{N_W \times 1}$ on the component gains:

$$\boldsymbol{m}_l = \begin{cases} 1 & \text{if } l \in C_{sel} \\ 0 & \text{else} \end{cases}$$

$$\hat{\boldsymbol{V}}_f = \gamma(M \otimes \hat{\boldsymbol{H}})\boldsymbol{W} \qquad \hat{\boldsymbol{V}}_b = ((1 - M) \otimes \hat{\boldsymbol{H}})\boldsymbol{W}$$

where $\gamma > 1$ is a gain on the foreground estimation. This is needed because part of the target source energy is actually spread in other pitch components that share harmonic relations, such as fifths and octaves.

Once we have the spectra estimates we calculate the actual foreground and background Discrete Fourier Transform (DFT) signals using Wiener filtering:

$$\hat{S}_f = \frac{\hat{\boldsymbol{V}}_f^2}{(\hat{\boldsymbol{V}}_f^2 + \hat{\boldsymbol{V}}_b^2)} S \qquad \hat{S}_b = \frac{\hat{\boldsymbol{V}}_b^2}{(\hat{\boldsymbol{V}}_f^2 + \hat{\boldsymbol{V}}_b^2)} S$$

where S is the original DFT of the mix signal. Note that even though the mask applied to the component gains is binary, the final mask applied to the DFT frames is actually a soft mask, resulting from the Wiener filtering. To go back to the time domain we apply a simple overlap and add technique. Finally we evaluate the performance of the separation by computing the Signal to Distortion Ratio (SDR) with the popular audio source separation evaluation toolbox BSS_{EVAL} (Vincent et al., 2006). We compared each method to a baseline obtained with a non-binary oracle separation (Vincent et al., 2007a). The values used in our experiments are the difference between

the measure of each algorithm and the oracle estimation measure, averaged for all audio examples in the dataset. The evaluation material consists of a dataset of 11 multitrack recordings with vocals, compiled from publicly available resources (MASS [4], SiSEC [5], BSS Oracle [6]).

Results

The STFT analysis is performed with a 92ms Blackman-Harris window ($F = 4096$ for signals at sample rate $S_r = 44100$Hz), a hop size of 46ms ($H = 2048$) and a DFT size of 4096 which results in $N_\omega = 2049$. As in other pitch estimation techniques, we apply whitening to the spectrum to enhance the high harmonics by applying a compression factor of $\eta = 0.75$ so that $Y = |S|^\eta$. We also apply this process to the components spectra of matrix $\boldsymbol{W}$. Regarding the parameters of the $\boldsymbol{W}$ matrix, we have set the number of filters $N_I = 12$, the lowest pitch frequency $f_l = 27.5$, the frequency overlap $\alpha = 0.5$, 60 pitches per octave covering a total of 6 octaves ($N_L = 60 \cdot 6 = 360$) and a maximum number of harmonics per component $N_h = 120$. This leads to a total number of components $N_W = 4332$. The factorization has been performed using the presented NMF solution (5.6) for the three objective functions in Eq. 5.3 and the proposed TR method with $\lambda = 10, 1, 0.1, 0.01$. Audio examples from our experiments are available online [7].

In Figure 5.3 we observe the evolution of the SNR with relation to the number of iterations of the NMF approaches. On the same figure we plot the SNR of the TR methods. The results for the NMF behave as expected, constantly growing with the iteration count. The SNR results of the TR approaches demonstrates reconstruction equivalent to NMF-based methods depending on the value of λ. As expected, lower values of λ lead to better reconstruction results. Methods to find optimal λ values will be considered in future work. We have tested the following separation parameter values: source estimation gain $\gamma = 1, 2, 4, 8$, component gains mask width $\Delta f = 0.1, 0.2, 0.4, 0.8$ and intervals for the mask $\Theta_f = \{0\}, \{0, -12, 12\}$. For each factorization method the best parameter combination has been selected for the plots and comparisons. In Figures 5.5 and 5.4 we show the results of our separation tests. As we can see the difference between TR and NMF methods is relatively small (< 2dB). In the supervised scenario

4. http://www.mtg.upf.edu/static/mass
5. http://sisec.wiki.irisa.fr/
6. http://bass-db.gforge.inria.fr/bss_oracle/
7. http://www.mtg.upf.edu/~rmarxer/papers/icassp12

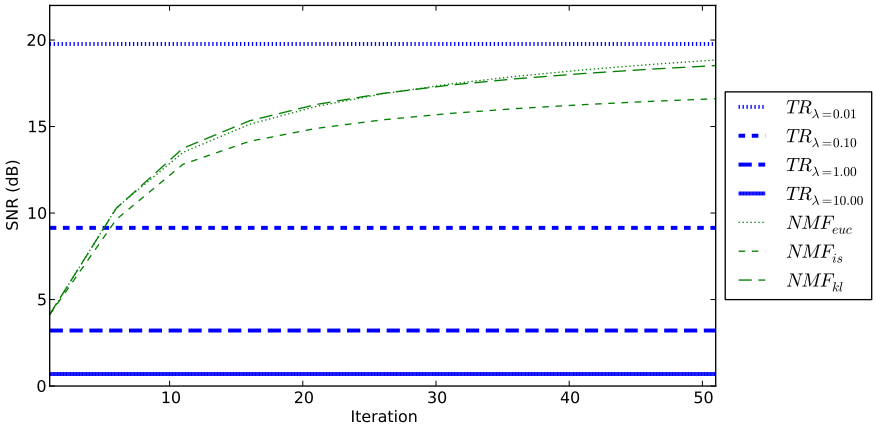

Figure 5.3: Reconstruction SNR versus the factorization method and number of iterations.

of Figure 5.4 we can observe a slightly better performance of NMF with respect to TR. However in the case where the pitch is estimated from $\hat{\boldsymbol{H}}$ the TR method performs better, this could be due to NMF finding and separating better other non-predominant pitches. The TR method has a much lower computational cost and is a closed-form solution that does not require iterations. This makes it much more attractive for low-latency and computation-limited contexts. Taking a closer look at the TR method, we observe that in contrast to the SNR case, lower values of λ do not necessarily lead to better separation.

Conclusions

We present a new spectrum model and factorization method with applications in source separation. This method is based on a Tikhonov regularization (TR) approach to the spectrum decomposition problem and offers a direct and closed-form solution with a significantly lower computational cost than NMF-based methods. We also present a comparative study between the TR approach and the NMF approach in the context of spectrum reconstruction and source separation. The study shows that TR can perform similarly to NMF with the proposed basis matrix.

The comparison has been limited to one single basis matrix in the current study. In future work we should compare the TR method to NMF-based ap-

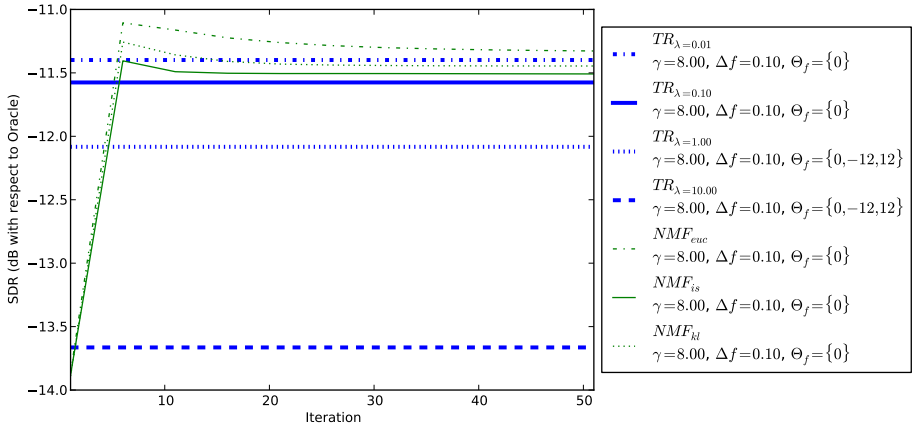

Figure 5.4: Separation SDR for the background source (non vocals track) in the supervised test where the pitch is extracted from the vocal track in isolation.

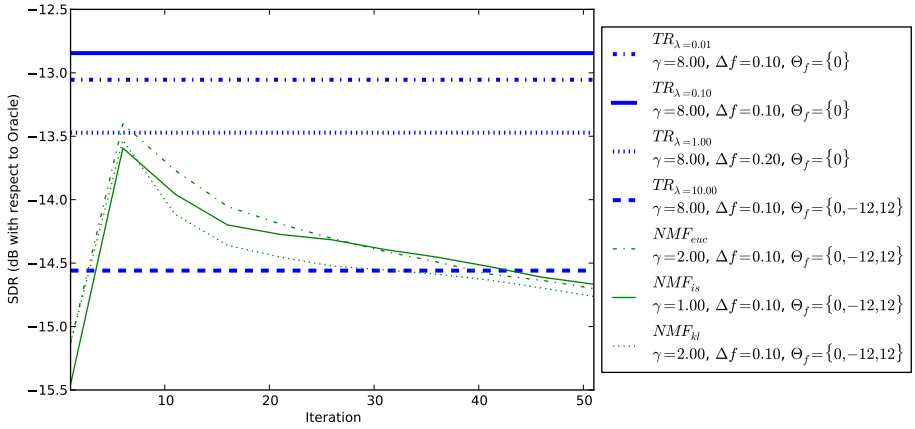

Figure 5.5: Separation SDR for the background source (non vocals track) in the unsupervised test where the pitch has been estimated from $\hat{G}$.

proaches using different basis matrices. Furthermore the flexibility of NMF should be taken into account when comparing the computational cost, for instance source-filter models for the basis matrix could lead to a significantly lower number of components. NMF with sparsity constraints (Raczyński et al., 2007) should also be taken into account. Another direction for future

research consists in exploring the choice of the Tikhonov matrix Γ. Finally other measures (SIR and SAR) should also be evaluated for a more complete comparison.

5.4 Drums Separation

Many recent approaches to musical source separation rely on model-based inference methods that take into account the signal's harmonic structure. To address the particular case of instantaneous percussion separation, we propose a method that combines a harmonic-based decomposition using a Tikhonov regularization-based algorithm with the transient analysis of spectral peaks from a single audio frame. The signal model allows the estimation of harmonic and non-harmonic sources. Later, as shown in the evaluation, adding transient peak information improves the Signal-to-Distortion Ratio (SDR). Compared to other existing methods, this approach achieves comparable performance while being suitable for low-latency conditions.

Introduction

Recent techniques allow for separating instrumental sources from a musical mixture signal. This process may have various application areas including musical production (e.g remixes), entertainment (e.g. karaoke), music analysis (e.g. transcription) or cultural heritage (e.g. restoration). This section addresses the particular case of separating non-harmonic percussion sources (e.g drums, cymbals) in musical mixtures.

In our scenario, we assume that the audio mixture contains one or more harmonic instrumental sources in addition to the percussion to be extracted. The timbre structure of the percussion source is difficult to model, since it might comprehend a large variety of instruments. However, a more distinguishable trait is its time signature, consisting of a sharp attack followed by an exponential decay.

Algorithms such as Non-negative Matrix Factorization (NMF) decompose an input time-frequency representation into basis components without prior knowledge, which allows blind source separation. Most approaches impose additional constraints in the factorization process. For example, some authors (Virtanen, 2007) add temporal continuity constraints, while other approaches force a source/filter decomposition with a set of harmonic patterns and filter banks (Klapuri et al., 2010; Durrieu et al., 2009b). More recently, Ozerov et al. (2010) have proposed a framework that combines

spectral patterns (source/filter model) and temporal patterns (attack and decay envelopes of hundreds of milliseconds). In this case, the factorization step estimates the patterns' gains.

We found methods that specifically address the problem of percussion separation. A two-step separation method, with NMF decomposition and SVM (Support Vector Machines) classification (Helén and Virtanen, 2005), classifies the separated components into drums or pitched. Another approach makes use of drum separation with NMF methods as a pre-process for its classification and transcription (Gillet and Richard, 2008). The Harmonic Percussion Sound Separation (HPSS) method (Ono et al., 2008a) provides an efficient and effective two-dimensional filtering of the spectrogram, to distinguish harmonic components (temporal continuity but spectral discontinuity) from percussive components (temporal discontinuity and spectral wide band energy). This method has proven effective as a pre-process for automatic music description tasks.

Our method combines transient estimation of spectral peaks with a model-based inference algorithm that decomposes the input signal into harmonic and non-harmonic magnitude spectra. The algorithm processes a single frame magnitude spectrum to estimate the two decomposed spectra. Compared to other approaches, our method is causal and therefore appropriate for low-latency situations.

Method

The separation process involves various steps, as shown in figure 5.6. First, the input audio signal is windowed and represented as a sequence of complex spectra using the Short-Time Fourier Transform (STFT). Next, we decompose the magnitude spectrum as a linear combination of basis spectral components with the Tikhonov regularization method presented earlier. Additionally, we extract temporal information by means of a transient analysis of spectral peaks in the current spectrum. Combining this information with the estimated non-harmonic spectrum, we can improve the separation of percussion sources from background noise and other harmonic components.

Signal model

The central part of the source estimation is the signal model, which is built from a set of spectral basis components. Our focus is on low-latency

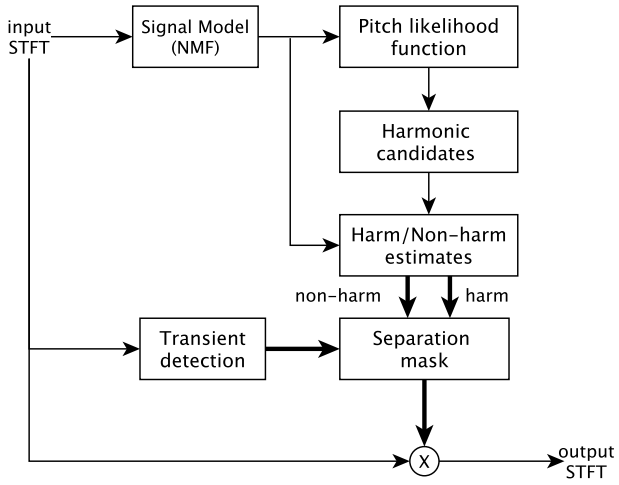

Figure 5.6: The separation mask combines the non-harmonic estimation and the transient peak analysis. Thick arrows represent spectral masks.

applications which require the decomposition of each spectrum frame to be done independently.

For each frame, we assume that the spectrum magnitude $\boldsymbol{V}$ can be decomposed into a linear combination of N_W elementary spectra, also named basis components. This can be expressed as $\boldsymbol{V} = \boldsymbol{W}\boldsymbol{H}$ where $\boldsymbol{V} \in \mathbb{R}^{N_\omega \times 1}$ is the spectrum at a given frame m, N_ω being the size of the spectrum. $\boldsymbol{W} \in \mathbb{R}^{N_\omega \times N_W}$ is the matrix whose columns are the basis components, also referred to as the basis matrix. $\boldsymbol{H} \in \mathbb{R}^{N_W \times 1}$ is a vector of component gains for the current frame.

Spectral basis components $\boldsymbol{W}$ are constant and fixed a priori. It consists of a set of N_L single-pitch multiple-harmonic spectra. In order to model different timbres we must allow different spectral envelopes. This is done by filtering the single-pitch components with a bank of N_I filters. To cope with all possible observed spectra (e.g. in the presence of percussive events or noise), we add a set of filters as wideband components similar to Wu et al. (2011). This results in a total of $N_W = (N_L + 1) \cdot N_I$ basis components. Detailed information about the creation of these spectral basis components can be found in Marxer (2011).

As described in 5.3, solving our spectrum decomposition problem with Tikhonov regularization consists in finding the best component gains vector

$\hat{\boldsymbol{H}}$ that minimizes the objective function based on the Euclidean distance with the norm regularization term. The solution $\hat{\boldsymbol{H}}$ can be computed using a closed form solution based on a simple matrix vector multiplication. Apart from reconstructing the input spectrum, we can use $\hat{\boldsymbol{H}}$ to compute a pitch likelihood function by summing the individual gains corresponding to a given pitch candidate in the basis components matrix $\boldsymbol{W}$.

Harmonic and non-harmonic source estimation

In the decomposition solution, we expect a harmonic instrumental source to contribute principally to specific candidates in the pitch likelihood function. In contrast, we expect percussion source contributions to be distributed over several candidates, both pitched and wide-band filter candidates. The consequence is that non-harmonic sources will show energy spread over the pitch likelihood function and not exclusively localized in individual candidates.

Then to reconstruct the harmonic component, we select K pitch candidates from the pitch likelihood function by means of a peak picking algorithm. Candidates with a likelihood value below an empirically defined threshold τ_1 are discarded. Figure 5.7 shows two pitch likelihood curves corresponding to different time instants of a polyphonic audio mixture, one with percussion and one without.

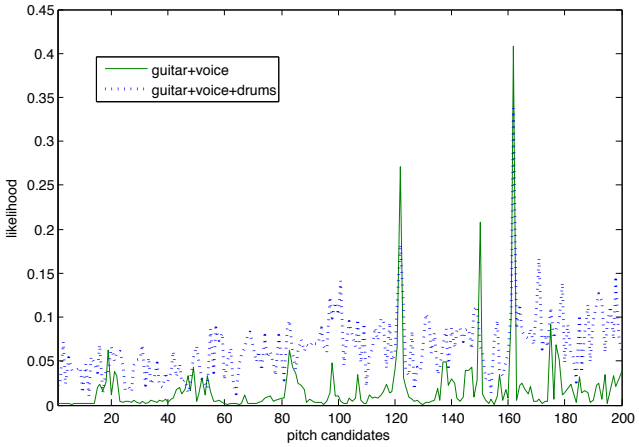

Figure 5.7: Pitch likelihood curves in two different time instants of an audio excerpt containing: vocals and guitar (solid green); and vocals, guitar and drums (dashed blue).

From the estimated vector $\hat{\boldsymbol{H}}$, we create a new vector $\hat{\boldsymbol{H}}_h$ containing non-zero values only at those selected candidates. Therefore, we can compute the harmonic signal estimation as $\hat{\boldsymbol{X}}_h = \boldsymbol{W}\hat{\boldsymbol{H}}_h$. In a complementary fashion, the reconstruction of the non-harmonic part takes a gains vector $\hat{\boldsymbol{H}}_{nh}$ containing non-zero values for the unselected pitch candidates plus the wideband filter banks. The non-harmonic source estimation is computed as $\hat{\boldsymbol{X}}_{nh} = \boldsymbol{W}\hat{\boldsymbol{H}}_{nh}$

With the estimated magnitude spectra $\hat{\boldsymbol{X}}_h$ and $\hat{\boldsymbol{X}}_{nh}$, we can recover a separated output complex spectrum by means of Wiener filtering, as used extensively in recent approaches (Benaroya et al., 2006). Equation 5.8 contains the spectral mask $\boldsymbol{m}_{nh}$, which is then multiplied element-wise by the input complex spectrum $\tilde{\boldsymbol{V}}(\omega, t)$ to reconstruct the non-harmonic signal.

$$\boldsymbol{m}_{nh} = \frac{\hat{\boldsymbol{X}}_{nh}^2}{\hat{\boldsymbol{X}}_h^2 + \hat{\boldsymbol{X}}_{nh}^2} \tag{5.8}$$

By informally listening to the separated non-harmonic signal, we realized that in the presence of a percussion event, the separated signal is weak and lacks clarity. The rationale behind this is that a percussion attack increases the spectrum's energy in the form of wide-band noise. The parameter estimation, instead of representing it exclusively with the wideband (unpitched) filter candidates, also assigns energy to the selected pitch candidates, to represent the percussion spectrum. To help the identification and separation of the percussion contribution in the spectrum, we propose including transient analysis.

Transient analysis

Our aim is to detect transient events in the signal, which should reveal the presence of percussion sources. This analysis can be achieved in the spectral domain by means of the temporal center of gravity of spectral peaks. Given a magnitude spectrum, spectral peaks are detected by localizing local maxima, and neighboring local minima, which determine the spectral peak width.

Röbel (2003) suggested computing the temporal center of gravity (COG) to treat transient events in a phase vocoder algorithm. The COG of an isolated spectral peak can describe how the energy of a given frequency is localized inside the temporal window. It is based on the group delay and it can be computed directly from the bins of a spectral peak in the STFT,

as shown in equation 5.9. If a spectral peak is part of a transient event, its energy will be concentrated at the rightmost part of the window, and will have a high COG value. A spectral peak p that corresponds to a sustained sound will have its energy spread over the whole window, having a COG value near 0.

$$p_{COG} = \frac{\int -\frac{\partial \phi(\omega, p_t)}{\partial \omega} A(\omega, p_t)^2 d\omega}{\int A(\omega, p_t)^2 d\omega} \tag{5.9}$$

Similar to the transient detection in Röbel (2005), which computes statistical measures of COG values of individual peaks p_{COG}, our approach defines $N_I = 14$ bands with a bandwidth of 1500Hz. For each band i, we compute the average of the COG value of individual spectral peaks, referred to as c_i. We create a transient spectral mask $\boldsymbol{m}_t$, in which the bins corresponding to all spectral peaks in a frequency band i are set to one if $c_i > \tau_2$ and set to zero otherwise.

However, the decay of a percussion sound can typically extend over hundreds of milliseconds. To handle the decay, our method keeps a history of N frames (e.g. covering 250 ms) of each band's COG average c_i. First, for a given frame t we compute the time derivative as $\Delta c_i[t] = c_i[t] - c_i[t-1]$ of all band's COG average. Then, for a band i, a binary transient decay value $d_i[t]$ is set to one if two conditions are fulfilled:

if $\max_n c_i[n] > \tau_2$ **and** $\Delta c_i[t] < 0$ **then**
  $d_i[t] = 1$
else
  $d_i[t] = 0$
end if

The leftmost condition in the above pseudo-code requires the presence of a transient event in that past N history frames ($t - N \leq n \leq t$). At the same time, by forcing a negative derivative value $\Delta c_i[t]$, we assure that the transient "is shifting to the left of the window". In order to take the transient decay into account, we compute a decay mask $\boldsymbol{m}_d$, in which the bins of all spectral peaks found in a frequency band i are set to the binary value d_i. Note that the transient analysis does not distinguish between harmonic and non-harmonic transients. Therefore the masks $\boldsymbol{m}_t$ and $\boldsymbol{m}_d$ would let through transients corresponding to harmonic instrumental sources.

Separation mask

Finally, to build the final percussion separation mask $\boldsymbol{m}_p$ for the current frame, we combine the partial masks previously computed ($\boldsymbol{m}_t$, $\boldsymbol{m}_d$, $\boldsymbol{m}_h$ and $\boldsymbol{m}_{nh}$). We have to take into account that, on the one hand, a percussion source will contribute greatly to the estimated harmonic mask $\boldsymbol{m}_h$. Therefore, we cannot achieve the separation only from the estimated non-harmonic mask $\boldsymbol{m}_{nh}$. On the other hand, applying only the transient masks $\boldsymbol{m}_t$ and $\boldsymbol{m}_d$ based on spectral peaks analysis, we would effectively separate the percussion but leaks from other harmonic transients (e.g. bass guitar, piano) would still be present.

To tackle this problem, for those spectral peaks classified as transients in $\boldsymbol{m}_t$, we compare the values of the harmonic mask $\boldsymbol{m}_h$ to a given threshold τ_3 at the center frequency of each spectral peak. Typically, when a spectral peak in the input spectrum corresponds to a harmonic source frequency partial, the estimated value in the harmonic mask at this specific frequency will be high. Hence, we can identify those harmonic transient peaks and not separate them as percussion. During the percussion decay, we proceed in a similar manner, but adding at the same time the estimated non-harmonic mask $\boldsymbol{m}_h$ to the final percussion separation mask $\boldsymbol{m}_p$. This process can be written as mask operations.

$$\begin{aligned} \boldsymbol{m}_\tau &= \begin{cases} 1, & \text{if } \boldsymbol{m}_h < \tau_3 \\ 0, & \text{if } \boldsymbol{m}_h \geq \tau_3 \end{cases} \\ \boldsymbol{m}_p &= \lfloor (\boldsymbol{m}_\tau \otimes \boldsymbol{m}_t) + (\boldsymbol{m}_\tau \otimes \boldsymbol{m}_d \otimes \boldsymbol{m}_{nh}) \rfloor_{0,1} \end{aligned} \tag{5.10}$$

In equation 5.10, a binary matrix $\boldsymbol{m}_\tau$ is computed by thresholding the harmonic mask $\boldsymbol{m}_h$. The operator $\otimes$ denotes Hadamard's (element-wise) product and $\lfloor\ \rfloor_{0,1}$ indicates a clipping of the mask values between 0 and 1. Finally, the separated percussion source $\widehat{\widetilde{\boldsymbol{X}}}_p(\omega, t)$ is computed from the input complex spectrum $\widetilde{\boldsymbol{V}}$ and the separation mask as $\widehat{\widetilde{\boldsymbol{X}}}_p(\omega, t) = \boldsymbol{m}_p \otimes \widetilde{\boldsymbol{V}}(\omega, t)$. The time-domain signal is recovered by means of the inverse Fourier transform and an overlap-add mechanism.

Evaluation

Source separation algorithms can be objectively evaluated if the original multi-track sources are available. We use the same measurements employed in the community evaluation campaigns such as SiSEC (Vincent et al.,

2006): SDR (Signal to Distortion Ratios), SIR (Source to Interference Ratios) and SAR (Sources to Artifacts Ratios). Evaluation material consists of a dataset of 18 multi-track recordings with presence of drums, compiled from publicly available resources (MASS [8], SiSEC [9] and BSS Oracle [10]) and an in-house multitrack dataset.

Table 5.2 shows the results of three configurations of our Transient Harmonic Percussion Separation (THPS) method. Different masks are used, $\boldsymbol{m}_{nh}$ for the non-harmonic separation (THPS-NH), $\lfloor \boldsymbol{m}_t + \boldsymbol{m}_d \rfloor_0^1$ for the transient separation (THPS-TD) and $\boldsymbol{m}_p$ for the final percussion separation (THPS).

We also include two state-of-the-art methods: a custom implementation of the HPSS method (Ono et al., 2008a) and the publicly available implementation of FASST [11] (Ozerov et al., 2010). Additionally we perform an approximately optimal binary mask (ORACLEBIN) used as a glass ceiling reference. We compared each method to a baseline obtained with the oracle separation (Vincent et al., 2007a). Error measures in table 5.2 are the difference between the oracle estimation measure and the measure of each algorithm, averaged for all audio examples in the dataset.

	SIR	**SAR**	**SDR**
THPS-TIK-TD	18.65	9.89	13.34
THPS-TIK-NH	17.62	9.99	11.50
THPS-TIK	15.32	11.83	10.26
FASST	20.01	8.07	11.33
HPSS	16.13	11.50	10.46
ORACLEBIN	-0.60	2.43	1.09

Table 5.2: Average error measures for various algorithms of the low latency drums separation.

For the experiments with the THPS algorithm, we performed an STFT analysis with a Blackman-Harris window 92ms long ($F = 4096$ for signals at sample rate $S_r = 44100$), a hop size of 11ms ($H = 512$) and a DFT size of 8192 which results in $N_\omega = 4097$. Regarding the parameters of the B matrix we have set the number of filters $N_I = 12$, the lowest pitch frequency $f_l = 35$

8. http://www.mtg.upf.edu/static/mass
9. http://sisec.wiki.irisa.fr/
10. http://bass-db.gforge.inria.fr/bss_oracle/
11. http://bass-db.gforge.inria.fr/fasst/

Hz, 40 pitches per octave covering a total of 5 octaves ($N_L = 40 \cdot 5 = 200$). This leads to a total number of components $N_W = 2412$. The number of NMF iterations is set to 15, and harmonic candidate threshold is $\tau_1 = 0.05$, the transient mask threshold is $\tau_2 = 0.3$, and the harmonic mask threshold is $\tau_3 = -30$dB.

The HPSS implementation separates the input signal into two sources: harmonic and percussion. The frame size in this process was set to 1024, which offered a good trade-off between audio quality and vocals/percussion separation. Regarding the FASST framework, we used its default configuration which separates the input signal into four sources: lead melody, bass, drums and other. In our experiment, we consider only the separated *drums* as percussion source.

The results show that our THPS algorithm is comparable to both state-of-the-art methods. It outperforms both partial configurations THPS-NH and THPS-TD, demonstrating the hypothesis of the proposed combination. Figure 5.8 illustrates the SDR error for the individual audio examples in the dataset [12]. It shows how depending on the audio example, one approach may work better than the others, explaining also the similar average results. A perceptual-based evaluation, either by subjective listening tests or using perceptual software toolkits (e.g. PEASS), was not possible to carry out for the current experiment.

Conclusions

This section presents a musical source separation approach specifically adapted to isolate percussion sources. It combines transient analysis with a spectrum decomposition based on a harmonic model. We show that the combination of these two strategies improves the separation quality.

In contrast to other state-of-the-art methods, this method features fast separation from a single audio frame, which makes this approach suitable for low-latency situations. The quantitative evaluation shows that it obtains performance very similar to offline methods meaning that quality is not lost due to the low-latency processing constraint. Nevertheless, the method still presents some limitations. In the presence of vocals in the mix, the separated percussion source contains residues of fricative phonemes. Also the separated percussion decay loses fidelity. We think that in both cases the

12. Audio examples are available online: http://www.mtg.upf.edu/~jjaner/presentations/icassp12.

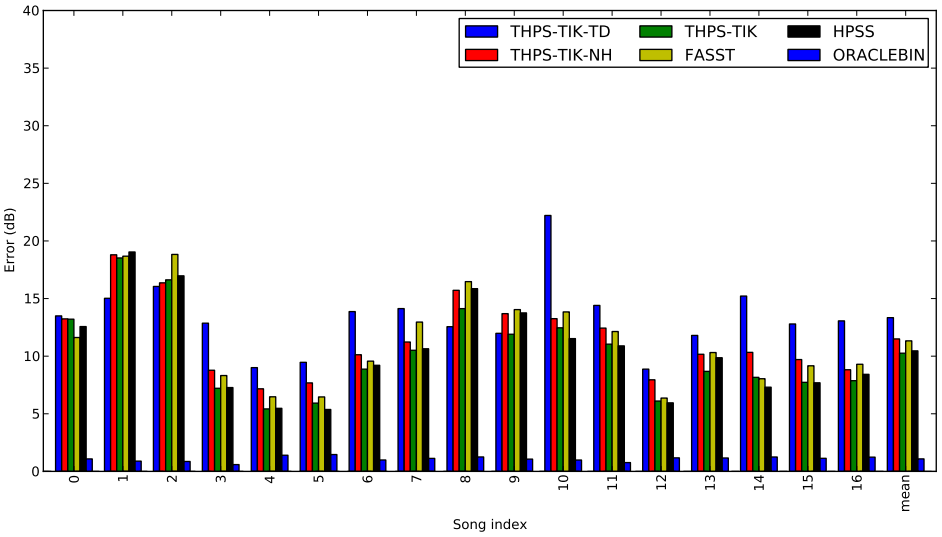

Figure 5.8: SDR error measures of individual audio examples for all the methods of low latency drums separation.

quality can be improved by including statistical modeling of these particular signals.

5.5 Bass Line Separation

In this section we explore the use of Tikhonov regularization spectrum decomposition for the task of bass separation under low-latency constraints. We test whether Tikhonov regularization is a valid alternative to NMF for spectrum decomposition when using fixed harmonic basis. Our experiment compares the separation performance of this method to a naive low-pass filter, a state-of-the-art NMF-based method and a near-optimal binary mask. The proposed low-latency method achieves results similar to the NMF-based high-latency approach at a lower computational cost. Therefore the method is valid for real-time implementations.

Introduction

In the rhythm section of popular western music, the bass line often fulfills the role of anchoring the harmonic framework and laying down the beat. The sound produced by the bass is predominantly harmonic with a

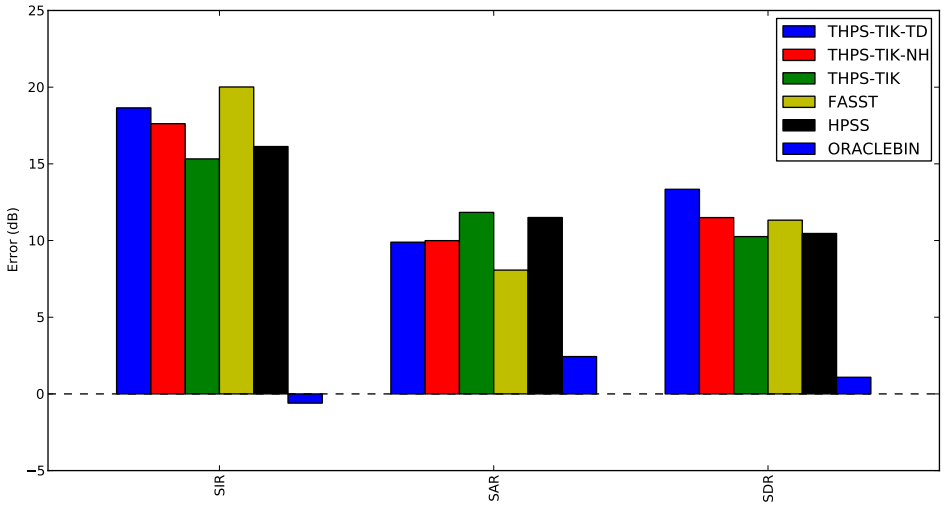

Figure 5.9: Average of all error measures for all the methods.

low fundamental frequency and usually has an impulsive excitation. Bass line estimation is a relevant case in musical source separation, since it can improve the separation of drums or the predominant melody from the mix.

In comparison to other instruments, bass line separation is difficult due to the low frequency and the presence of the bass drum that shares a similar spectrum distribution. FitzGerald et al. (2005) demonstrated the use of Non-negative Tensor Factorization for isolating the bass guitar from other instruments in multichannel synthetic mixtures. Ozerov et al. (2010) employed their general source separation framework for the isolation of the bass line in professionally recorded music.

Nowadays, with increasing availability of music in online streaming services, it is often necessary to process audio data as it is received by the system. And with the increase of embedded devices in our everyday lives, limiting memory requirements is often important. These factors motivate the development of low-latency methods. We propose an extension to the method presented in Section 5.4 with modifications to the signal model in order to better represent bass line components. An evaluation is conducted where the proposed method is compared to a baseline naive method and to FASST, a state-of-the-art high latency and computationally expensive method.

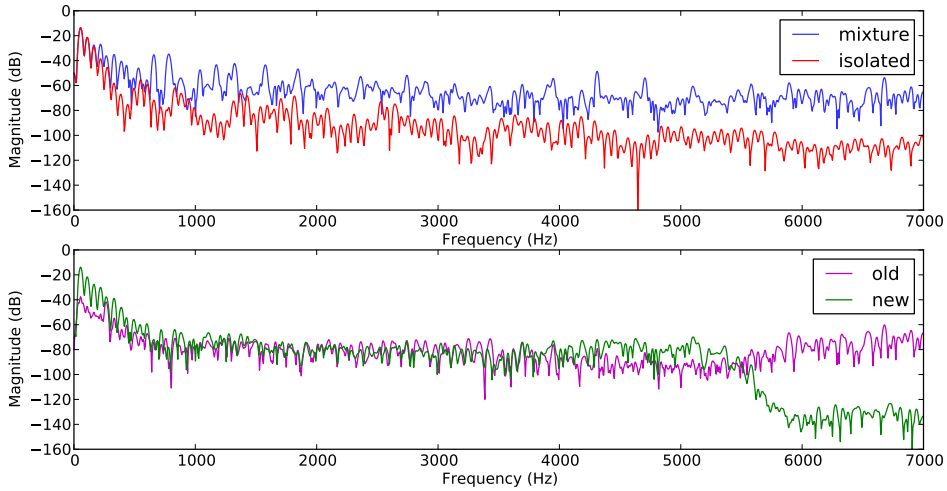

Figure 5.10: Example of a spectrum of the bass in a mixture and the bass in isolation *(top)*. Separated bass using the old signal model presented in Section 5.3 and with the new proposed signal model *(bottom)*.

Method

In Section 5.4 we introduced a low-latency drum separation method based on harmonic decomposition using single-frame Tikhonov regularization. In Section 5.3 we presented a signal model that contains spectrum patterns to represent both wideband and narrowband pitched components.

The bass guitar is mainly a pitched instrument, and at first sight the narrowband components in the signal model would seem to be sufficient. However bass drums quite often present a narrowband spectrum with a resonance of high magnitude and low frequency similar to that of the first partial of the bass guitar. Due to the low pitch of the bass guitar and the limited size of the analysis window of the STFT, the partials in its spectrum are often very close together (see Figure 5.10). This leads to a harmonic comb with less contrast. These components are similar to certain wideband components such as drums or sustained background noise. This causes problems, especially in the high frequency range, where the bass spectrum has very low energy. To solve this issue, the signal model contains specific basis components for non-harmonic wideband spectra and the bass guitar components are constrained to represent their specific timbre.

Bass Specific Signal Model

We employ the same signal model used in Section 5.3, in which pitched sources are modeled as various components of band-filtered harmonic oscillators. Non-pitched sources are incorporated into the model as wideband noise components. The main modification to the signal model is to account for the usual spectral shape characteristics of the bass guitar and bass line in western music. The lowest note in a bass guitar is E1 (41.20Hz) and usually the pitch rarely goes higher than 120Hz. The harmonic envelope of the bass guitar is mainly restricted to the frequency range from 0Hz to 5000Hz.

To achieve this behavior in our signal model, the pitch components that would correspond to the bass are limited in frequency by setting the magnitude of high frequency partials to zero. Using the same notation as in Section 5.3 we can redefine the source-filter model of the pitched components of the basis matrix by adding a function $a[l,\omega]$ that serves as an excitation envelope:

$$\begin{aligned} \vartheta[l,n] &= f_{low} H N_L \frac{2^{\frac{iH-N_T/2+n}{HN_L}} - 1}{S_r \ln(2)} \\ E_l[\omega] &= \mathcal{F}\left\{ \sum_{h=1}^{N_h} a[l, hf_l] sin\left(2\pi h \vartheta[l,n] n\right) \right\} \\ \boldsymbol{W}_{l,i}[\omega] &= \begin{cases} U_i[\omega] E_l[\omega] & \text{if } l \leq N_L \\ U_i[\omega] & \text{if } l = N_L + 1 \end{cases} \end{aligned} \tag{5.11}$$

with $H = (1-\alpha)N_T$. Where α is a coefficient to control the frequency overlap between the components, N_T is the frame size, S_r the sample rate, $\mathcal{F}$ is the Discrete Fourier Transform (DFT), N_h is the number of harmonics of our components, $\boldsymbol{W}_{l,i}$ is the spectrum of the component of l^{th} pitch filtered by the i^{th} filter. U_i is the spectrum of the i^{th} filter in our filterbank. U_i is constructed as a sequence of N_I Hann windows, linearly distributed in the Mel scale and with a 50% overlap. $f_l = \vartheta[l, N_T/2]$ is the center fundamental frequency of the l^{th} pitch. $\vartheta[l,n]$ is the instantaneous frequency function of the l^{th} pitch component.

In order to restrict the use of bass pitched components during the decomposition, we force their excitation envelope to a function decreasing to zero after a given cutoff frequency f_{cut}. The bass pitched components are defined

as those having a fundamental frequency lower than f_{0bass}:

$$a[l,\omega] \quad = \quad \begin{cases} r(\omega)/\omega & \text{if } f_l \leq f_{0bass} \\ 1/\omega & \text{else} \end{cases} \tag{5.12}$$

where $r(\omega)$ is a function of ones that ramps down linearly to 0 from f^s_{cut} to f^e_{cut}:

$$r(\omega) \quad = \quad \begin{cases} 1 & \text{if } \omega \leq f^s_{cut} \\ 1 - \frac{\omega - f^s_{cut}}{f^e_{cut} - f^s_{cut}} & \text{if } f^s_{cut} < \omega \leq f^e_{cut} \\ 0 & \text{if } \omega > f^e_{cut} \end{cases} \tag{5.13}$$

In our experiments we fix the size of the ramp and only control the start frequency $f^s_{cut} = f_{cut}$ and $f^e_{cut} = 1.3 f_{cut}$.

Bass Source Estimation

Using Tikhonov regularization as in Section 5.3 with the modified signal model, we can derive the pitch likelihood L from the gains vector $\hat{\boldsymbol{H}}^{TR}$. The next step is the selection of the components belonging to the bass line.

Instead of using a pitch tracking algorithm as in Section 4.3 that would add complexity and latency to the method, we rely here on a simple peak detection and picking algorithm. The proposed method is simple and has a low computational cost. In order to select the pitch of the bass line, at every frame we select the highest peak in the pitch likelihood function under a certain frequency value f_{0bass}. We assume that only one pitched source will be present in this low frequency range, and that this source will be the targeted bass guitar or bass line.

The peak picking is performed by selecting local maxima in the pitch likelihood $\boldsymbol{L}$:

$$\omega_i \in \left\{ \omega | \underset{j=\omega - W_\omega \ldots \omega + W_\omega}{\arg\max} \boldsymbol{L}(j) \text{ and } \boldsymbol{L}(\omega - 1) < \boldsymbol{L}(\omega) \geq \boldsymbol{L}(\omega + 1) \right\} \tag{5.14}$$

where $2W_\omega$ is the size of the local neighborhood for the peak local maximum.

However as we previously explained, the basis components of the bass described in Section 5.5 are similar to those of wideband components such as the bass drum or other background sources present in the low frequency range. This leads to pitch likelihood functions with a high energy distribution in the low pitch components that do not necessarily correspond to pitched instruments.

As presented in Section 4.4, pitched sources in the spectrum can be modeled as Gaussians in the pitch likelihood function $\boldsymbol{L}$. The width of the Gaussian is related to the chirp ratio of the fundamental frequency of the pitch. If the source is wideband and not pitched (e.g. drums), it can be regarded as a limit case of the partials widening and forming a smooth spectrum with no harmonic structure. Empirical observations show that wideband non-pitched sources that are not decomposed into the wideband components of our signal model, appear as wide noisy Gaussians in the pitch likelihood function.

To distinguish between pitch likelihood peaks corresponding to a pitched bass and those related to other wideband sources, for each pitch likelihood peak p we define a measure of peak contrast c_p. The peak contrast feature is computed using the difference between the height of the peak and the heights of the local minima around it:

$$c_p = \max\Big(\boldsymbol{L}(\omega_p) - \boldsymbol{L}(\omega_p^l), \boldsymbol{L}(\omega_p) - \boldsymbol{L}(\omega_p^r)\Big) \tag{5.15}$$

where ω_p is the position of the p^{th} peak, ω_p^l is the first local minimum under ω_p and ω_p^r is the first local minimum over ω_p.

The bass component in the pitch likelihood ω_b is defined as the position of the highest peak, with frequency under f_{0bass} and whose contrast is over a given threshold $\boldsymbol{L}_{th}$.

As in Section 5.4 we create a new vector $\hat{\boldsymbol{H}}_b$ containing non-zero values only at those bins corresponding to the selected bass pitch.

$$\hat{\boldsymbol{H}}_b[\omega] = \begin{cases} \hat{\boldsymbol{H}}[\omega] & \text{if } |\omega_b - \omega| < \Delta\omega \\ 0 & \text{otherwise.} \end{cases}$$

where $\Delta\omega$ controls the amount of selected pitch components around ω_b. Therefore, we can compute the bass signal estimation as $|\hat{\boldsymbol{X}}_b| = \boldsymbol{W}\hat{\boldsymbol{H}}_b$. In a complementary fashion, the reconstruction of the non-harmonic part takes a gains vector $\hat{\boldsymbol{H}}_{nb}$ containing non-zero values for the unselected bass pitch plus the wideband filter banks. The non-harmonic source estimation is computed as $|\hat{\boldsymbol{X}}_{nb}| = \boldsymbol{W}\hat{\boldsymbol{H}}_{nb}$.

With the estimated magnitude spectra $|\hat{\boldsymbol{X}}_b|$ and $|\hat{\boldsymbol{X}}_{nb}|$ we perform a Wiener filtering to obtain the mask that isolates the bass component:

$$\boldsymbol{m}_b = \frac{|\hat{\boldsymbol{X}}_b|^2}{|\hat{\boldsymbol{X}}_b|^2 + |\hat{\boldsymbol{X}}_{nb}|^2} \tag{5.16}$$

Finally the estimated bass spectrum is simply the result of multiplying the input complex spectrum with the previously presented mask $\hat{\boldsymbol{X}}_b = \boldsymbol{m}_b \otimes \hat{\boldsymbol{V}}$. The output time-domain signal is recovered by means of the inverse STFT and an overlap-add process.

Evaluation

We employ the evaluation techniques used in community evaluation campaigns such as SiSEC (Vincent et al., 2006) to measure the performance of the proposed method. We compute the following measures using the BSSEval toolbox: SDR (Signal to Distortion Ratios), SIR (Source to Interference Ratios) and SAR (Sources to Artifacts Ratios). Evaluation material consists of a dataset of 12 multi-track recordings containing bass guitar or a bass line compiled from publicly available resources (MASS [13], SiSEC [14]) and two in-house professional recordings. The audios were downmixed to mono to avoid using pan information in the separation, since that is out of the scope of this work. The sampling rate of the audio examples is 44.1 kHz, and the spectral analysis uses a frame size of 4096 and a hop-size of 512 samples.

The proposed method, Tikhonov Regularization Bass Separation (TRBS), is compared to several existing techniques. A low frequency filter (LOWP) is used as a baseline trivial method. The publicly available implementation of FASST [15] (Ozerov et al., 2010) serves as a state-of-the-art high-latency option. Finally an oracle separation (Vincent et al., 2007a) using a binary mask is tested as a glass ceiling for spectral bin classification techniques (see Section 5.2). We compared each method to a reference obtained with the soft mask oracle separation. All values presented are error measures: the difference between the soft mask oracle estimation measure and the measure of each algorithm. Thus, the lower the value the closer it is to the oracle estimator meaning better quality.

In a first experiment we perform a parameter exploration for the LOWP and TRBS methods. For the low pass filter we studied the effect of the cutoff frequency. For the TRBS method we studied the effect of varying parameter f_{0bass} that controls the threshold under which a pitch may be considered as belonging to the bass. A second experiment consisted of a

13. http://www.mtg.upf.edu/static/mass
14. http://sisec.wiki.irisa.fr/
15. http://bass-db.gforge.inria.fr/fasst/

comparative study of all the selected methods, where the best parameters for the LOWP and TRBS methods were used.

Results

In Figure 5.13 we observe that the artifacts error (SAR) of the LOWP method is very low. This is expected because the low pass filter does not add new components such as musical noise. The frequency response of the LOWP method is very smooth compared to all other methods, including the oracle soft mask that is used as reference, which explains the negative value of this error measure. We also see that the interference error (SIR) of the LOWP method is high. This method is used as a trivial baseline, and in fact it does not target the bass guitar source, it simply separates low frequency components without making any discrimination. Another observation to be made is that the cutoff frequency parameter controls the tradeoff between artifacts and interferences. A local minimum of distortion error (SDR) is found at 250Hz, even though the average error continues to descend to 75Hz. The results of the individual excerpts, not presented here, show that the SDRs of some songs increase significantly between 250Hz and 120Hz. For that reason 250Hz was chosen as the optimal parameter value.

	SAR	SDR	SIR
60	-17.53	16.68	15.30
120	-28.46	17.30	14.75
250	-32.14	17.17	15.77
500	-37.88	17.47	16.95
1000	-40.54	18.05	17.86
2000	-41.43	18.48	18.41
5000	-42.29	18.76	18.76

Table 5.3: Average error measures for various values of the cutoff frequency parameter (in Hz) of the LOWP method.

Table 5.4 and Figure 5.12 show best performance for both artifacts errors (SAR) and interference errors (SIR) when the f_{0bass} parameter is around 100Hz. Therefore 100Hz was selected as the best parameter value for f_{0bass} for the comparative study. We see a significant difference in the errors (≈ 3 dB) depending on the parameter value. This leads us to think that a more precise pitch selection method could further reduce separation error.

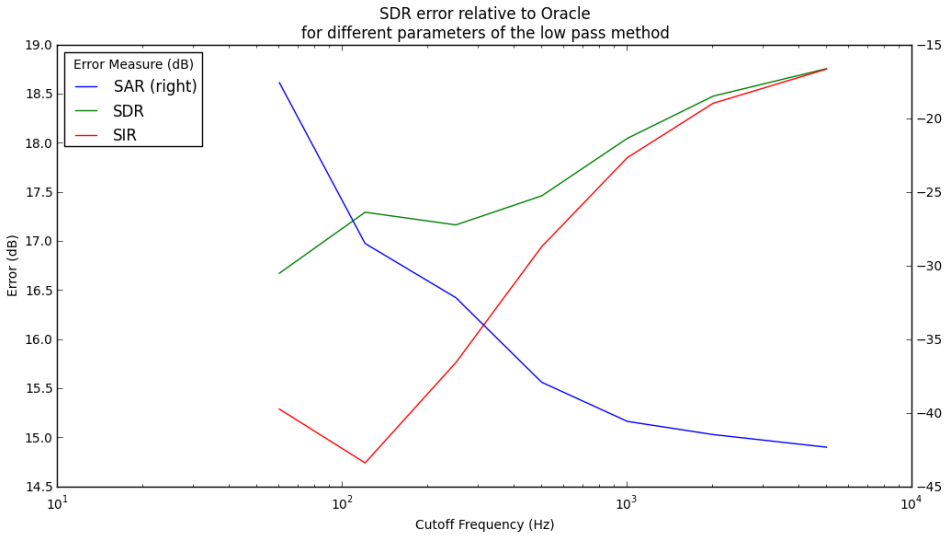

Figure 5.11: Average error measures for various values of the cutoff frequency parameter (in Hz) of the LOWP method.

Several conclusions can be drawn from the results of the comparative study. Table 5.5 and Figure 5.13 show that the proposed method performs similarly to state-of-the-art techniques such as FASST. While FASST achieves a lower artifact error (SAR) separation, TRBS has less interference error (SIR). Another observation is that the oracle binary mask scores a slightly negative SIR error measure. This means that on average the binary mask produces less interference than the soft mask oracle. However, this improvement is balanced by the artifacts error (SAR), where the oracle binary mask reveals the highest error level of all methods.

In Figure 5.14 we can see that this behavior is consistent on all the individual excerpts. In listening to the separated sources, we found that these quantitative results seem to correctly reflect the perceived differences between the methods. A web page [16] with audio examples illustrates the results obtained with our method.

Conclusion

We show that the Tikhonov regularization spectrum decomposition method can be used successfully to perform low latency bass guitar/base line sep-

16. http://www.dtic.upf.edu/~rmarxer/dafx13/bass

	SAR	SDR	SIR
60	12.89	13.14	14.03
70	10.41	13.08	13.54
80	8.43	12.87	12.37
90	6.85	13.01	12.41
100	6.13	12.95	12.40
110	6.43	12.96	12.34
120	6.63	13.05	12.50
130	7.22	13.37	13.18
140	7.73	13.46	13.65
160	8.15	13.83	14.13
180	8.54	13.98	14.60

Table 5.4: Average error measures for various values of $f_{0_{bass}}$ parameter (in Hz) of the TRBS method.

	SAR	SIR	SDR
LOWP-250	-32.14	15.77	17.17
FASST	4.35	15.02	14.33
TRBS-100	6.13	12.40	12.95
ORACLEBIN	9.26	-1.98	6.53

Table 5.5: Average error measures for the evaluated algorithms.

aration of western music signals. Furthermore the use of pitch likelihood peak contrast and specific bass timbre models allows us to produce separation results comparable to state of the art high latency methods, such as FASST. Quantitative results show the accuracy of the separation in contrast to baseline trivial methods such as low pass filters. The proposed method was also compared to approximations of the best possible performance of binary masks by using BSS oracle techniques.

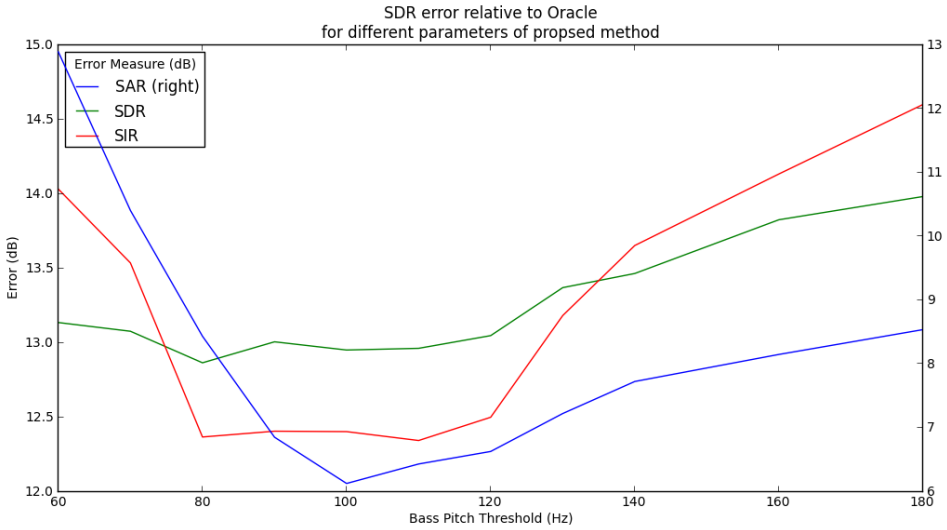

Figure 5.12: Average error measures for various values of f_{0bass} parameter (in Hz) of the TRBS method.

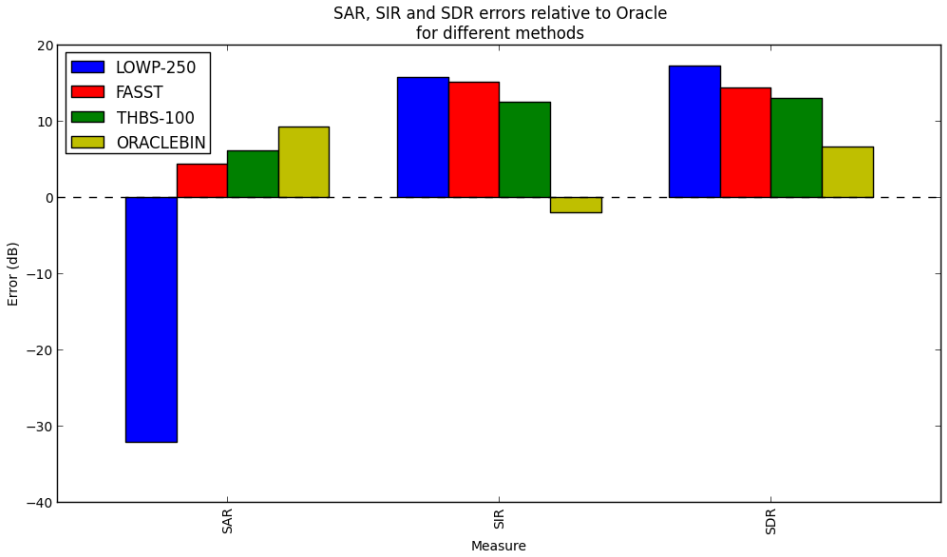

Figure 5.13: Average error measures (x-axis) for the evaluated algorithms. The LOWP-250 method presents very low artifacts (SAR) but a worse global separation (SDR).

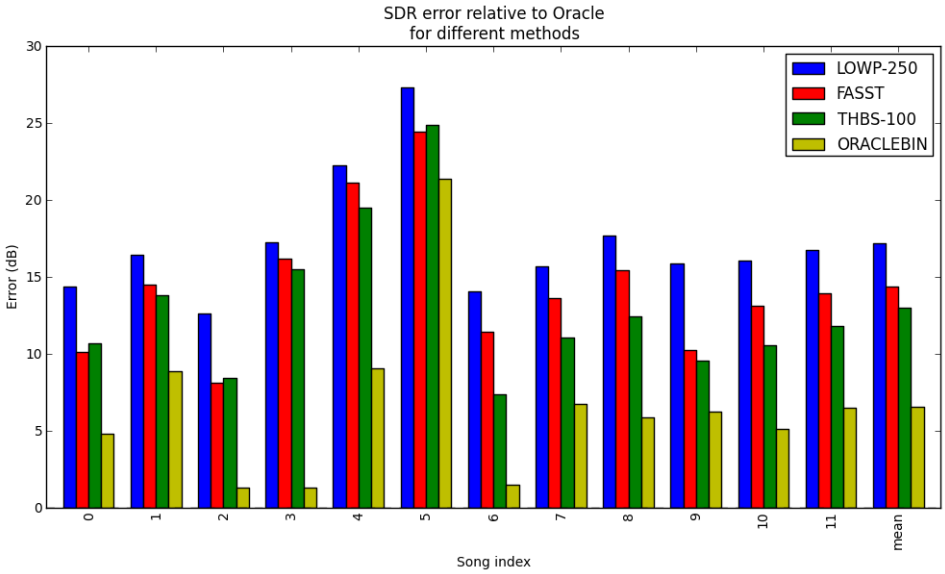

Figure 5.14: SDR error measures of individual audio examples for the methods.

Part III

Chapter 6

High Latency Audio Source Separation

6.1 Introduction

This chapter is dedicated to insights on audio source separation in high latency scenarios acquired during our work on separation under low latency constraints. While developing low-latency pitch estimation and source separation methods some source characteristics were not exploitable or had to be approximated and simplified due to the temporal locality of the process. Some of this knowledge is presented here as proof-of-concept and for comparison. The work shown in this chapter should also serve as a reference glass ceiling for future low latency source separation developments. Furthermore the research presented here is also of interest for future work on low-latency contexts. Many of the techniques proposed can be adapted in the future to low latency situations by performing simplifications and/or approximations.

The structure of this chapter is similar to that of Chapter 5. Two specific musical instruments are considered, the singing voice and drums, due to their presence in western popular music. We also consider multiple-pitch-based separation. Current state-of-the-art high-latency multipitch estimation methods achieve higher accuracy than low-latency methods. Therefore the assumption of accurate multipitch transcription availability of the mixture signal is much more plausible.

Furthermore high-latency scenarios allow the inclusion of manual interven-

tion and user-assisted semiautomatic methods. This motivated us to consider semi-supervised versions of some of the methods presented.

6.2 Comparing Low and High Latency Separation Methods

In this section we first analyze the limitations of performing music source separation under low-latency constraints. We then present a short review of current state-of-the-art high-latency methods with the focus on singing voice and drums separation. We emphasize the Smoothed Instantaneous Mixture Model (SIMM) method due to its flexibility, intuitivity and performance in the community-based source separation evaluation campaigns.

Low-latency Separation Limitations

The low latency constraint imposed in some source separation scenarios implies a series of limitations on the methods. A low latency constraint is defined by a limit on the availability of future signal during the processing. Musical signals usually have large temporal correlations in the form of rhythmic, tonal and timbral structures. A history and/or a future of the signal is usually needed to exploit them. Furthermore, western music is often composed of sections that contain significantly different cues, often the case with the verse and chorus divisions. Methods using only past signal will often fail to take into account changes in sections. The use of prediction-based methods could solve these issues, however low-latency prediction models are still limited and complex, and they are rarely applied to source separation tasks. Regarding the melody line, low-latency constraints often limit the use of future regions of the signal to apply continuity and best path estimation techniques. State-of-the-art high accuracy pitch estimation methods have failed to show good results under low-latency conditions. Perhaps the largest issue with low-latency source separation methods in the case of singing voice, is the temporal limitation in correctly estimating timbre information. Due to the nature of singing voice production, the timbre is far from constant, presenting large variations over time (different phonemes and voice characteristics). Even though in many cases these transitions may be smooth in time, a history of the signal is often necessary for their correct estimation. All the low-latency models that were presented in previous chapters use a set of constant basis components for the decomposition of the mixture. By combining the different basis components several varia-

tions can be approximately reconstructed, however due to insufficient signal history or excessive computational cost, the basis cannot be learned from the data and therefore the approximation will be restricted to the rank of the constant basis set. Another limitation of low-latency methods which we are easily able to overcome in high-latency situations is the inability to add regularizations on the temporal evolutions of the gains. Low-latency methods are also limited in allowing user-assistance because that requires integrating information that is unavailable as the audio frames arrive. Manual intervention in audio processing often requires the user to access the whole audio waveform or at least large regions of it. Certain works on human-assisted source separation (Smaragdis and Mysore, 2009) allow the user to input information in realtime (e.g. by humming), however the separation is performed later as a batch process.

Existing High-latency Techniques

Currently most research on audio source separation for music is performed without taking into consideration latency constraints. Therefore most methods presented in the literature address the problem with relaxed high-latency constraints. Several approaches consider the case of main instrument separation from the accompaniment, typically focusing on singing voice extraction (Virtanen et al., 2008b; Durrieu et al., 2009b; Hsu and Jang, 2010b). Other works address the separation of harmonic and percussion components, which has applications ranging from music transcription to remixing (Helén and Virtanen, 2005; Ono et al., 2008b; Rafii and Pardo, 2011). On the other hand, Ozerov et al. (2010) take a general approach accounting for multiple types of spectra during the signal decomposition. Recent approaches have shifted from purely blind source separation towards incorporating supervised source separation. They focus on obtaining high quality results with the help of additional information such as a musical score (Ewert and Müller, 2012), timbre training (Carabias-Orti et al., 2011; Rodriguez-Serrano et al., 2012) or pitch information manually provided by the user (Smaragdis and Mysore, 2009; Durrieu and Thiran, 2012).

Most high-latency audio source separation methods are based on signal decomposition techniques (see Section 3.4). Here we focus on those of special relevance to our work.

Virtanen (2007) used regularization terms on the basis and gains for the spectral decomposition. He proposed a regularization based on the temporal continuity of the gains and another based on the sparseness in frequency

of the basis. This solution increases the physical interpretability of the estimated basis, allowing better reconstruction of the target sources.

Rafii and Pardo (2011) estimated the repeating background spectrum using beat estimation techniques in order to isolate the lead singing voice. The background spectrum was estimated by averaging spectra taken from beat period positions. This method was extended in Liutkus et al. (2012) with the use of beat tracking methods to account for beat changes during the song.

Virtanen et al. (2008b) also proposed a separation of the lead singing voice by estimation of the accompaniment spectrum using NMF decomposition. The method is characterized by the use of the estimated pitch to mask regions that must not contribute to the NMF decomposition. The proposed method consists in creating a harmonic binary mask to avoid the contribution of bins where vocals are predominant in the multiplicative update rules for the NMF estimation of music accompaniment. This way the bins where the vocals are predominant do not bias the estimation of the musical accompaniment spectra. The estimated musical accompaniment spectra is then subtracted from the original to isolate the targeted pitch line.

Durrieu et al. (2009b) also made use of the estimated pitch to create a binary mask. However in this case the mask is used to enforce a monophonicity constraint. Similarly, as in Virtanen et al. (2008b) the spectrum of the mixture was modeled as the addition of a lead voice and accompaniment. However Durrieu et al. (2009b) went further, decomposing the lead voice component using a source-filter model. In this way the estimation of the lead voice spectra and the accompaniment spectra are performed jointly during the NMF decomposition. Furthermore the source-filter model for the lead singing voice imposes other constraints on the smoothness of the filter and the harmonic structure of the source.

Ozerov et al. (2010) took the spectrum modeling approach further by creating a general set of factors that describe independently the frequency structure and the temporal evolution of the spectral patterns. The music signal was thus decomposed into harmonic sustained, harmonic transient, wideband sustained and wideband transient components. The authors proposed using Generalized Expectation-Maximization (GEM) to solve the decomposition and derived multiplicative update rules inspired by those used in NMF.

The previously presented methods are based on a priori general knowledge of music signals characteristics, such as the smoothness of the singing voice,

timbre and the temporal regularities of the accompaniment. Recent approaches are starting to focus on the use of knowledge specific to the signals analyzed, such as the specific instruments or the score of the song.

Carabias-Orti et al. (2011) and Rodriguez-Serrano et al. (2012) derived spectral models using audio recordings of the instruments in isolation. The training of the models is done by performing NMF decompositions of the instruments' spectra using a source-filter spectrum model. This results in a set of source and filters spectra that can be used to represent the instruments' timbres during the decomposition of the mixture spectra.

Ewert and Müller (2012) also proposed using external information by integrating knowledge originating from the score of the analyzed audio recording into the NMF decomposition of the mixture spectra. Like Virtanen et al. (2008b) and Durrieu et al. (2009b) the fundamental frequencies of the notes in the score are used to restrict the basis to harmonic structures. Additionally the note starts of the scores are used to restrict the gains of a set of wideband basis components that serve to reconstruct note attacks. This method was successfully applied to piano audio recordings.

Finally another type of information used in high-latency methods is that supplied by the user. Smaragdis and Mysore (2009) and Durrieu and Thiran (2012) investigated the use of human-assisted annotations of specific processed recordings to enhance the separation. In both cases, the user assists the estimation of the melody line, which is then used to perform the separation using NMF-based harmonically constrained source separation methods.

Smoothed Instantaneous Mixture Model

The Smoothed Instantaneous Mixture Model (SIMM) introduced by Durrieu et al. (2011) is especially interesting for us because most of the work presented here is an extension of it. We chose SIMM as a base due to its flexibility and simplicity. Furthermore SIMM has proven successful in music source separation tasks in several community evaluation campaigns and can therefore be used as a reference in further work. More recent methods, such as Flexible Audio Source Separation Toolbox (FASST) by Ozerov et al. (2010) have a more flexible and general spectrum model, however they also have increased complexity and higher computational cost.

SIMM is an iterative parameter estimation approach based on NMF exploiting a source/filter model for the predominant instrument. The code

implementing it is available online [1].

This method approximates the mixture magnitude spectrum $\boldsymbol{V}$ as the sum of the lead singing voice and the accompaniment spectra $\hat{\boldsymbol{V}} = \hat{\boldsymbol{X}}_v + \hat{\boldsymbol{X}}_m$. These components are further factorized. The accompaniment is modeled as the non-negative combination of a set of N_{Wm} constant basis components $\hat{\boldsymbol{X}}_m = \boldsymbol{W}_m \boldsymbol{H}_m$. The singing voice spectrum is approximated as an elementwise multiplication ($\otimes$) of a smooth filter and a monophonic harmonic excitation $\hat{\boldsymbol{X}}_v = \boldsymbol{X}_\Phi \otimes \boldsymbol{X}_{f_0}$. The factor corresponding to the filter is modeled as a combination of constant spectral shapes that are smooth in frequency $\boldsymbol{X}_\Phi = \boldsymbol{W}_\Phi \boldsymbol{H}_\Phi$. To ensure smoothness, the spectral shapes $\boldsymbol{W}_\Phi$ are modeled as a non-negative linear combination of band-limited spectra $\boldsymbol{W}_\Phi = \boldsymbol{W}_\Gamma \boldsymbol{H}_\Gamma$. The monophonicity of the excitation is achieved by modeling it as a non-negative combination of harmonic spectral templates $\boldsymbol{X}_{f_0} = \boldsymbol{W}_{f_0} \boldsymbol{H}_{f_0}$, where all the gains $\boldsymbol{H}_{f_0}$, except a region limited in frequency around the estimated predominant pitch f_0, are set to 0. In this work we use the low-latency method presented in Section 4.3 to estimate the source's pitch f_0. To sum up, the observed mixture spectrum $\boldsymbol{V}$ is approximated by the spectrum model $\hat{\boldsymbol{V}}$ in the following way:

$$\begin{aligned} \hat{\boldsymbol{V}}_{SIMM} &= (\boldsymbol{W}_\Gamma \boldsymbol{H}_\Gamma \boldsymbol{H}_\Phi) \otimes (\boldsymbol{W}_{f_0} \boldsymbol{H}_{f_0}) \\ &\quad + \boldsymbol{W}_M \boldsymbol{H}_M \end{aligned} \tag{6.1}$$

where $\otimes$ is the Hadamard product (an elementwise multiplication of the matrices). $\boldsymbol{W}_\Gamma$ and $\boldsymbol{W}_{f_0}$ are fixed matrices and the rest are estimated from the data. $\boldsymbol{W}_{f_0}$ is composed of harmonic spectra with a magnitude decay computed using the Klatt glottal model. $\boldsymbol{W}_\Gamma$ is a set of band-limited filters, modeled with Gaussians centered at frequencies distributed uniformly on the spectrum. The author derived a set of multiplicative update rules for the other components which are detailed in Section 3.4.

Current Limitations

While high-latency source separation methods address some of the shortcomings of low-latency techniques, limitations still remain with current state-of-the-art approaches. In singing voice separation it is well known that most current source separation techniques focus on the harmonic or voiced part of the signal, rarely addressing other components such as unvoiced consonants, breathiness, growl and other noises or wideband sounds

1. http://durrieu.ch/phd/software.html (last accessed on January 3, 2011)

that accompany the voice. These components of the singing voice are very noticeable to humans due to the information they carry and to correlations with the voiced components. Because of this their presence or absence is often perceived in the source separation results even when their energy is relatively low. Another important shortcoming of current high-latency methods is that user assistance has just begun to be taken into consideration. Currently most methods concentrate on manual correction and selection of pitch tracks (Smaragdis and Mysore, 2009; Durrieu and Thiran, 2012) or on the use of scores to achieve this (Ewert and Müller, 2012; Bosch et al., 2012). However using other sources of information such as the instants of the drums or the polyphonies present in a recording have rarely been studied. Ewert and Müller (2012) showed the potential of using similar information such as note starts and fundamental frequencies, however only in a very limited scenario with a single instrument and style of music.

In the following sections we present a series of experiments and methods that explore these limitations and propose solutions for some of them.

6.3 Singing Voice with Breathiness

Most current source separation methods only target the voiced component of the singing voice, failing to remove the breathiness. The remaining breathiness is very noticeable to humans and it retains much of the phonetic and timbral information from the singer. Breathy voice is a phonation mode in which the vocal cords vibrate as they do in normal (modal) voicing, but are held further apart, so that a larger volume of air escapes between them producing an audible noise. We propose a method for estimating the spectrum of the breathiness component and taking it into account when isolating the singing voice source from the mixture. The breathiness component is derived from the detected harmonic envelope in pitched vocal sounds. The separation of the voiced components is used in conjunction with an existing iterative approach based on spectrum factorization. Finally, we conduct an evaluation that demonstrates the separation improvement.

Introduction

Breathiness is an aspect of voice quality that is difficult to estimate or analyze due to its stochastic nature and wideband spectral characteristics. In western music mixture signals, this component often overlaps with other wideband components such as drums or transients. To our knowledge there

are no music source separation methods that have focused on this component of the singing voice. However, in the field of speech analysis and synthesis, the decomposition and manipulation of the breathiness component has been done in a variety of areas such as text-to-speech synthesis, speech encoding, and clinical assessment of disordered voices.

For example, Nordstrom et al. (2005, 2008) studied the relations between the vocal tract and the glottal source in human speech signals. Mehta and Quatieri (2005) focused on the analysis of the breathy component of speech voice and proposed a modulation-based model where the noise component of the voice is modulated by the glottal waveform. This model is used to analyze, synthesize and transform isolated voice recordings. Degottex et al. (2011) proposed an extension to the source-filter model that takes into account turbulence at the glottal level and the radiation at the lips and nostrils level. Their model, Separation of the Vocal-tract with the Liljencrants-fant model plus Noise (SVLN), shows benefits in pitch transformation and breathiness control tasks for singing voice synthesis.

All of this work focused on voice signals in isolation and did not consider either the source separation problem or the analysis of mixture signals.

Proposed Estimation Method

Our method can be integrated into any source separation approach that approximates the mixture spectrum as the sum of the lead singing voice spectrum and the accompaniment spectrum $\hat{\boldsymbol{V}} = \hat{\boldsymbol{X}}_v + \hat{\boldsymbol{X}}_m$. It is appropriate for both low-latency and high-latency situations since it only requires a single audio frame.

The estimation of the breathiness component is based on the approximation of a pitched voice spectrum (with pitch f_0) as a filtered composition of two additive components: a glottal excitation $\boldsymbol{X}_v$ and a wideband component (due to the glottal air flow) $\boldsymbol{X}_{vr}$, both filtered by the vocal tract. The magnitude of the voice spectrum can be expressed in the following manner (Degottex et al., 2011):

$$\begin{aligned} \boldsymbol{X}_{v'}[\omega] &= \boldsymbol{X}_v[\omega] + \boldsymbol{X}_{vr}[\omega] && (6.2)\\ &= \boldsymbol{L}[\omega]\boldsymbol{U}[\omega]\boldsymbol{S}[\omega]\boldsymbol{H}[\omega] + \boldsymbol{L}[\omega]\boldsymbol{U}[\omega]\gamma && (6.3)\\ &= \boldsymbol{L}[\omega]\boldsymbol{U}[\omega](\boldsymbol{S}[\omega]\boldsymbol{H}[\omega] + \gamma) && (6.4) \end{aligned}$$

where $\boldsymbol{S}[\omega]\boldsymbol{H}[\omega]$ is the spectrum of the excitation, $\boldsymbol{S}[\omega]$ is the excitation envelope, $\boldsymbol{H}[\omega]$ is a harmonic comb of unity magnitude, $\boldsymbol{L}[\omega]\boldsymbol{U}[\omega]\gamma$ is the

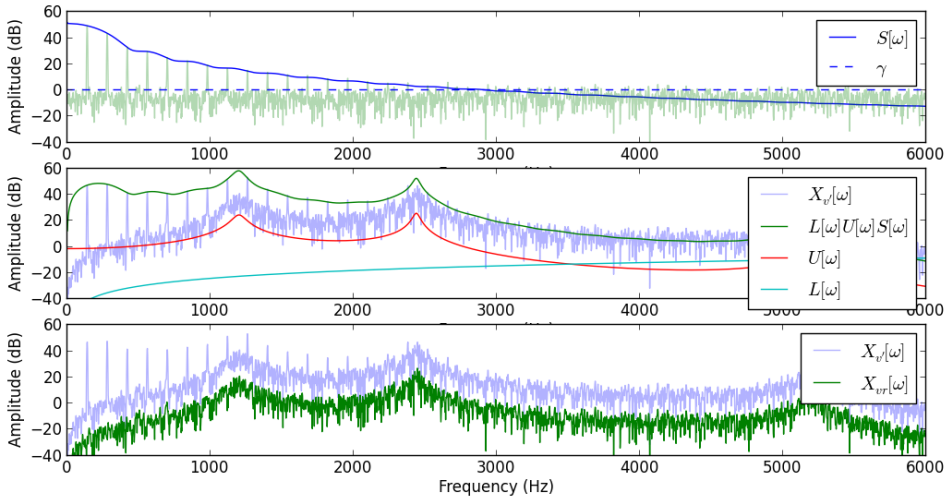

Figure 6.1: Representation of the different components of the singing voice model given a synthetic spectrum.

magnitude spectrum of the breathiness, $\boldsymbol{U}[\omega]$ is the magnitude of the frequency response of the vocal tract filter, γ is the gain of the breathiness spectrum relative to the pitched component, and $\boldsymbol{L}[\omega]$ is the component due to lips and nostrils radiation. Here we approximate the wideband component as a constant spectrum filtered by the vocal tract. This is equivalent to modeling the glottal air flow as white noise, which is realistic especially for a mid-range frequency region.

The human voice excitation envelope can be modeled, as proposed in Klatt and Klatt (1990), using a linear decay in the decibel/octave scale:

$$\boldsymbol{S}[\omega] = C \cdot \omega^{m/20\log_{10}(2)} \tag{6.5}$$

where C is a scaling factor, ω is the frequency in Hz, and m is the slope of the excitation envelope in decibels per octaves (dB/octave).

In our scenario the vocal source spectrum $\boldsymbol{X}_{v'}$ is unavailable, only the mixture spectrum $\boldsymbol{V}$ is accessible. Therefore we cannot directly estimate the breathiness spectrum $\gamma\boldsymbol{U}[\omega]$ using Equation 6.2. Instead, we exploit the fact that at harmonic positions lf_0 of the singing voice pitch we can consider the vocals spectrum predominant $\boldsymbol{V}[lf_0] \approx \boldsymbol{X}_{v'}[lf_0]$ for all harmonic indices $l > 0$. If we additionally consider the vocal tract filter smooth in frequency, as is done in previous works (Durrieu et al., 2011), we can then use Akima

interpolation (Akima, 1970) between the harmonic positions to estimate the harmonic envelope $e_h[\omega] = \boldsymbol{L}[\omega]\boldsymbol{U}[\omega]\boldsymbol{S}[\omega]$ as done in Section 4.3. By assuming the magnitudes (in the decibel scale) of $\boldsymbol{L}[\omega]\boldsymbol{U}[\omega]$ to be drawn from a Gaussian distribution, we can make an estimation $\hat{\boldsymbol{S}}[\omega]$ using least squares fitting of the model from Eq. 6.5 on the harmonic envelope $e_h[\omega]$. The least squares can be linear if the envelope and frequencies are first translated to logarithmic scales. However this must be done on a limited region $[\omega_{lo}, \omega_{hi}]$ of the spectrum where the vocals source is usually predominant and the estimated $e_h[\omega]$ is reliable. Finally we whiten the harmonic envelope $e_h[\omega]$ using the excitation envelope $\hat{\boldsymbol{S}}[\omega]$ derived from the excitation slope:

$$\boldsymbol{L}[\omega]\hat{\boldsymbol{U}}[\omega] = e_h[\omega]/\hat{\boldsymbol{S}}[\omega] \tag{6.6}$$

The model of Equation 6.5 is only valid for the mid frequency region and the estimation of $\boldsymbol{S}[\omega]$ is based on the region where the harmonics are present and predominant. In order to overcome this limitation, whitening is only performed under ω^w_{hi} and is limited to $\hat{\boldsymbol{S}}[\omega] = \hat{\boldsymbol{S}}[\omega^w_{lo}]$ for $\omega < \omega^w_{lo}$. In our proposed method, γ is a parameter that is not estimated from the data. This parameter controls the gain of the breathiness relative to the harmonic component. In Section 6.3 we explore the effect of this parameter on the separation performance.

Figure 6.2 illustrates the intermediate results of the breathiness estimation on a spectrum of a song that contains pitched singing voice. The breathiness envelope is derived from the spectral envelope sampled at the harmonic partial frequencies.

At this point we have estimated the breathiness component as $\hat{X}_{vr} = \boldsymbol{L}[\omega]\hat{\boldsymbol{U}}[\omega]\gamma$. The next section describes how to use it in conjunction with an existing separation approach to obtain the final isolated singing voice .

Integration into a Separation Approach

We integrate our proposed breathiness method into the SIMM separation approach for the reasons previously cited in Section 6.2.

The reader might observe that the SIMM method already provides an estimation of the smoothed filter as $\boldsymbol{X}_\Phi$. Theoretically this filter should have a spectral shape similar to the estimated breathiness component found in section 6.3. However, the goal of this work was not to extend the SIMM method but to provide a general breathiness estimation valid for various separation approaches, even in low-latency conditions. The separation of

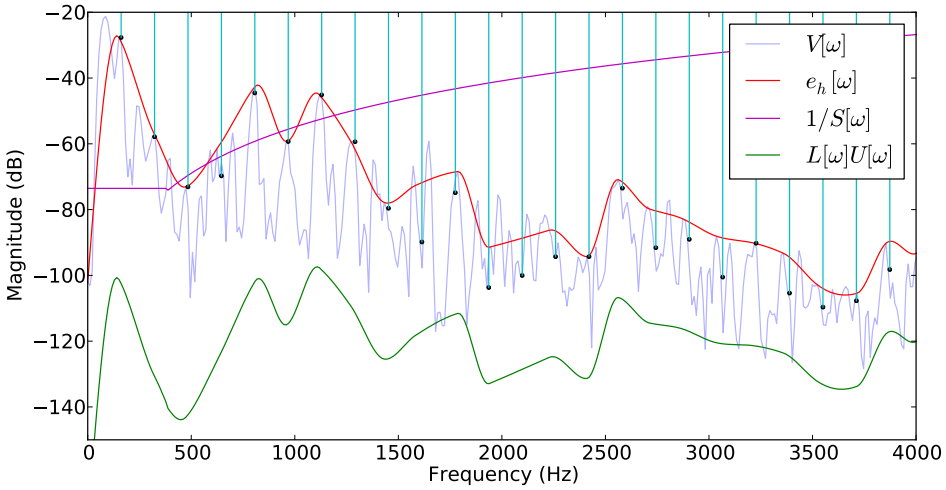

Figure 6.2: Spectrum, harmonic envelope, source-based whitening and the estimated breathiness.

the vocals is done using a Wiener filter as in Durrieu et al. (2011). The estimated breathiness spectrum is added to the estimation of the harmonic part of the voice. Using a notation similar to that used in Section 6.2, the spectrum model becomes $\hat{\boldsymbol{V}} = \hat{\boldsymbol{X}}_{v'} + \hat{\boldsymbol{X}}_m$, which leads to the following mask:

$$\boldsymbol{m}_{v'} = \frac{\hat{\boldsymbol{X}}_{v'}}{\hat{\boldsymbol{X}}_{v'} + \hat{\boldsymbol{X}}_m} \tag{6.7}$$

where $\hat{\boldsymbol{X}}_{v'}[\omega] = \boldsymbol{X}_v[\omega] + \boldsymbol{L}[\omega]\hat{\boldsymbol{U}}[\omega]\gamma$ is the estimated vocal source spectrum and $\hat{\boldsymbol{X}}_v$ is estimated following the procedure described in Section 6.2. The mask is then applied to the mixture complex spectrum to compute the estimated source complex spectrum $\widehat{\widetilde{\boldsymbol{X}}}_{v'} = \boldsymbol{m}_{v'} \otimes \widetilde{\boldsymbol{V}}$. Then a simple overlap-add technique is used to achieve the output waveform signal.

Experiments

We prepared a dataset of multitrack recordings containing singing voice to evaluate the effect of integrating breathiness estimation into the SIMM source separation method. The multiple tracks of each recording were combined forming two sources: the vocals, and the accompaniment music created by mixing all the other tracks.

The evaluation material consists of a dataset of 14 multitrack pop-rock recordings with vocals, compiled from publicly available resources (MASS [2], SiSEC [3], BSS Oracle [4]) and 2 in-house multitrack recordings.

A quantitative evaluation was done using the BSSEval toolbox to calculate the SDR (Signal-to-Distortion Ratio), SIR (Signal-to-Interference Ratio) and SAR (Signal-to-Artifacts Ratio) values. On a first inspection of the results, we saw that the objective measures did not reflect the perceived differences. Additionally we also computed the perceptually motivated objective measures from PEASS: OPS (Overall Perceptual Score), TPS (Target-related Perceptual Score), IPS (Interference-related Perceptual Score), APS (Artifact-related Perceptual Score).

For all the excerpts we also computed the near-optimal time-frequency mask-based separation using the BSS Oracle framework. The evaluation measures of the oracle versions of each excerpt were used as references to reduce the dependence of the results on the difficulty of each audio. Therefore the values shown are error values (lower is better) with respect to the near-optimal version.

In the experiments we set the frequency limits for the excitation slope estimation to $\omega_{lo} = 200$Hz and $\omega_{hi} = 4000$Hz. The whitening limits were set to $\omega^w_{lo} = 400$Hz and $\omega^w_{hi} = 15000$Hz. Audio examples have a sampling rate of 44.1kHz, and the spectral analysis used a frame size of 4096 without zero-padding and a hop-size of 512 samples respectively.

Discussion

In an informal listening test we noticed that in the samples where the vocals are predominant over the background music our approach achieved its objective of maintaining the breathiness in the isolated voice. The downside, however, is that in some cases a dynamic low pass filtering is applied, which reduces the brightness of drums and cymbals in the mute version. In examples where the vocals are fast and the background is loud with relation to the vocals, the breathiness removal is less noticeable.

Looking at the objective quantitative results (not shown here), the BSSEval evaluation results show very little variation (< 0.2dB) for the different values of γ. However this does not reflect the perceived differences in the informal

2. http://www.mtg.upf.edu/static/mass
3. http://sisec.wiki.irisa.fr/
4. http://bass-db.gforge.inria.fr/bss_oracle/

listening procedure. This is probably due to the fact that the differences are in frequency bands with low energy, such as the regions between the partials.

	APS	IPS	OPS	TPS
0.0	64.03	7.53	28.89	60.32
0.1	62.66	8.64	28.41	56.83
0.2	62.57	9.31	28.15	55.95
0.3	62.44	9.89	28.29	55.09
0.4	62.61	10.12	28.31	54.69
0.5	62.71	10.61	28.24	54.46
0.6	62.86	11.01	28.17	54.30
0.7	63.04	11.31	28.20	54.03
0.8	63.17	11.27	28.42	53.78
0.9	63.17	11.54	28.42	53.46
1.0	63.32	11.73	28.31	53.09
1.1	63.43	11.81	28.35	52.86
1.2	63.73	11.84	28.33	52.65
1.3	63.82	11.97	28.45	52.50
1.4	63.99	12.14	28.37	52.34
1.5	64.07	12.40	28.42	52.19

Table 6.1: Average error values of PEASS measures for various values of γ.

In the PEASS results (Table 6.1) we observed a larger change in the performance scores, however the differences in scores remained small. This could be due to limitations of the auditory model used in PEASS. Shrivastav and Sapienza (2003) showed the need for special care with voice breathiness quality in objective measures based on perceptual ratings.

In any case these results reflect the conclusions extracted from the informal listening tests rather than the BSSEval results. We see a separation improvement for the OPS, APS and TPS measures, with an optimal parameter value of around $\gamma = 0.2$ for the breathiness estimation gain. While the improvement on the Overall Perceptual-related Score error is small ($\tilde{0}.74$ decrease), the proposed method does perform significantly better with respect to other measures such as APS and TPS.

Figures 6.3 and 6.4 show the trends of the different perceptually-motivated

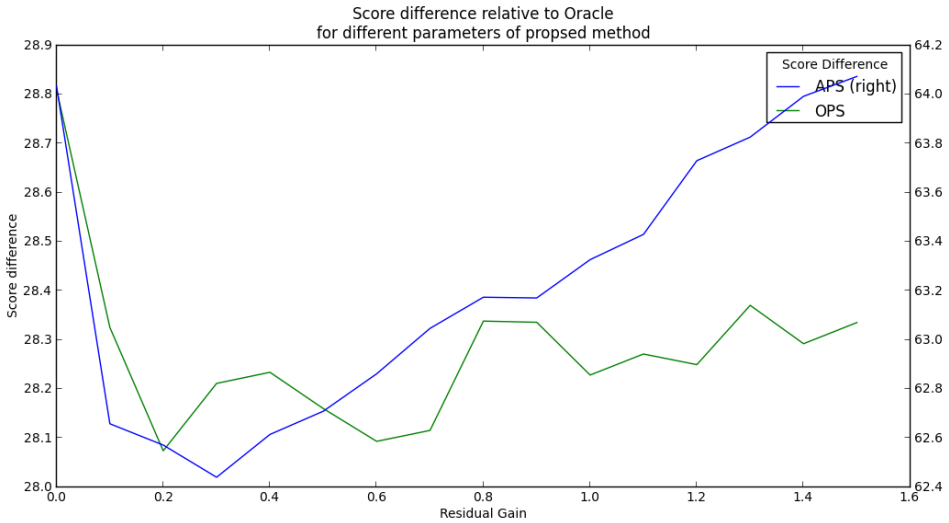

Figure 6.3: PEASS OPS and APS results for different parameters of the breathiness gain γ.

separation performance measures. Figure 6.4 shows that the tradeoff between the interference and the target scores can be controlled with this gain parameter. OPS and APS curves in Figure 6.3 show a global minimum corresponding to the optimal gain for the breathiness estimation, after which the errors slowly increase with γ. The OPS curve has several local minima which could mean that the optimal value of γ depends on the song. From the results of the individual songs in Figure 6.5, we observe that for excerpts 5, 6 and 10 there is a clear improvement in OPS. On the other hand excerpts 2 and 12 show a significant decrease in performance when using the proposed method. Manual inspection of these instances reveal that excerpts 5 and 6 belong to the same song, with a voice containing a high degree of breathiness. Excerpt 2 shows a large number of pitch errors that could explain the large increase in errors. Finally, excerpt 12 presents a vocal track with almost no breathiness component, which would imply, as the results show, a gradual increase in errors with the increase in the parameter γ.

Another observation is that each excerpt presents a point of minimum error at a different value of γ, this shows the desirability of developing methods for estimating the optimal value γ, and thus the strength of the breathiness, from the mixture data. In a practical implementation, we suggest a user-

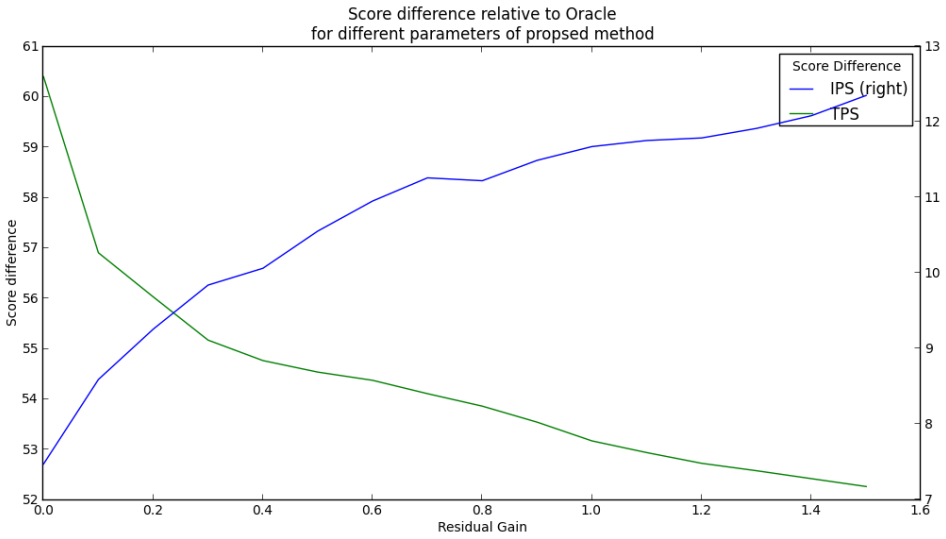

Figure 6.4: PEASS TPS and IPS results for different parameters of the breathiness gain.

controllable parameter γ that can be adapted to the audio content. To demonstrate the subjective improvement in the singing voice separation of our approach, we prepared a web page [5] with several audio examples.

Conclusions

We propose a method to estimate the breathiness component of the singing voice in a professional music mixture. The method extends the source-filter model in a way similar to Degottex et al. (2011). The spectrum model for the source is decomposed into harmonic deterministic narrowband and stochastic wideband components. The harmonic envelope of singing voice and a regression with a Klatt model is used to estimate the spectral shape of the breathiness. The spectral shape is then scaled with a parameter empirically adjusted to control the gain of the breathiness spectrum. An experiment shows that this breathiness estimation method can be used in conjunction with the SIMM method to improve the isolation of the singing voice. Additionally, the parameter exploration of the breathiness shows that estimating the scale of the breathiness from the mixture could further improve the performance of the separation process. Future work could also

5. http://www.dtic.upf.edu/~rmarxer/dafx13/breath

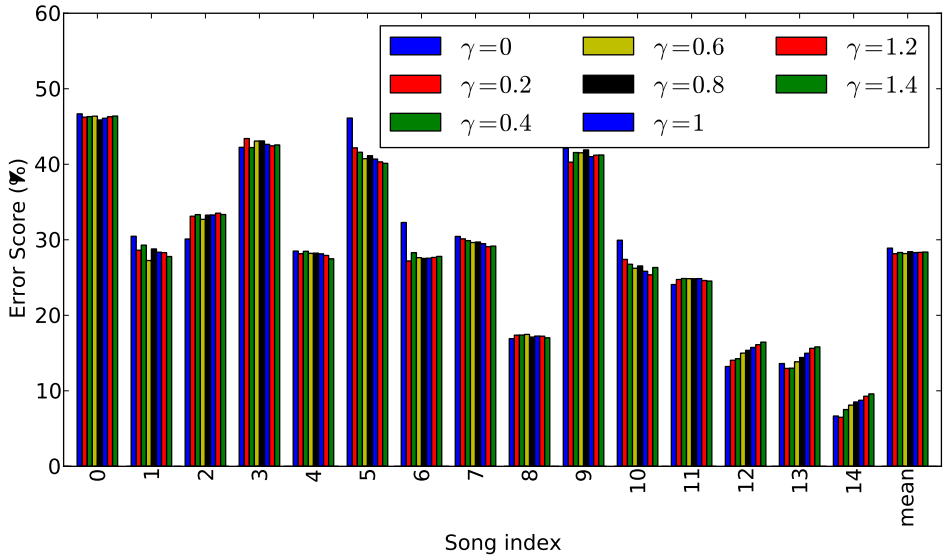

Figure 6.5: PEASS OPS error score (relative to Oracle) for individual songs.

be dedicated to estimating the optimal high-pass filter that models the radiation effect from lips and nostrils as well as the distribution of the glottal turbulence noise, both of which are currently empirically parametrized.

6.4 Singing Voice Fricatives

Separating the singing voice from a musical mixture is a problem widely addressed due to its various applications. However, most approaches do not tackle the separation of unvoiced consonant sounds, which causes a loss of quality in vocal source separation algorithms, and is especially noticeable for unvoiced fricatives (e.g. /s/, /ʃ/, /θ/) due to their energy level and duration. Fricatives are consonants produced by forcing air through a narrow channel made by placing two articulators close together. We propose a method to model and separate unvoiced fricative consonants based on a semi-supervised Non-negative Matrix Factorization, in which a set of spectral basis components are learned from a training excerpt. We implemented this method as an extension of an existing well-known factorization approach for singing voice (SIMM). An objective evaluation shows a small improvement in the separation results. Informal listening tests show a significant increase of intelligibility in the isolated vocals.

Introduction

In the context of musical audio source separation, we do not find many references in the literature addressing the problem of unvoiced phonemes in singing voice. Usually, removing the unvoiced (e.g. fricative) components of the singing voice in a polyphonic mixture is addressed as a joint problem in the signal modeling step. For example, NMF approaches that use harmonic basis sometimes integrate a flat spectrum component to capture the unvoiced parts of the lead vocals (Durrieu et al., 2011). While not specifically addressing singing voice separation, the technique by Wong et al. (2007) performs spectral subtraction to obtain the enhanced vocal signal. Then a multilayer perceptron (MLP) is used to segregate the vocal from the non-vocal segments taking as input the spectral flux, the Harmonic Coefficient (HC), the Zero Crossing Rate (ZCR), the Mel-frequency Cepstral Coefficients (MFCCs), the amplitude level and the 4Hz modulation energy. Finally, the Dynamic Time Warping (DTW) algorithm is used to align the two sequences.

Hsu and Jang (2010a) specifically addressed the problem of unvoiced singing voice separation. A first step segments the signal into accompaniment, voiced and unvoiced predominant frames by means of a Hidden Markov Model (HMM) using 39 MFCC features computed directly from the STFT (taking energy and the first and second derivative of the cepstral coefficients). A second step uses a Gaussian Mixture Model (GMM) classifier to perform an "Unvoiced-Dominant Time-Frequency (T-F) Unit Identification" within only the unvoiced frames. T-F units are computed by means of a gammatone filter-bank of 128 channels. In the training stage, each T-F unit is labeled as unvoiced-dominant or accompaniment-dominant, depending on the energy ratio in the training mixture examples (during training the source singing voice and accompanying signals are known). The Gaussian mixture model consists of 32 components with a diagonal covariance matrix. This approach seems promising after listening to the results. One drawback of this method is the large number of parameters to learn (39 features x 32 GMM components x 128 channels), which requires a lot of training data. However, in addition to audio examples they do provide a publicly available dataset for training [6].

Recent work has shown that a semi-supervised variation of the NMF can be useful for detecting and modeling individual phonemes in speech. Schmidt and Olsson (2006) approached the speech separation problem using semi-

6. http://sites.google.com/site/unvoicedsoundseparation

supervised sparse NMF. The basis components are previously learned from training data in this technique. The authors showed an improvement in the separation when the basis components learned are phoneme-specific. Raj et al. (2011) proposed a phoneme-dependent, exemplar-based NMF model for speech enhancement of monaural mixes. The authors created a set of basis components for each phoneme by drawing spectral vectors from segments of speech recordings that contained the target phoneme.

Lately NMF constraints have been widely used in music separation tasks. Heittola et al. (2009), Durrieu et al. (2009a) and Hennequin et al. (2010) proposed harmonicity and monophonicity constraints by initializing basis or gain bins to zero where the target source is known to have a low or no energy contribution. Ewert and Müller (2011) used musical scores to apply harmonicity, fundamental frequency and note-start constraints on the basis and gains of an NMF decomposition of the mixture spectrum. Note attacks were modeled using wideband learned basis components with the gains initialized to 0 at all time frames except those where the notes started.

We propose a method to detect and suppress unvoiced fricative consonants of the singing voice in music mixture recordings. The method extends SIMM with semi-supervised NMF and additional constraints on the factors in order to take into account unvoiced fricative consonants during the singing voice separation. This technique serves as a proof of concept and could be extended to other singing voice components as well as unvoiced components of other musical instruments. The method is tested on a dataset of multitrack music recordings and shows an improvement on the objective perceptual-based separation measures.

Proposed Method

We propose an extension to SIMM that approximates the fricative consonants as an additive wideband component to the singing voice spectrum. Using the same notation as the SIMM spectrum model results in $\hat{\boldsymbol{V}} = (\hat{\boldsymbol{X}}_v + \hat{\boldsymbol{X}}_{fric}) + \hat{\boldsymbol{X}}_m$, where $\hat{\boldsymbol{X}}_{fric}$ is interpreted as the spectrum of the fricative consonants of the singing voice, and $\hat{\boldsymbol{X}}_{v'} = \hat{\boldsymbol{X}}_v + \hat{\boldsymbol{X}}_{fric}$ is the full spectrum of the singing voice comprising voiced and fricative components. Similarly to what is done for the accompaniment spectrum $\hat{\boldsymbol{X}}_m$ in SIMM, we use an NMF decomposition to model the fricative spectrum $\hat{\boldsymbol{X}}_{fric} = \boldsymbol{W}_{fric}\boldsymbol{H}_{fric}$. However in this case the $\boldsymbol{W}_{fric}$ are learned during a training stage and set constant during the separation stage.

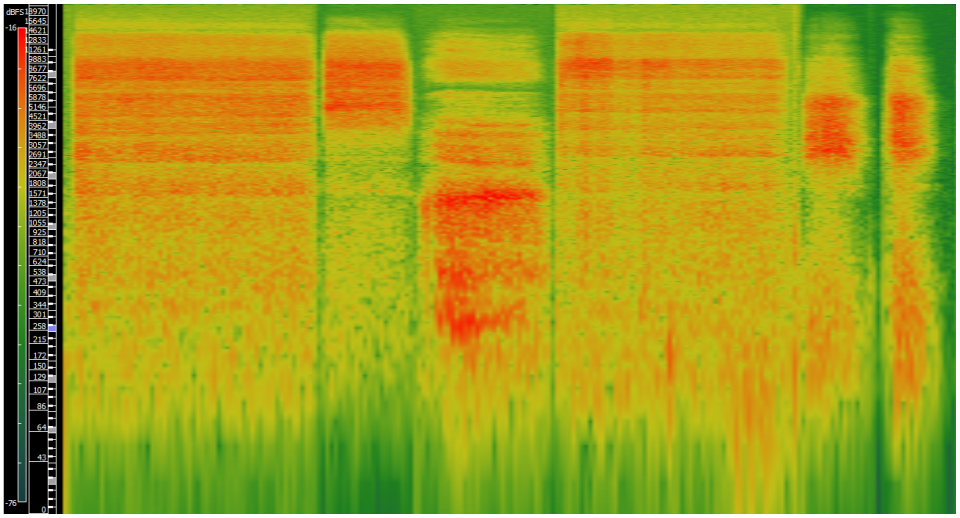

Figure 6.6: Spectrogram of the unvoiced fricative sounds used in the NMF training stage. Frequency axis shown in a logarithmic scale

The proposed method has two steps: 1) training a model of NMF basis; and 2) separating the fricatives with the learned NMF basis. We provide a recording to train the NMF basis which contains only the target sounds. We record a sequence of several unvoiced (voiceless) fricative sounds (/s/, /f/, /ʃ/, /θ/, /h/, /tʃ/) by a single subject using a Shure SM-58 microphone (see Figure 6.6) and apply a low-shelf filter with cutoff frequency at 200Hz to remove the blowing noise (low frequency). We refer to the resulting processed waveform as $x^t_{fric}[t]$.

We then specify the number of basis components $N_{W\,fric}$ to be learned and perform an NMF decomposition of the training audio spectra $\boldsymbol{X}^t_{fric} = \boldsymbol{W}^t_{fric}\boldsymbol{H}^t_{fric}$ into a set of $N_{W\,fric}$ basis components and the corresponding gains $\boldsymbol{H}^t_{fric}$. Both the basis components $\boldsymbol{W}^t_{fric}$ and the gains $\boldsymbol{H}^t_{fric}$ are learned from the data.

The resulting $\boldsymbol{H}^t_{fric}$ can be disregarded, since they are only applicable to the specific training input spectrum $\boldsymbol{X}^t_{fric}$. However we can assume $\boldsymbol{W}^t_{fric}$ to be a good generic basis for general vocal fricative instances.

Since the fricative spectrum is assumed additive and independent from other factors, the multiplicative update rules are trivial to derive. The update rules of all the components except $\boldsymbol{H}_{fric}$ are computed in the same manner as for SIMM (see Section 3.4 Equations 3.91, 3.94, 3.92, 3.93 and 3.95). We

must take into account that in the proposed method the estimation of the mixture spectrum $\hat{\boldsymbol{V}}$ also includes the estimated fricative spectrum $\hat{\boldsymbol{X}}_{fric}$. The multiplicative update rules for the $\boldsymbol{H}_{fric}$ become:

$$\boldsymbol{H}_{fric} \leftarrow \boldsymbol{H}_{fric} \otimes \frac{{\boldsymbol{W}^t_{fric}}^{\top} \left(\hat{\boldsymbol{V}}^{(\beta-2)} \otimes \boldsymbol{V} \right)}{{\boldsymbol{W}^t_{fric}}^{\top} \hat{\boldsymbol{V}}^{(\beta-1)}} \tag{6.8}$$

Applying the multiplicative rules for a given number of iterations, we obtain the estimated gains of the fricatives $\hat{\boldsymbol{H}}_{fric}$.

The separation of the voice is then done by performing a Wiener filter where the target source is composed of the voiced and fricatives spectrum:

$$\boldsymbol{m}_{v'} = \frac{\hat{\boldsymbol{X}}_{v'}}{\hat{\boldsymbol{X}}_{v'} + \hat{\boldsymbol{X}}_{m}} \tag{6.9}$$

where $\hat{\boldsymbol{X}}_{v'}[\omega] = \boldsymbol{X}_{v}[\omega] + \boldsymbol{X}_{fric}[\omega]$ is the estimated vocal source spectrum. The mask is then applied to the complex spectrum of the mixture to compute the estimated source complex spectrum $\widehat{\widetilde{\boldsymbol{X}}}_{v'} = \boldsymbol{m}_{v'} \otimes \hat{\widetilde{\boldsymbol{V}}}$. Then a simple overlap-add technique is used to achieve the output waveform signal.

After initial examination of the results, we realized that the main drawback of this approach for estimating the spectrum of the fricatives is the use of the fricatives basis components to reconstruct other wideband sources such as hi hats, cymbals or even snare drums. This is due to the similarity of the spectra of these sources. The main difference between these two sources is the transient nature of the sounds. Drums generally create sounds with a very fast increase in energy, which are referred to as transient sounds. On the other hand, fricatives are usually more sustained, with a very slow onset and termination.

In order to overcome this problem we propose using the transient quality of the spectrum frames to differentiate between the fricatives and the drums. Using the same transient estimation method presented in Section 5.4 we extract from the audio mixture a set of N_J transient timepoints t^{tr}_j for $j \in [1 \dots N_J]$.

We assume that at transient positions the fricative presence will be negligible compared to other sources such as drums or other attacks. These timepoints are then used as constraints on the gains of the fricatives by

initializing the corresponding columns to zero:

$$\boldsymbol{H}^{T}_{fric}[w,t] = \begin{cases} 0, & \text{if } |t - t_j^{tf}| < \tau \forall j \\ \gamma, & \text{else} \end{cases} \tag{6.10}$$

where $\gamma > 0$ is a random positive value and τ is a parameter that controls the size of the vicinity of the transient timepoint.

Following a similar rationale we assume unvoiced fricatives will not be present at the same instants as the pitched singing voice component. Therefore we can define a different initialization based on the estimated singing voice pitch $p[t]$. By following the same convention as in Section 4.3 where an unvoiced frame is defined by $p[t] <= 0$ we can determine the initialization constraint based on pitch as:

$$\boldsymbol{H}^{P}_{fric}[w,t] = \begin{cases} 0, & \text{if } p[t] > 0 \\ \gamma, & \text{else} \end{cases} \tag{6.11}$$

Finally we propose another initialization constraint that combines the two previous ones:

$$\boldsymbol{H}^{PT}_{fric}[w,t] = \begin{cases} 0, & \text{if } p[t] > 0 \text{ or } |t - t_j^{tf}| < \tau \\ \gamma, & \text{else} \end{cases} \tag{6.12}$$

Each of these initialization constraints will lead to a different factorization and separation result. From now on, we will use the names FRIC-T, FRIC-P and FRIC-PT respectively for these separation methods. The method that does not apply any constraint on $\boldsymbol{H}_{fric}$ will be called FRIC.

Experiment

We hypothesize that in the context of singing voice separation the use of trained basis components with transient and pitch-based gains constraints can improve the estimation of unvoiced fricative consonants with the SIMM model. In the experimental setting we test this hypothesis by evaluating the separation results of the SIMM method with the proposed extensions (FRIC, FRIC-T, FRIC-P and FRIC-PT) and without extensions (NON-FRIC). The evaluation is performed on the same 14-excerpt dataset as in Section 6.3 and using the same performance measures. The results of these tests are shown and discussed in the following section. The main parameter of the proposed method has been set empirically to $\tau = 75$ms for all the tests.

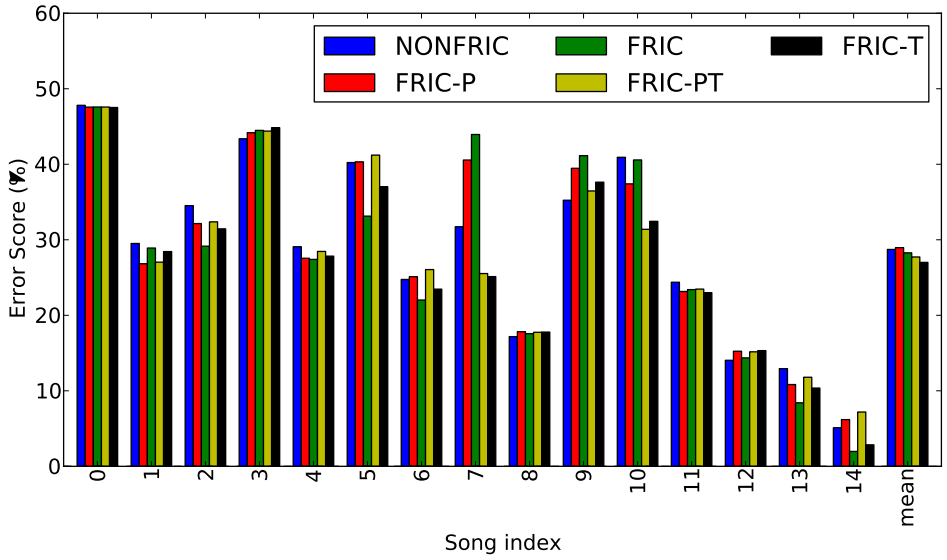

Figure 6.7: OPS error measures of individual audio examples for the vocal fricatives separation methods.

Results

As shown in Figure 6.7, results are better on the Overall Perceptual Score in the singing voice isolation task with most of the fricatives estimation methods. This improvement is present in almost all the excerpts, and in those where the proposed methods do not improve, the penalty on the OPS is relatively small. The FRIC-P method, which uses the pitch as a constraint for the fricative gains is the only method that does not improve over the no fricative estimation (NONFRIC).

In Figure 6.8 and in Table 6.2 we can observe that the overall separation improvement is mainly due to a decrease in interference, and the consequent reduction of the IPS error. We also note that the different constraint methods (FRIC-T, FRIC-P, FRIC-PT) have a large influence on the score errors. The use of pitch-based constraints on the fricative gains degrades significantly the separation performance in terms of TPS and APS. Listening examination of the results show that this is mainly due to false positive pitches at fricative positions. Fricatives are often positioned close to voiced phonemes and the pitch estimation used often extends the resulting pitch tracks to these regions due to large analysis windows. Future work could

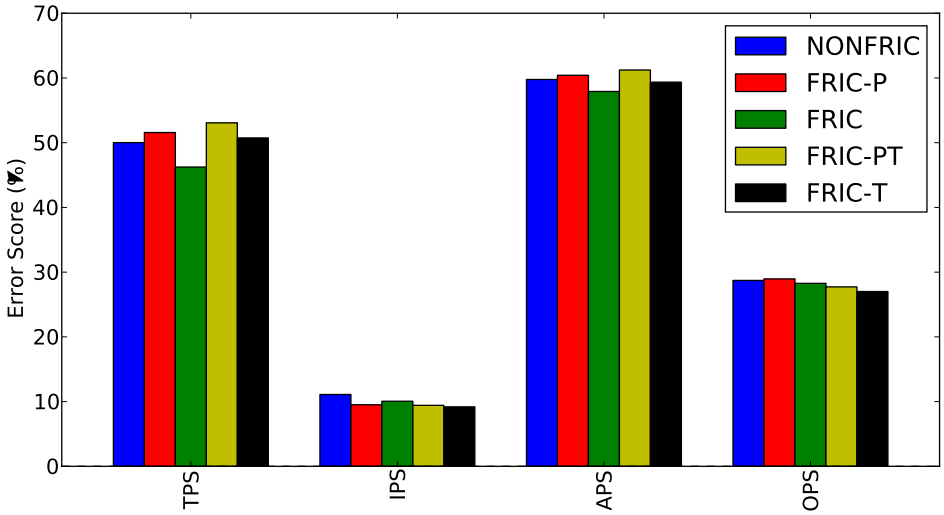

Figure 6.8: Average error measures for the vocal fricatives separation methods.

be devoted to studying the adaptation of the pitch estimation and tracking methods to avoid such situations.

Transient-based constraints (FRIC-T and FRIC-PT) improve the overall separation results in comparison to using no constraints at all (FRIC). The improvement of the FRIC-T method comes in the form of a tradeoff between interference and target/artifact errors, possibly due to transient constraints being binary thereby leading to non-smooth changes in the time-frequency masks. Informal listening to the results shows mainly a reduction of drums interference in the isolated singing voice, which was the desired effect of the constraint.

	TPS	**IPS**	**APS**	**OPS**
NONFRIC	50.01	11.10	59.77	28.72
FRIC-P	51.57	9.51	60.42	28.96
FRIC	46.24	10.06	57.91	28.27
FRIC-PT	53.07	9.42	61.23	27.72
FRIC-T	50.74	9.19	59.35	27.01

Table 6.2: Average error values of PEASS measures for all the fricative estimation methods.

We also conducted preliminary tests using basis components trained on "plosive" and "trill" consonants, however the singing voice separation did not improve. Possibly this was due to the lack of characterization of the temporal evolution of their spectra.

Conclusions

We propose an extension to the SIMM spectrum model and separation technique that takes into account unvoiced fricative consonants when isolating the singing voice. The method uses semi-supervised NMF to train a set of basis components on audio recordings of isolated fricative consonants. The resulting components are then used in the separation stage. Two types of constraints on the factorization were evaluated. Transient analysis was used to distinguish between percussive events and fricatives. Pitch-based constraints were used to restrict the estimation of unvoiced fricatives to regions without pitch. Although the improvement of the objective separation measures is small the perceived difference in informal listening tests is significant. The method is capable of retaining many of the unvoiced fricatives present in the mixture. The transient-based constraints improved the separation by disambiguating between fricatives and drums. However the pitch-based constraints had a negative effect on the separation results, probably due mainly to pitch estimation errors. This research shows the potential of combining semi-supervised NMF with model-based factorization such as SIMM. Future work could focus on non-fricative unvoiced consonants such as "plosives" and "trills" to better understand the limitations of the current spectrum model and factorization technique. The use of constraints could be further studied by adapting the pitch estimation techniques to this particular use-case and by testing the methods on ground truth pitch annotations. The use of regularization could also be an interesting alternative to the constraints, and could reduce the musical artifacts by avoiding the binary masks on the gains matrices.

6.5 Drums Separation

Drums separation is especially interesting for high-latency scenarios because percussive events are highly characterized by their temporal evolution. Current research has often focused on drums separation without specifically modeling the other sources present in the signal. Other approaches perform a joint modeling of the multiple sources present in the signal but being very general and generic they may be too restrictive to achieve low interference

rates. We compare the use of signal-specific information, in our case drum hits positions, combined with specifically modeling other non-target sources such as the singing voice.

Introduction

Drums transcription has been regarded as an important task by the Music Information Retrieval (MIR) community and in the past decade there has been increasing interest in developing techniques for separating the drums track from music mixes. Zils et al. (2002) derived a method based on synthetic drums sound pattern matching. The matching is performed using the correlation as the objective function. Barry (2005) computed the presence of percussive events based on the temporal derivative of the spectral magnitudes on the decibel scale. The separation is then performed by spectral modulation, weighting the spectral bins by the individual bin derivatives previously computed. Ono et al. (2008b) proposed another method based on spectrotemporal features. However in this case both the temporal and frequency derivatives are taken into account. Gillet and Richard (2005) decomposed the signal into a basis of Exponentially Damped Sinusoids (EDS) using a noise subspace projection approach. This leads to a harmonic/noise decomposition that is used to extract the percussive sources. Yoshii et al. (2005a) used a template-based pattern matching technique to estimate and separate the drums spectrum from the rest. The authors showed several applications such as remixing, drum timbre modification and rhythmic sources equalization. Helén and Virtanen (2005) proposed the use of Non-negative Matrix Factorization (NMF) and Support Vector Machine (SVM) classification to perform drum separation. The technique consists in performing an NMF decomposition of the spectrogram of the mixture and classifying the basis components of the factorization using Mel-Frequency Cepstrum Coefficients (MFCC) and an SVM previously trained using isolated drums and harmonic audio recordings. Yoo et al. (2010) proposed a similar approach where Nonnegative Matrix Partial Co-Factorization (NMPCF) is used to avoid training harmonic components. In Ozerov et al. (2010) the authors proposed using the Flexible Audio Source Separation Toolbox (FASST) to perform an isolation of the drums components in a mixture. FASST is based on non-negative factorization of a complex spectrum model that contains templates for specific spectral and temporal patterns which are able to reconstruct harmonic and percussive components when combined.

The use of temporal constraints on NMF is not new and has proven useful

in several scenarios. Ewert and Müller (2011) used score-based temporal restrictions on the gains of an NMF decomposition to estimate piano notes attacks.

Here we address the separation of drums in polyphonic music mixtures, typically containing lead vocals. We integrate our proposed drums separation method into the SIMM separation approach for the reasons previously cited in Section 6.2.

In Section 6.4 we showed the benefits of using temporal constraints on the gains of a SIMM-based spectrum model component to improve the estimation of fricatives.

Proposed Method

We propose an extension to the SIMM method that includes an extra additive spectral component to represent percussive events. The proposed spectrum model can be defined as $\hat{\boldsymbol{V}} = \hat{\boldsymbol{X}}_v + \hat{\boldsymbol{X}}_{m'} + \hat{\boldsymbol{X}}_d$, where the additional component $\hat{\boldsymbol{X}}_d$ corresponds to the estimation of the drums. The lead vocals spectrum $\hat{\boldsymbol{X}}_v$ is decomposed into multiple factors representing a source-filter harmonic model, the other components are decomposed into two factors $\hat{\boldsymbol{X}}_{m'} = \boldsymbol{W}_{m'}\boldsymbol{H}_{m'}$ and $\hat{\boldsymbol{X}}_d = \boldsymbol{W}_d\boldsymbol{H}_d$. It is trivial to show that without any further modifications and with a specific ordering of the multiplicative updates, the proposed spectrum model is equivalent to SIMM with $\boldsymbol{W}_m = [\boldsymbol{W}_{m'}; \boldsymbol{W}_d]$ and $\boldsymbol{H}_m = [\boldsymbol{H}_{m'}, \boldsymbol{H}_d]$.

As in the original SIMM the actual separation is performed by Wiener filtering using the drums spectrum estimation $\hat{\boldsymbol{X}}_d$. Thus the time-frequency mask becomes:

$$\boldsymbol{m}_d = \frac{\hat{\boldsymbol{X}}_d}{\hat{\boldsymbol{X}}_v + \hat{\boldsymbol{X}}_{m'} + \hat{\boldsymbol{X}}_d} \tag{6.13}$$

In the following sections we show different techniques to achieve the differentiation between drums and other musical accompaniment sources in $\hat{\boldsymbol{X}}_d$ and $\hat{\boldsymbol{X}}_{m'}$. First we present a method based on NMF regularizations and then one that uses information specific to the processed signal to apply constraints to the factorization.

Training

We study the use of semi-supervised NMF in conjunction with the SIMM method for the separation of drums sources. In this scenario we first learn

a set of basis components $\boldsymbol{W}^t_{drums}$ using recordings of drums in isolation and then use these components during the separation stage. The learned components will be used as constants and complemented with $N_{W\,d'}$ basis components that will be free and learned during the separation $\boldsymbol{W}_d = [\boldsymbol{W}^t_{drums}; \boldsymbol{W}_{d'}]$.

For the proposed method we trained for two different types of drums: snare drums and cymbals. The result is two sets of learned basis components $\boldsymbol{W}^t_{drums} = [\boldsymbol{W}^t_{snare}; \boldsymbol{W}^t_{cymbal}]$.

Regularizations

Virtanen (2007) (see Section 3.4) proposed the use of temporal continuity and sparseness regularizations on the gains of an NMF process to isolate sustained harmonic sources. We extend these regularization terms to the the basis factor and integrate them into the proposed spectrum model based on SIMM.

In our proposed method we apply different regularizations to the factors $\hat{\boldsymbol{X}}_{m'}$ and $\hat{\boldsymbol{X}}_d$ in order to disambiguate between drums and other musical accompaniment. Drums are characterized by their wideband smooth spectral shape and their sparseness in the time axis, since they are often transient sounds with a short decay and a shorter attack. On the other hand we assume the spectral evolution of the other musical accompaniment to be smooth in time. We define two additional regularization terms to include this prior knowledge into the factorization. We propose a regularization on the basis that penalizes frequency domain discontinuities in the spectra. The smoothness regularization on the temporal axis of the gains proposed by Virtanen (2007)(Equation 3.57) is modified to evaluate the smoothness on the frequency domain of the basis, the resulting frequency continuity regularization is defined as:

$$\boldsymbol{J}^{fc}_{\boldsymbol{W}}(\boldsymbol{W}) = \sum_{w}^{N_W} \frac{1}{\sigma_w^{\omega\,2}} \sum_{\omega}^{N_\omega} \Big([\boldsymbol{W}]_{\omega,w} - [\boldsymbol{W}]_{\omega-1,w} \Big)^2 \tag{6.14}$$

where the standard deviation of the components is estimated as $\sigma^\omega_w = \sqrt{(1/N_\omega)\sum_\omega^{N_\omega}([\boldsymbol{W}]^2_{\omega,w})}$.

The gradient of the regularization then becomes:

$$\left[\varphi^{fc}_{\boldsymbol{W}}(\boldsymbol{W})\right]_{\omega,w} = 2N_\omega \frac{2\left[\boldsymbol{W}\right]_{\omega,w} - \left[\boldsymbol{W}\right]_{\omega-1,w} - \left[\boldsymbol{W}\right]_{\omega+1,w}}{\sum_i^{N_\omega}\left[\boldsymbol{W}\right]^2_{w,i}} - N_\omega \frac{2\left[\boldsymbol{W}\right]_{\omega,w}\sum_{i=2}^{N_\omega}\left(\left[\boldsymbol{W}\right]_{i,w} - \left[\boldsymbol{W}\right]_{i-1,w}\right)^2}{\left(\sum_i^{N_\omega}\left[\boldsymbol{W}\right]^2_{i,w}\right)^2} \quad (6.15)$$

which can easily be expressed as an addition of positive and negative terms $\varphi^{fc^+_{\boldsymbol{W}}}$ and $\varphi^{fc^-_{\boldsymbol{W}}}$. The term w represent the basis index, ω the frequency index and t the time index (columns in $\boldsymbol{H}$).

We also propose a regularization on the drums activation matrix $\boldsymbol{H}_d$ that penalizes gains that are non-sparse in time. The regularization is a simple variation on that proposed by Virtanen (2007) (Equation 3.58).

$$\boldsymbol{J}^{ts}_{\boldsymbol{H}}(\boldsymbol{H}) = \sum_t^{N_T}\sum_w^{N_W} g(\left[\boldsymbol{H}\right]_{w,t}/\sigma_t) \quad (6.16)$$

where $g(\cdot)$ is a function that penalizes non-zero gains, in our case $g(x) = |x|$. The only difference between the regularization term proposed in Virtanen (2007) and the one we propose is that the standardization is done with respect to each time frame instead of each basis. The gradient then becomes:

$$\left[\varphi^{ts}_{\boldsymbol{H}}(\boldsymbol{H})\right]_{w,t} = \frac{1}{\sqrt{\frac{1}{N_W}\sum_i^{N_W}\left[\boldsymbol{H}\right]^2_{w,t}}} - \sqrt{N_W}\frac{\left[\boldsymbol{H}\right]_{w,t}\sum_i^{N_W}\left[\boldsymbol{H}\right]_{i,t}}{\left(\sum_i^{N_W}\left[\boldsymbol{H}\right]^2_{i,t}\right)^{3/2}} \quad (6.17)$$

Due to the additive nature of the spectrum model and regularizations, the derivation of the multiplicative update rules are quite straightforward. The multiplicative update rule for accompaniment $\boldsymbol{W}_{m'}$ remains the same as for $\boldsymbol{W}_m$ in the original SIMM method. The update rules for the $\boldsymbol{H}_{m'}$ $\boldsymbol{W}_d$ and $\boldsymbol{H}_d$ become:

$$\boldsymbol{H}_{m'} \leftarrow \boldsymbol{H}_{m'} \otimes \frac{\boldsymbol{W}^\top_{m'}\left(\hat{\boldsymbol{V}}^{(\beta-2)} \otimes \boldsymbol{V}\right) + \varphi^-_{\boldsymbol{H}_{m'}}}{\boldsymbol{W}^\top_{m'}\hat{\boldsymbol{V}}^{(\beta-1)} + \varphi^+_{\boldsymbol{H}_{m'}}} \quad (6.18)$$

$$\boldsymbol{H}_d \leftarrow \boldsymbol{H}_d \otimes \frac{\boldsymbol{W}_d^\top \left(\hat{\boldsymbol{V}}^{(\beta-2)} \otimes \boldsymbol{V}\right) + \varphi^-_{\boldsymbol{H}_d}}{\boldsymbol{W}_d^\top \hat{\boldsymbol{V}}^{(\beta-1)} + \varphi^+_{\boldsymbol{H}_d}} \tag{6.19}$$

$$\boldsymbol{W}_d \leftarrow \boldsymbol{W}_d \otimes \frac{\left(\hat{\boldsymbol{V}}^{(\beta-2)} \otimes \boldsymbol{V}\right) \boldsymbol{H}_d^\top + \varphi^-_{\boldsymbol{W}_d}}{\hat{\boldsymbol{V}}^{(\beta-1)} \boldsymbol{H}_d^\top + \varphi^+_{\boldsymbol{W}_d}} \tag{6.20}$$

where the gradient terms are defined as follows:

$$\begin{aligned} \varphi^-_{\boldsymbol{H}_{m'}} &= \alpha_{tc} \varphi^{tc-}_{\boldsymbol{H}_{m'}} \\ \varphi^+_{\boldsymbol{H}_{m'}} &= \alpha_{tc} \varphi^{tc+}_{\boldsymbol{H}_{m'}} \end{aligned} \tag{6.21}$$

$$\begin{aligned} \varphi^-_{\boldsymbol{H}_d} &= \alpha_{ts} \varphi^{ts-}_{\boldsymbol{H}_d} \\ \varphi^+_{\boldsymbol{H}_d} &= \alpha_{ts} \varphi^{ts+}_{\boldsymbol{H}_d} \end{aligned} \tag{6.22}$$

$$\begin{aligned} \varphi^-_{\boldsymbol{W}_d} &= \alpha_{fc} \varphi^{fc-}_{\boldsymbol{W}_d} \\ \varphi^+_{\boldsymbol{W}_d} &= \alpha_{fc} \varphi^{fc+}_{\boldsymbol{W}_d} \end{aligned} \tag{6.23}$$

and the parameters $\alpha_{tc} \in \Re^+$, $\alpha_{ts} \in \Re^+$ and $\alpha_{fc} \in \Re^+$ control the enforcement of the temporal continuity of the accompaniment gains $\boldsymbol{H}_{m'}$, the temporal sparseness of the drums gains $\boldsymbol{H}_d$ and the frequency continuity on the drums basis $\boldsymbol{W}_d$ respectively.

The regularizations can improve the separation between the musical accompaniment and the percussive components in the SIMM method. This separation is performed in an unsupervised manner since no signal-specific knowledge is needed. However the parameters controlling the regularizations may have a large influence on the results.

Constraints

Another extension proposed to the SIMM method for isolating the percussive instruments is the use of constraints. In this extension we assume the temporal positions of the drum events are known. This information is used to restrict the activation of the gains of the percussive components, reducing the degrees of freedom of the factorization problem. The constraints are performed in a manner similar to Ewert and Müller (2011).

We consider a set of percussive sources $m_d \in [1, N_{Md}]$. We denote $t_e^{m_d}$ for $e \in [1, N_e]$ the frame indices of the attacks of the events of the percussive source m_d. The dictionary $\boldsymbol{W}_d$ is the set of basis components for all the percussive sources, with ${N_W}^s$ components assigned to each percussive source. The constraints are set in the form of initializations to 0 in the corresponding gains matrix $\boldsymbol{H}_d$:

$$\boldsymbol{H}_d[w,t] = \begin{cases} \gamma, & \text{if } t_e^{m_d} - (1-\alpha)\tau < t < t_e^{m_d} + \alpha\tau \\ & \quad \text{and } (m-1){N_W}^s < w < m{N_W}^s \quad \forall m_d, t_e \\ 0, & \text{else} \end{cases} \tag{6.24}$$

where $\gamma > 0$ is a random positive value, τ is a parameter that controls the size of the event region and α controls the position of the active region around the event position.

We examine two different ways of supplying the drum event positions $t_e^{m_d}$. We propose an unsupervised approach using a transient estimation method and two scenarios with user-supplied annotations.

Transient Analysis The transient analysis used to evaluate the constraint-based unsupervised method is the same one used in Section 5.4. It is based on the work by Röbel (2003) where the spectral peak center of gravity is used as a measure of transient quality. This measure is coupled with a band analysis and thresholding in order to extract a frame-level decision about the presence of a percussive event attack. This method for drum event estimation is quite straightforward and serves as a baseline for constraint-based blind drums separation methods. State-of-the-art drum estimation techniques can achieve much better results, probably leading to improved separation.

Annotations Two main scenarios for user-supplied annotations are considered. The first consists in creating different annotations sets for each of the drum sounds (bass drum, snare drum, closed hi-hat, open hi-hat,...). This implies having multiple drum sources $N_{Md_{ind}} > 1$ in our spectrum model. The second technique uses a single set of annotations, by merging all the drum sounds together $N_{Md_{join}} = 1$, in order to keep both approaches comparable, the number of basis components used in the second approach is ${N_W}^s_{join} = N_{Md_{ind}}{N_W}^s$. The annotations of the drum events were manually performed by an amateur experienced drum player using the Son-

icVisualiser software application [7]. The annotations were created using the isolated drum tracks in order to evaluate near-optimal separation using a constraint-based method. The annotations dataset has been made publicly available online [8].

Experiments

We used the same dataset of multitrack audio recordings with drums as in Section 5.4 to evaluate the proposed methods. We conducted measurements using the PEASS and BSSOracle frameworks to obtain quantitative results for the separation. We performed two series of experiments (regularization and constraints) evaluating the results of the different methods.

The first set of tests consisted of parameter explorations of the regularization-based methods (REG). In these experiments we tested the separation for multiple values of the time continuity regularization $\alpha_{tc} = 25$ (SM25), $\alpha_{tc} = 50$ (SM50), $\alpha_{tc} = 75$ (SM75), $\alpha_{tc} = 100$ (SM100) for the non-percussive accompaniment basis $\boldsymbol{W}_m$. We also evaluated the effect of employing a sparseness regularization $\alpha_{ts} = 10$ (SP10) on the drums gains. The regularizations for the frequency continuity of the non-percussive accompaniment has been kept to a fixed value $\alpha_{fc} = 1$. These tests were conducted in an unsupervised scenario (UNS) where all the drum basis components are learned during the separation and a semi-supervised (SUP) case where the basis components are learned previously using training data with the drums in isolation.

In a second series of experiments we evaluated three constraint-based methods. We compared a blind transient analysis method (CON-TR) to two annotated methods: individual sources model (CON-AN-I), and joint sources model (CON-AN-J). We explored the influence of the main parameter ${N_W}^s$ on each method and the effect of using the SIMM lead voice model with an external annotated pitch (CON-TR-NP, CON-AN-I-NP, CON-AN-J-NP). Finally we performed a comparative evaluation with state-of-the-art methods THPS-TIK (from Section 5.4), HPSS (Ono et al., 2008b) and FASST (Ozerov et al., 2010). The best parameter combination found in the parameter exploration was used in the comparative tests.

7. http://www.sonicvisualizer.org
8. http://mtg.upf.edu/download/datasets/dreanss

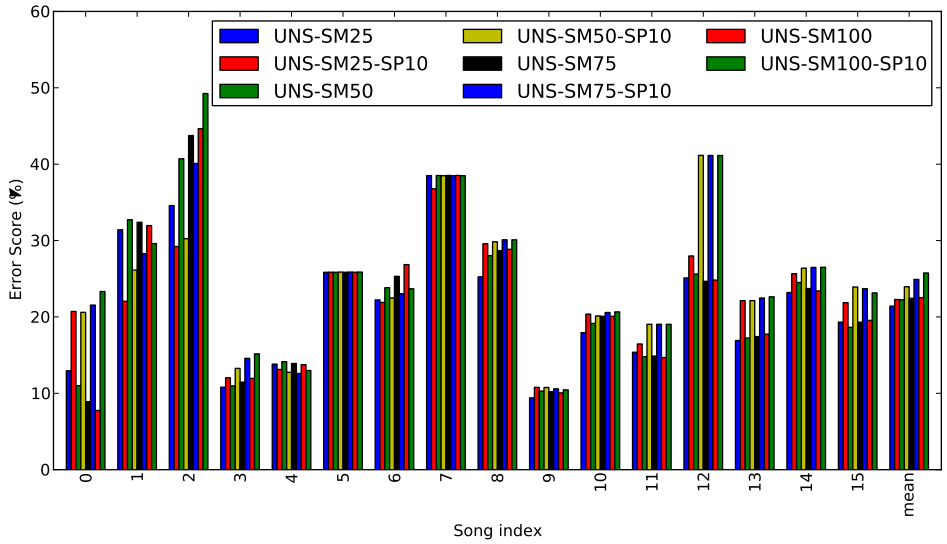

Figure 6.9: Individual OPS error measures for the drums separation unsupervised scenario with relation to the regularizations applied.

Discussion

Regularizations Experiments

Figure 6.9 and Figure 6.10 show the Overall Perceptual Score (OPS) errors relative to Oracle, for the individual excerpts in the unsupervised and semi-supervised configurations. We can appreciate that in both scenarios the results are not conclusive, since the OPS error varies a lot with changes in the regularization parameters. For the unsupervised configuration, on average we observe an increase of the error with the amount of temporal continuity regularization applied to the accompaniment gains. The average result also shows that the application of the sparseness is detrimental since it increases the separation errors. The average results show very little variation for the semi-supervised scenario.

However we do notice that for certain excerpts, such as for excerpt 0 in the unsupervised case, temporal continuity regularization causes a significant improvement. This improvement for individual excerpts is more visible still for the temporal sparseness regularization parameter of the drums gains $\boldsymbol{H}_d$.

In Figures 6.11 and 6.12 we plot a histogram of the improvements from

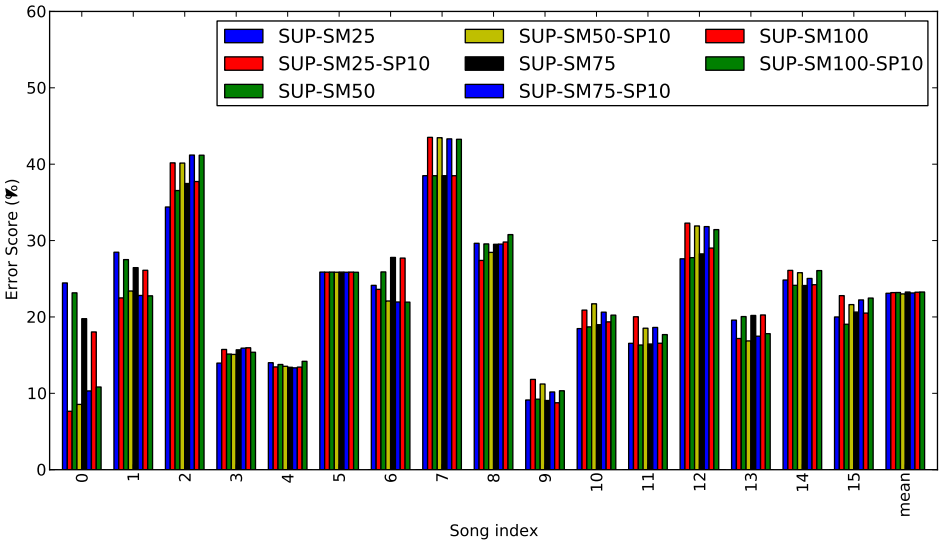

Figure 6.10: Individual OPS error measures for the drums separation semi-supervised scenario with relation to the regularizations applied.

adding sparseness regularization. This value is computed as the difference of OPS error obtained with the method using sparseness regularization and that obtained without using it. These values are computed for all the values of the temporal continuity regularization. The histograms show a large variance in improvement. In some cases the use of $\alpha_{ts} = 10$ creates a large improvement and in others the opposite.

These results suggest the utility of future investigation of the dependence of optimal regularization parameters on the data, and the potential for deriving methods to estimate optimal regularization for each excerpt to be analyzed.

Informal listening to the results confirms the findings that we show here. In some excerpts the regularization improves the separation while in others it is disadvantageous. We can also see that the regularizations behave as expected, controlling the desired spectro-temporal qualities of the estimated sources. In general we also observe that semi-supervised separation maintains the bass drum and snare sources better. Unsupervised separation tends to produce a filtered signal keeping only mid-high components. A drawback of the semi-supervised version is the greater interference between lead vocals and the bass line.

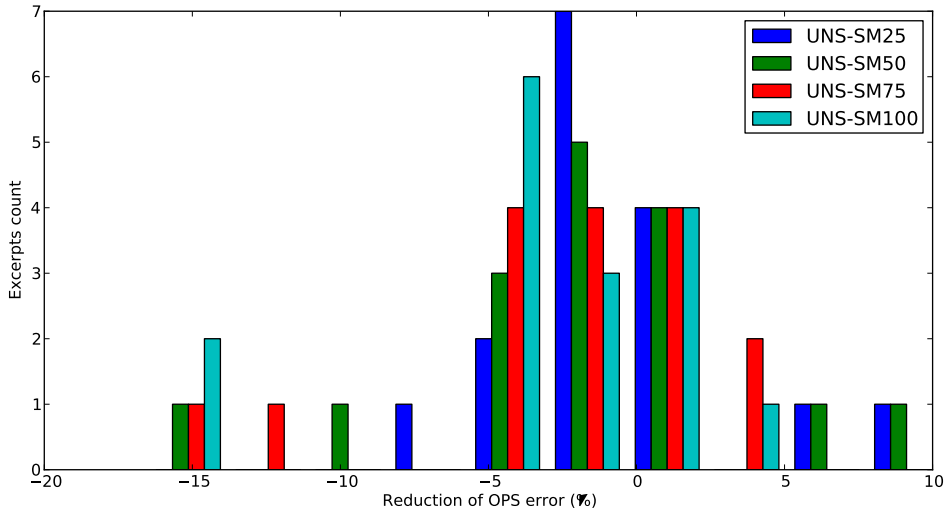

Figure 6.11: Histogram of the OPS improvement by using the sparseness regularization (SP10) in the unsupervised scenario.

Constraints Experiments

Figures 6.13 and 6.14 show the ${N_W}^s$ parameter exploration experiment for the constraint-based method that uses annotations of the individual drums sources (CON-AN-I). This represents the method with the most prior information supplied about the mixture and can serve as a maximum for our proposed constraint-based methods. The plot of the OPS and APS score errors shows that the results vary slightly depending on the number of basis components assigned to each drum source ${N_W}^s$. There are several local minima implying that there is no unique optimal value for all excerpts and drum sources.

In terms of TPS and IPS the number of basis components ${N_W}^s$ controls the tradeoff between target fidelity and interference. This is an expected result, since a large number of basis components to reconstruct drum components could lead to overfitting of the mixture spectra causing the capture of other non-percussive components thus increasing the interference while at the same time better reconstructing the target drums.

Figures 6.15 and 6.16 show a similar trend when the constraints are based on generic drums annotations $N_{Md_{join}}$, without making a distinction between drum sounds (CON-AN-J). In future work we should investigate optimizing

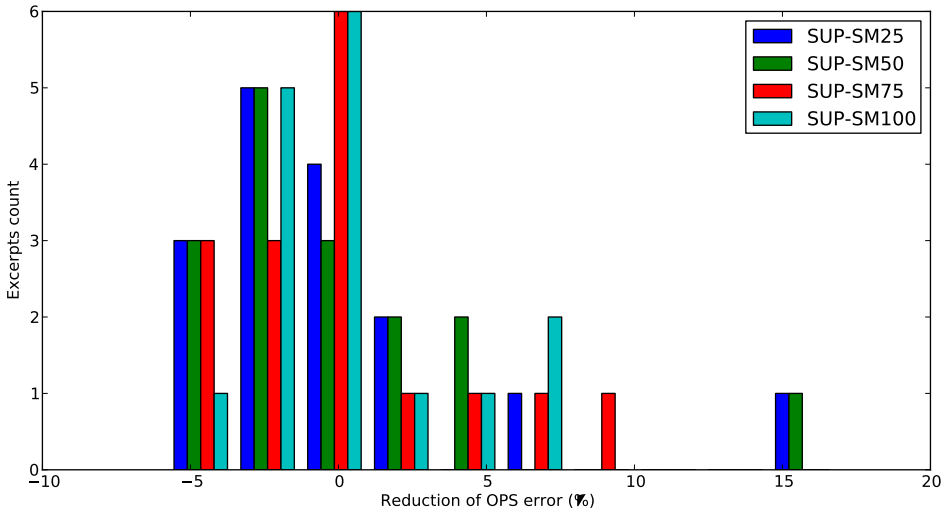

Figure 6.12: Histogram of the OPS improvement by using the sparseness regularization (SP10) in the supervised scenario.

the parameter for each drum type and its dependence on the number of occurrences in the excerpt.

Figure 6.17 shows the effect of implementing constraint-based methods as extensions of the SIMM approach in contrast to not performing the lead voice estimation (NP). There is a reduction of the OPS error in all the constraint-based methods. This improvement is mainly due to a decrease in interference and informal listening to the results confirms this finding. The lead voice is often an energetic component and by specifically modeling it we significantly reduce the parts of it that are counted as drum sounds.

Figure 6.18 shows how these constraint-based methods relate to other state-of-the-art drums separation approaches. The annotation-based informed source separation methods show a clear improvement in OPS over the blind techniques. This indicates that the development of better temporal estimation of the drum event positions could lead to significant improvements in blind drums separation. The difference between annotations of individual drum sources (CON-AN-I) and generic drum sources (CON-AN-J) is insignificant, from which we can conclude that estimation of general drums events should be sufficient.

The artifact-based scores (APS) show unexpected results where the FASST

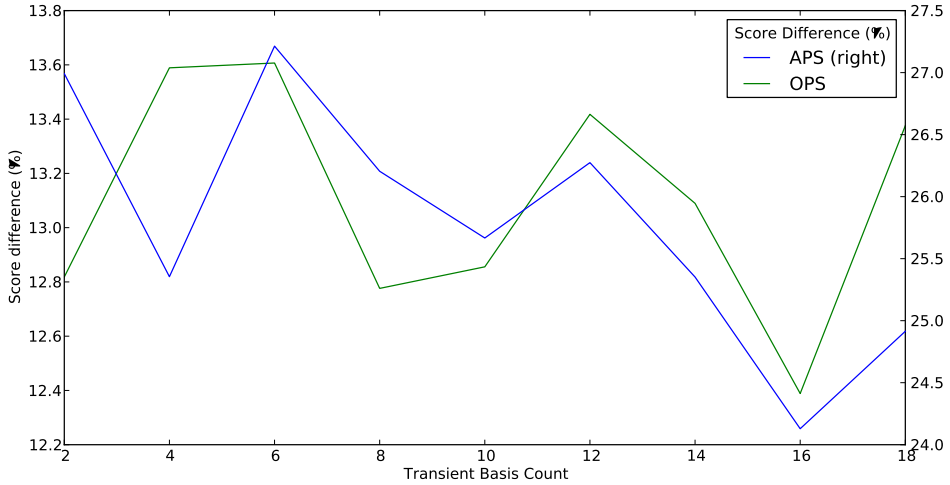

Figure 6.13: OPS and APS score errors with relation to ${N_W}^s$ for the constraint-based individual annotation method (CON-AN-I).

method achieves better average scores (negative score difference) than the Oracle version. This is probably due to the perceptual-inspired relations in the PEASS framework, since the non-perceptual-related BSSEval results in Figure 6.19 do not present this behavior.

Finally, we can observe that the naive blind transient constraint-based method (CON-TR-J) does not achieve results comparable to other blind techniques. Note that the transient detection is not adapted to drums and thus is prone to false positives caused by other instruments.

Subjective assessment by informal listening to the comparative study confirms the trend presented in Figure 6.18. The main shortcoming of the constraint-based methods is that the full decay of the drums is often not preserved. Increasing the parameter τ could help reduce this issue, however it would also increase the amount of noise in the learning process of the drums component basis during the factorization. In the future, studying the relations between τ and N_{Wd} might be useful since together they influence the amount of overfitting and underfitting of the problem.

Conclusions

We propose and study an extension to the SIMM method to perform drums separation. The proposed extension makes use of regularizations and con-

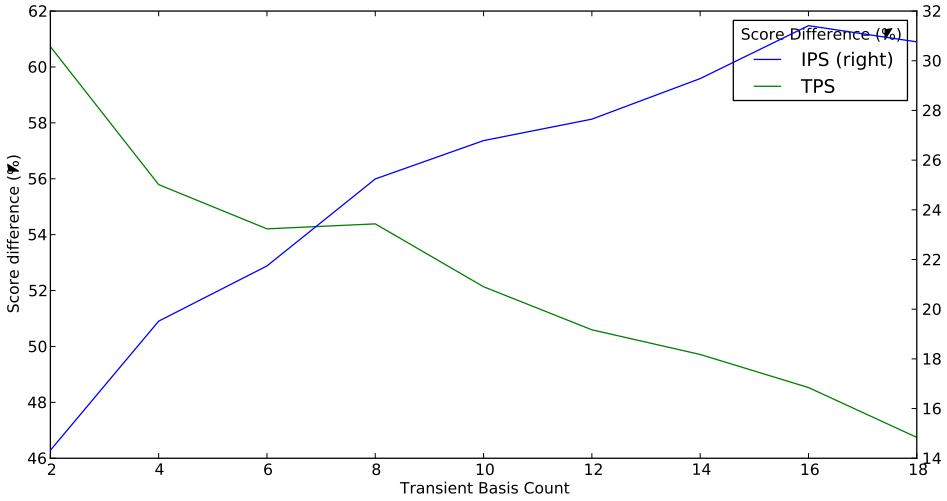

Figure 6.14: TPS and IPS score errors with relation to ${N_W}^s$ for the constraint-based individual annotation method (CON-AN-I).

straints to separate percussive from non-percussive music accompaniment. We propose two new regularization terms that consist in small variations on those proposed by Virtanen (2007). The proposed regularizations control the frequency smoothness of the basis components and the temporal sparseness of the gains. These regularizations are used together with the temporal continuity regularization of the gains to perform blind drums separation. We also study the effect of using a set of pre-trained basis components for drums sources. The experiments show that the optimal value for the strength of the regularizations is highly dependent on the excerpt.

We evaluate the use of temporal constraints on the gains to perform drums separation. The technique consists of using the positions of the drums events in the mixture to limit the regions of activation of the drums basis. This technique is tested using both ground truth manual annotations from the isolated tracks and automatically extracted transients from the mixture. This allows us to assess both a glass ceiling and a baseline for this type of approach. Results show that a simple transient estimation technique is insufficient for this task, compared to the method with manual annotations or other state-of-the-art methods. Additionally we test how the number of basis components assigned to each drum source affects the quality of the separation. The results show that the overall performance and the artifacts related score do not vary much with respect to this parameter.

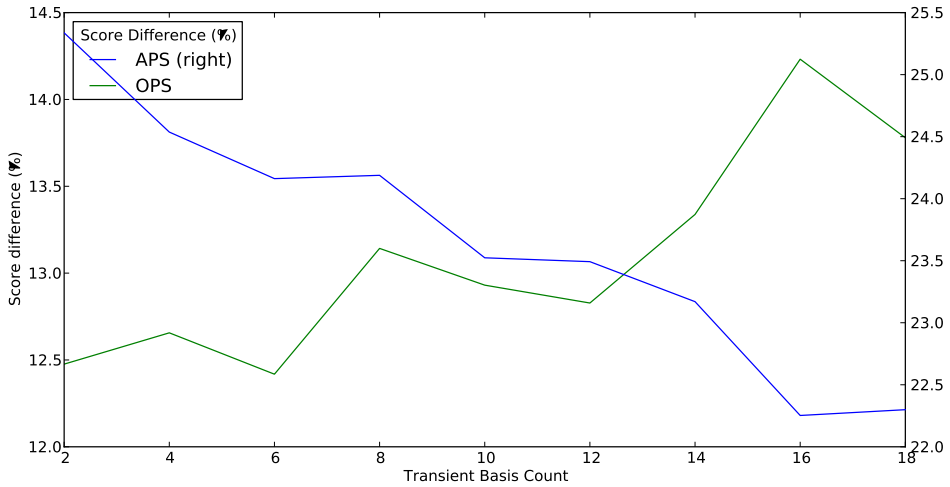

Figure 6.15: OPS and APS score errors with relation to ${N_W}^s$ for the constraint-based joint annotation method (CON-AN-J).

This parameter controls the tradeoff between interference and target related scores.

We also observe that it may not be of much benefit to estimate the positions of the individual drum sounds (closed hi-hat, open hi-hat, snare drum,...) since this does not significantly improve the separation results. However it remains to be tested whether using different parameter values per type of drum sound enhances the results. Furthermore the use of frequency domain constraints specific to each drum type could also improve the separation. Another possible future direction could be to perform a two step strategy, where a subset of the drum positions are first used to estimate the basis components, and a second step in which the separation is done by loosening the temporal constraints.

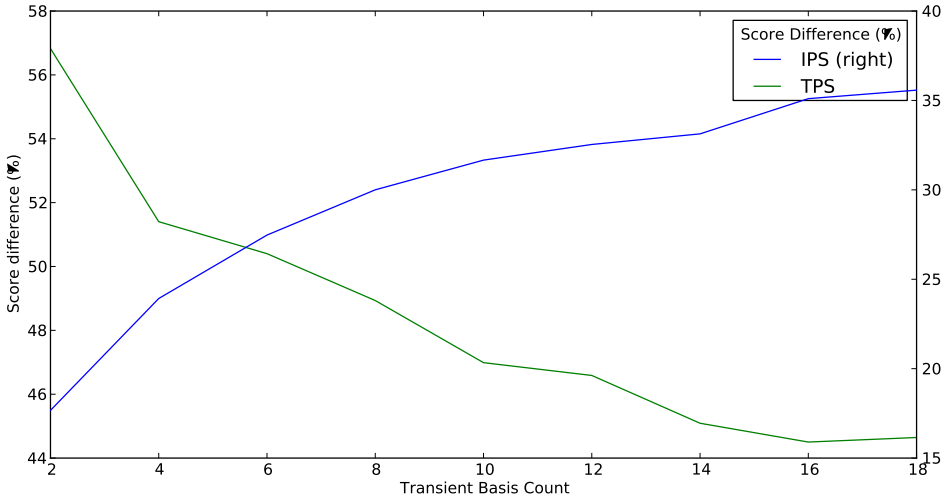

Figure 6.16: TPS and IPS score errors with relation to ${N_W}^s$ for the constraint-based joint annotation method (CON-AN-J).

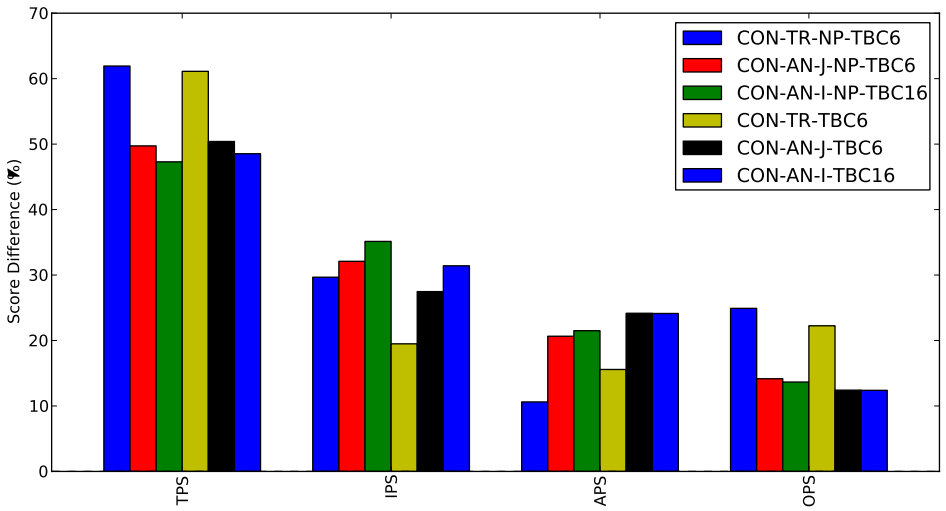

Figure 6.17: Effect of the lead voice estimation on the constraint-based methods, using ${N_W}^s = 6$.

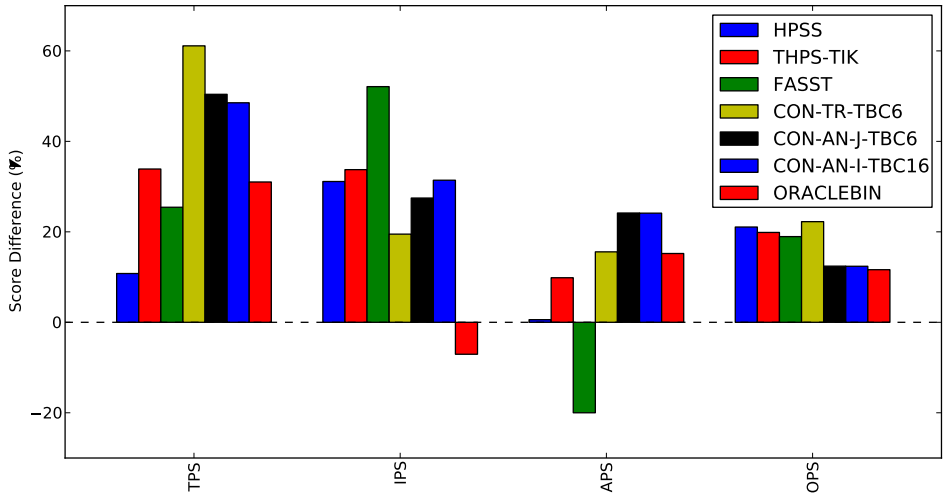

Figure 6.18: PEASS results of the comparative study of the constraint-based methods for drums separation.

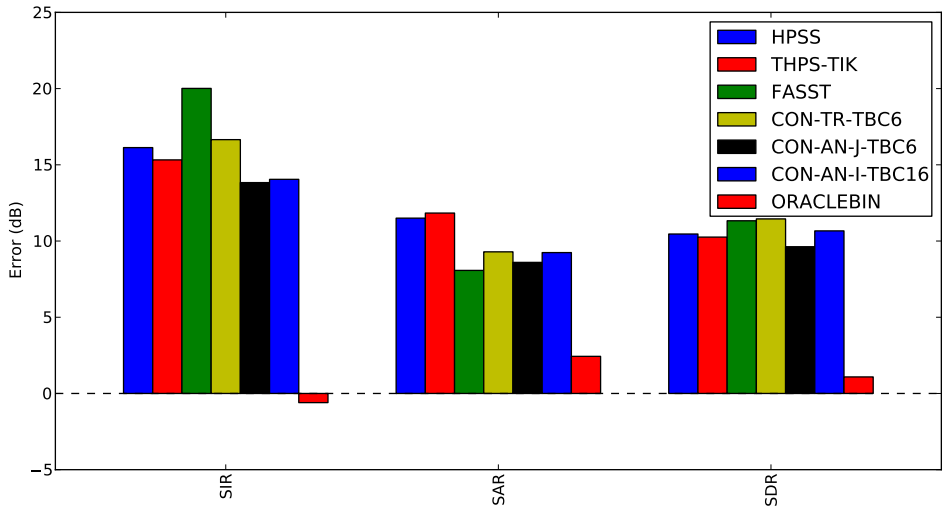

Figure 6.19: BSSEval results of the comparative study of the constraint-based methods for drums separation.

6.6 Multiple Instruments Separation

This section addresses the modeling of multiple monophonic instrument voices in a polytimbral audio source separation problem. We present a generalization of the Smoothed Instantaneous Mixing Model (*SIMM*) method which improves results for mixtures of multiple instruments. It is compared to the single-source filter approach and to a multiple-source, single filter extension as well as to the original *SIMM*. The method is evaluated on polyphonic audio recordings in three datasets with different mixtures: wind instruments, vocal choir and multi-talker speech. In a pitch-informed experiment, where pitch is extracted from the isolated tracks, we find that the proposed methods cope better with simultaneous timbres in the mixtures. We obtain an improvement of 3dB SDR on objective measures and a clear gain in listening quality.

Introduction

Source separation in polyphonic and polytimbral music mixtures is a complex task which has attracted much research interest in recent years. Several approaches consider the case of main instrument separation from the accompaniment, typically focusing on singing voice extraction (Durrieu et al., 2009b; Virtanen et al., 2008b; Hsu and Jang, 2010b). Other works address the separation of harmonic and percussion components, which has applications ranging from music transcription to real-time remixing (Helén and Virtanen, 2005; Ono et al., 2008b; Janer et al., 2012).

Most of these approaches are based on methods of automatic pitch estimation and tracking. While automatic predominant pitch tracking has proven successful (Marxer et al., 2012), automatic multipitch tracking has not been widely used in blind source separation tasks due to the complexity of accurate polyphonic multiple f0-frequency estimation (Klapuri, 2003).

Recent approaches have shifted from purely blind-source separation towards incorporating supervised source separation. They focus on obtaining high quality results with the help of additional information such as a musical score (Ewert and Müller, 2012), timbre training (Carabias-Orti et al., 2011; Rodriguez-Serrano et al., 2012) or pitch information manually provided by the user (Smaragdis and Mysore, 2009; Durrieu and Thiran, 2012).

Our work addresses the separation of multiple monophonic instrument mixtures that are *pitch-informed*, i.e. pitch contour is provided for each source in the mix. The assumption is that there is a multipitch detector, able

to accurately track the pitch contour of each instrument. In our experiments, the pitch contour is estimated from the isolated tracks by means of a monophonic pitch detection algorithm (de Cheveigné and Kawahara, 2002).

Proposed Extensions To SIMM

In a polyphonic mixture containing several harmonic instruments, one limitation of the SIMM method described in section 6.2 is that only a single instrument is represented by the harmonic model. We propose two extensions to the *SIMM* in order to model multiple pitched instruments.

Multiple Excitation Single Filter SIMM

The first extension to the *SIMM* is to add multiple excitations. This does not modify the spectrum model in any way. However we do not apply a *monophonicity* constraint to $\boldsymbol{H}_{f_0}$. Instead we apply a *polyphonicity* constraint that zeroes all the excitations other than those around multiple given pitches. When estimating the target source spectrum, we set to zero all the gains $\boldsymbol{H}_{f_0}$ of the excitations except those around the target instrument pitch. This method implies all instruments are to be modeled with the same filter basis $\boldsymbol{W}_\Gamma \boldsymbol{H}_\Gamma$. In contrast to the existing method *SIMM*, here we increase the number of filter basis components $K_I = K \cdot N_I$, proportional to the number of sources in the mixture.

Multiple Excitation-Filter SIMM

The second extension, the multiple-excitation filter *SIMM* model, is a generalization of *SIMM* which assumes that we have multiple source filter models. Under this assumption the spectrum model becomes:

$$\begin{aligned} \hat{\boldsymbol{V}}_{msfSIMM} &= \sum_{l}^{N_L} \Big((\boldsymbol{W}_\Gamma \boldsymbol{H}_\Gamma^l \boldsymbol{H}_\Phi^l) \otimes (\boldsymbol{W}_{f_0} \boldsymbol{H}_{f_0}^l) \Big) \\ &\quad + \boldsymbol{W}_M \boldsymbol{H}_M \end{aligned} \tag{6.25}$$

Monophonicity constraints are applied to each $\boldsymbol{H}_{f_0}^l$, assigning the l^{th} tracked pitch to $\boldsymbol{H}_{f_0}^l$. The update rules of this method remain the same as for the *SIMM* method (Durrieu et al., 2011). However we must point out that we have more matrices to update since for each source $\boldsymbol{H}_\Gamma^l$,$\boldsymbol{H}_\Phi^l$ and $\boldsymbol{H}_{f_0}^l$ are different. The mixture spectrogram estimation $\hat{\boldsymbol{V}}_{msfSIMM}$ is also different

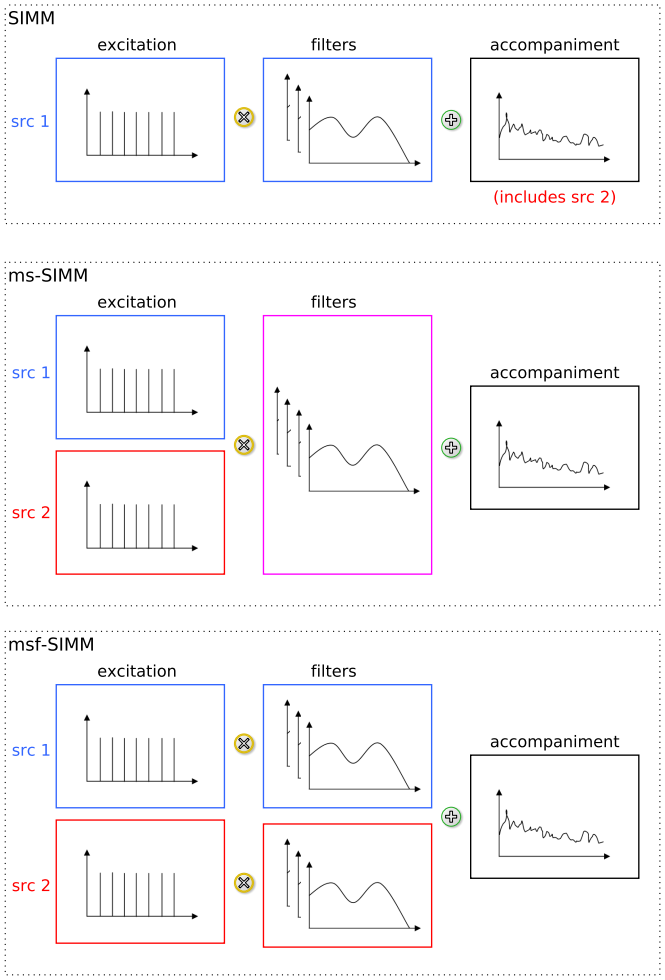

Figure 6.20: Examples of the source/filter configurations for the existing (*SIMM*) and the two proposed methods (*ms-SIMM* and *msf-SIMM*), for the case of a two-source mixture.

from that of the *SIMM*. When estimating the spectrum of the target source we use only the gains $\boldsymbol{H}^l_\Gamma$,$\boldsymbol{H}^l_\Phi$ and $\boldsymbol{H}^l_{f_0}$ of the corresponding pitch.

Experimental Setup

The experimental setup proposed in this study is similar to that proposed in Vincent et al. (2010) and Carabias-Orti et al. (2011). We evaluate the

separation performance of different spectrum models given the pitch annotations of the different instruments in the mixture.

Test Data

We consider three different multi-track datasets: *wind* (a wind instruments quintet), *choir* (a vocal quartet) and *speech* (multi-talker speech).

The first dataset is the wind instruments database for the Multiple Fundamental Estimation task of the Third Music Information Retrieval Evaluation Exchange (MIREX2007). This dataset is composed of a woodwind quintet recording of the fifth variation from Beethoven's Variations for String Quartet Op.18 No. 5. Each instrument (flute, oboe, clarinet, horn, and bassoon) was recorded separately while the performer listened to the other parts (recorded previously) through headphones. The mixtures are generated by mixing the recordings of the individual instruments, with polyphonies ranging from 2 to 5. This combinatorial process results in a total of 26 individual mixtures.

The second dataset consists of recordings of four voices (bass, tenor, alto and soprano) of the choir composition "Water Night", composed by Eric Whitacre. Isolated solo recordings were downloaded from the Virtual Choir site [9]. The goal is to observe the effect of having sources with similar timbre (singing voice), pitch contours with overlapping harmonic partials and polyphonies ranging from 2 to 4. The dataset has a total of 11 mixtures.

We also use a non-musical sample with multi-talker speech signals to observe the influence of pitch consonance. Speech shows a varying pitch contour driven by the prosody, and in a situation with simultaneous talkers no consonance is expected. To build the multitalker dataset, we randomly selected 10 sentences of four talkers from the GRID corpus (Cooke et al., 2006). generating four separate signals (one per talker) with a duration of 20 seconds. Here also the polyphonies range from 2 to 4, and the dataset has a total of 11 mixtures.

Annotation Data

The pitch annotation of each track is carried out automatically using a monophonic pitch estimation method (de Cheveigné and Kawahara, 2002)

9. These recordings are copyrighted and available online: http://ericwhitacre.com/the-virtual-choir/resources

on the individual recordings. The same pitch range (30 − 1800 Hz) and voiciness threshold are used to process all recordings. Pitch data is computed using a frame rate of 86 fps.

Algorithm parameters

In this experiment, we use the following algorithm parameters for all compared methods and datasets. Input signal has a sampling rate of 44.1 kHz. The spectrogram is computed using a Sinebell window of 4096 samples, with a hop size of 512 samples, without zero-padding. The separation algorithm works block-wise, processing consecutive blocks of 1000 frames (11.61 seconds). For the spectrum model we use excitation basis ($\boldsymbol{W}_{f_0}$) with frequencies ranging from 27.5Hz to 2489Hz. The smooth filters ($\boldsymbol{W}_{\Gamma}$) are composed of 60 linearly distributed Gaussians. The timbres ($\boldsymbol{H}_{\Phi}$) are modeled using a linear combinations of K_I smooth filters with $K = 50$. Finally, we use 40 accompaniment basis components in $\boldsymbol{W}_M$.

Evaluation measures

We use the measures found in the BSSEval toolkit proposed by Vincent et al. (2006) to objectively compare the results of the different separation methods. The measures used in this study are: SDR (Signal to Distortion Ratios), SIR (Source to Interference Ratios) and SAR (Sources to Artifacts Ratios). To more consistently compare the results of mixtures of different complexity (datasets and polyphony), we compare the values for each separated track to an oracle estimator (Vincent et al., 2007a) as a reference baseline. As in Marxer and Janer (2012), we use an error measure (e.g. SDR error) which is the difference between the measure obtained by the oracle estimator and the measure obtained by each separation algorithm.

Results

At first glance the results show a lot of space for improvement in SDR to achieve results equivalent to the Oracle solution (see Tables 6.3 6.4 6.5): 10dB for speech signals and 16dB for musical signals. However in this work we focus only on the harmonic components of signals, without attempting to separate other transient or wideband parts.

There is clear improvement in absolute separation performance of musical signals with respect to speech signals. This is due to the fact that the

Polyphony	*2*			*3*			*4*		
Method	SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR
SIMM	11.2	8.0	17.6	10.3	6.9	17.3	9.9	6.7	17.4
ms-SIMM	10.9	9.0	14.1	9.8	8.0	13.7	9.3	7.8	14.3
msf-SIMM	10.6	9.2	12.9	9.4	8.1	11.3	8.8	7.8	11.6

Table 6.3: BSSEval Results for the *speech* dataset

Polyphony	*2*			*3*			*4*			*5*		
Method	SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR
SIMM	18.6	18.5	20.2	17.5	17.1	21.3	17.1	16.6	21.3	16.7	16.2	21.1
ms-SIMM	17.9	18.7	13.8	15.8	16.8	14.8	15.0	16.2	14.9	14.5	15.9	14.9
msf-SIMM	16.8	17.5	12.9	14.4	15.4	12.5	13.3	14.6	11.9	12.6	14.2	11.3

Table 6.4: BSSEval Results for the *wind* dataset

Polyphony	*2*			*3*			*4*		
Method	SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR
SIMM	18.5	17.9	20.2	16.2	15.5	21.0	15.9	14.9	22.9
ms-SIMM	18.4	18.5	14.8	16.4	16.7	15.1	15.2	15.6	14.9
msf-SIMM	15.5	15.1	11.7	13.7	13.0	13.7	12.8	12.2	14.0

Table 6.5: BSSEval Results for the *choir* dataset

musical signals we focus on are dominated by harmonic components which lead to larger disjointness of the sources.

The proposed *msf-SIMM* method outperforms the existing *SIMM* and the proposed *ms-SIMM* method for all the polyphonies and all the datasets, however for speech signals the improvement is minimal. This can be explained by the fact that speech signals have very few regions where the partials of the sources overlap. Western music, on the other hand, is characterized by large regions of pitch consonance, where the sources have pitches with small integer ratios of frequency. This results in a significant number of partials overlapping, which is where the multiple source-filter model excels. This effect is most visible in the choir signals where even the *ms-SIMM* fails to show better results than the traditional *SIMM* method.

Another important result is that the SDR error relative to the Oracle decreases as the polyphony increases (see Figure 6.21). At first glance this result is surprising since it would seem the source separation problem is

more difficult when more sources are present. In Figure 6.22 we see that the absolute values of the SDR decrease as the polyphony increases. This is also the case for the Oracle solution, however this decrease is more pronounced than for the source separation methods.

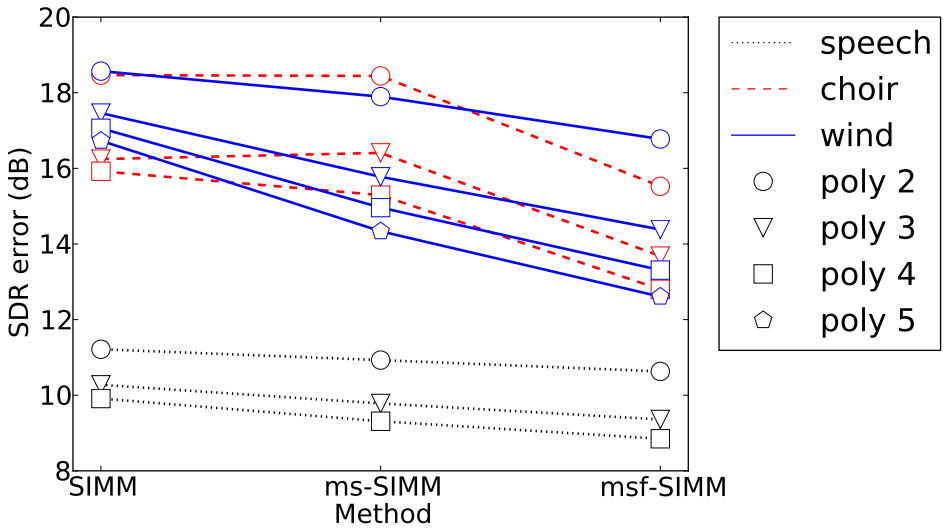

Figure 6.21: SDR error by method. Relative separation **error** decreases with the proposed methods *ms-SIMM* and *msf-SIMM*.

For demonstration purposes, we prepared a web site [10] containing all the processed audio examples. The quality improvement obtained by the proposed methods is especially noticeable with the *wind* dataset, where the instrument timbre is better preserved when using the *msf-SIMM* method.

Conclusions

We study a very specific audio source separation scenario where the pitch contours of the different sources in a mixture are known in advance. Although this is not a common real world case, user- and score-assisted multipitch estimation techniques are becoming more common. We propose a source separation method especially adapted to these situations. We also show the types of signals for which this method gives the best performance

10. Audio examples can be found in:
http://www.dtic.upf.edu/~rmarxer/papers/multipitch_separation/

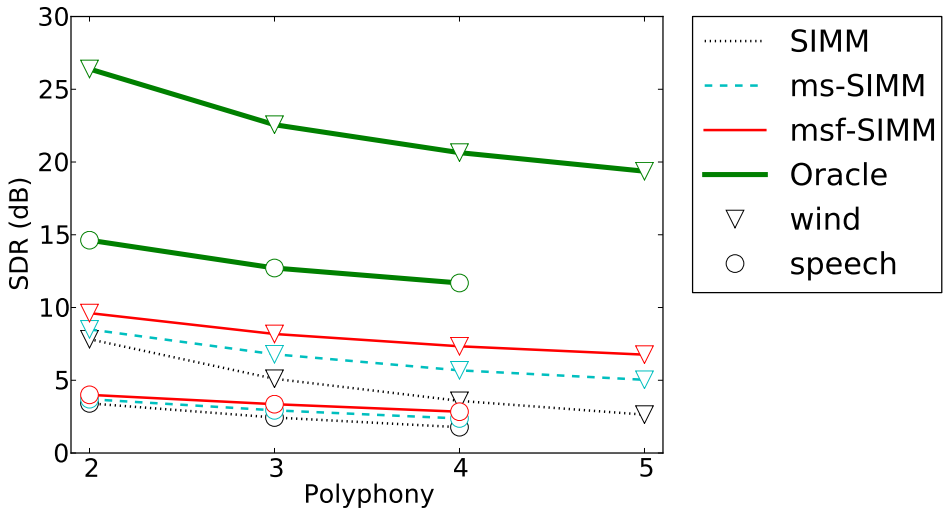

Figure 6.22: SDR by polyphony. Absolute separation **performance** decreases with the polyphony.

improvement over current techniques. The proposed method (*msf-SIMM*) generalizes the well known *SIMM* (Durrieu et al., 2011) to multiple source-filter models. This differs from the work by Rodriguez-Serrano et al. (2012) since they perform the pitch estimation automatically and assume knowledge about the instrument's timbre gained by performing a previous training stage. In our case the timbre is learned automatically from the signal, however the pitch of the different sources must be provided. *msf-SIMM* presents an improvement with all types of signals, especially music signals where an improvement of up to 3dB can be observed in the objective evaluations. This improvement can also be clearly perceived by listening to the results, showing the timbres of the separated sources to be much more stable and similar to the original.

Part IV

Chapter 7

Conclusions

The work presented in this dissertation was motivated by the need for audio source separation methods that can be used in a wide range of real-world audio applications. We focus on two main issues with current separation techniques: the high latency and computational cost of some techniques and the low audio quality of others. The specific scenario addressed is the separation of monaural and stereo instantaneous mixtures of western commercial music. We set our focus on the singing voice, the bass and the drums, due to their diversity from a signals perspective and their importance in western popular music. We propose a number of novel approaches that improve over the state-of-the-art in both low-latency and high-latency situations.

In low-latency settings we propose the use of Tikhonov regularization as a low-computational-cost and easy-to-implement technique for spectrum decomposition. To our knowledge this technique has not been previously used to perform audio spectrum factorization. This method is then proven useful in a series of tasks covering target source predominant pitch estimation, multiple pitch estimation, and singing voice, bass and drums separation. In the high-latency case, singing voice has been extensively studied in source separation research. Therefore it is harder to make significant improvements in this scenario. We propose novel methods to handle specific components of the singing voice that have rarely been tackled in the source separation literature. These methods focus on breathiness and unvoiced fricatives and they provide observable improvements in separation quality. We also propose a series of methods for separating drums in both unsupervised and supervised contexts. Finally we focus on multiple monophonic pitch-informed

separation.

7.1 Summary of results and contributions

Here we compile the main contributions and results in the order they are presented in the dissertation. For further details the reader is referred to the conclusions sections in the relevant chapters.

Novel Spectral Decomposition Approach based on Tikhonov Regularization The starting point of this work is the proposal to use Tikhonov Regularization as a spectral decomposition method (Section 4.2). The proposed method targets cases in which the basis matrix is fixed and independent from the signal, and in which spectral frames are analyzed one at a time. Processing the frames one at a time makes this method especially interesting in low-latency scenarios. The proposed decomposition approach is then used throughout the low-latency part of this dissertation with basis matrices specifically designed for each task. The main advantage of this method is the availability of a closed-form solution, in contrast with the iterative non-unique solution of the NMF approach. The implementation of the Tikhonov Regularization is very simple. The processing of each spectral frame consists in performing a common matrix multiplication which can be easily performed in many embedded architectures and significantly optimized.

Low-latency Singing Voice Pitch Tracking using Timbre Models The first task considered for the application of the Tikhonov Regularization spectral decomposition method is the estimation of the pitch likelihood (Section 4.3). The basis matrix for the decomposition is composed of a set of harmonic and wideband spectra. In the context of pitch tracking, the proposed pitch likelihood estimation method is compared to an approach based on the well known Harmonic Summation technique. The results show similar performance for both methods, with the proposed approach having the added advantage that it also estimates the spectrum of the pitched source. We also propose a pitch tracking method based on an online Viterbi algorithm with a limited latency on an HMM (Section 4.3). Furthermore the tracking algorithm integrates information about the timbre of each pitch candidate to enforce tracking the target instrument. The timbre information is based on MFCC features of the harmonic envelope and a trained SVM classifier. This method tracks the predominant pitch of a specific instrument

for which the training was performed. The method is compared to pitch tracking without using timbre information in a task of singing voice pitch estimation. The proposed method significantly decreases the number of false positives which is especially interesting in source separation tasks to avoid the removal/isolation of non-targeted instruments.

Low-latency Multiple Pitch Tracking After the single pitch estimation task we focus on the estimation and tracking of multiple simultaneous pitches. We present a prototype of a low-latency multiple pitch tracking system composed of tracking and selection stages. Based on the previous results, we propose a method to improve the selection and characterization of candidates in the pitch likelihood function (Section 4.4). The proposed technique targets pitch likelihood functions created using generative approaches such as spectral decomposition. The method consists in modeling the peaks in the likelihood function as Gaussian functions, and using the divergence between these as a transition probability in the HMM of the tracking stage. Empirical tests show significant improvements in the predominant pitch tracking task, however a quantitative evaluation remains to be done. For the multiple pitch tracking stage, we conceive an extension to the previous Viterbi algorithm that uses multiple copies of the HMM and iterative cancellation of the best paths. For the selection stage a set of pitch contour features that can be computed incrementally is proposed. Among them we put special attention on one that characterizes octave error probability of the contour, which is a common issue in multiple pitch estimation using generative models, such as the spectral decomposition based on Tikhonov regularization. Multipitch estimation and tracking is not a main objective in this dissertation and the proposed system is not quantitatively evaluated, however initial tests show promising results.

Real-time Singing Voice Separation using Tikhonov Regularization In the context of audio source separation we first focus on the singing voice (Section 5.3). We develop two low-latency singing voice separation systems based on the Tikhonov regularization decomposition of the spectra. In the first, we use the previously proposed pitch estimation method and perform a binary harmonic TF mask separation. A notable result is that the use of timbre models in the pitch tracking is able to improve the separation significantly when compared to other real-time methods such as pan-based TF masks. Another significant result is that there is still room for improvements to achieve the quality of state-of-the-art high-latency ap-

proaches. The second system tries to improve over the harmonic binary mask, by performing Wiener filtering using the Tikhonov regularization decomposition. To cope with the varying timbres of the singing voice we propose using a basis matrix composed of filtered harmonic components that, linearly combined, can reconstruct multiple harmonic envelopes. In this case the number of basis components increases the computational cost as well. However the proposed Tikhonov regularization decomposition remains significantly less computationally costly than standard NMF while achieving similar separation quality in a simple low-latency singing voice isolation task.

Real-time Drums Separation using Tikhonov Regularization Another contribution of our work is the use of Tikhonov regularization spectral decomposition in a drums separation task (Section 5.4). We propose a system that estimates the percussion spectrum with a single frame of latency based on the harmonicity of the spectral bins and the transient quality of the spectral peaks. The harmonicity is estimated by picking peaks in the pitch likelihood function and computing the respective harmonic TF masks. A transient mask is computed using the center of gravity of the spectral peaks. The combination of both masks is used to separate the percussive source. Objective evaluation and informal listening tests show that the proposed method achieves separation quality similar to other existing methods while presenting significantly lower latency and computational cost. Another important result is that the artifacts error of the proposed method remains significantly higher than that of the high latency separation approach.

Real-time Bass Separation using Tikhonov Regularization The last contribution presented in the field of low-latency source separation is a technique to separate the bass source in real-time (Section 5.5). This method is similar to the previously presented technique for drums separation. In this case the basis matrix is adapted to better represent bass sources. The energy of the harmonic components of bass pitches is limited above a given cutoff frequency. Additionally the bass pitch selection makes use of the likelihood peak contrast in order to disambiguate between bass and other similar spectra. The method achieves separation results similar to a state-of-the-art high-latency method and significantly better than simply applying a low-pass filter. As in the case of the drums, there is still room for improvement in terms of artifact errors with respect to high latency methods.

High-latency Breathiness Estimation in Singing Voice Separation In the context of improving existing high-latency source separation methods our first contribution is a method to estimate and separate the breathiness component of the singing voice (Section 6.3). The spectral shape of the breathiness component is estimated using the harmonic envelope and an estimation of the glottal source based on the harmonic magnitudes. The relative energy of the breathiness with respect to the harmonic component of the voice is manually set with a parameter. The results show that integrating the estimated breathiness component into a state-of-the-art singing voice separation method improves the objective perceptual-related separation measures. This objective result is confirmed in informal listening tests. The proposed method can also be integrated into other source separation techniques, including low-latency approaches since only the pitch value of a single frame is needed.

High-latency Unvoiced Fricatives Estimation in Singing Voice Separation Another contribution in the context of high-latency source separation methods is the separation of unvoiced fricative consonants (Section 6.4). To account for unvoiced fricatives in the singing voice we propose using semi-supervised NMF in conjunction with the existing SIMM method. The basis components to represent fricatives are learned from isolated audio recordings. We propose transient-based NMF constraints to disambiguate between fricatives and similar percussive sources in the mixture. The method slightly improves the separation of the singing voice in perceptual-oriented objective tests and in listening to the results we notice that many of the fricatives are conserved in the isolated signal.

High-latency Drums Separation using NMF Regularizations and Transient / Annotation-based Constraints The other instrument considered in the high-latency separation scenario is percussion. We present two different methods that extend SIMM to separate drums. The first method uses a temporal continuity NMF regularization on the accompaniment gains to separate them from the drums components. Furthermore we test using temporal sparsity and frequency smoothness regularizations for the drums gains and basis respectively. The second method is based on NMF constraints derived from estimated transients or human annotations. The use of regularizations for the drums does not show any overall performance improvement over simply regularizing the accompaniment gains, but we observe that this behavior is highly dependent on the analyzed signal.

The evaluation conducted on the second method shows that using manual annotations of the drum positions significantly improves separation performance with respect to existing methods.

Study the Effect of Multiple Source-Filter Models in the High-latency SIMM-based Separation Finally we extend the SIMM method to perform separation of mixtures with multiple monophonic sources in a pitch-informed scenario (Section 6.6). We propose three different approaches. The first is the original SIMM with a single source-filter model. The second is a model with multiple sources but a single set of filters for all. The third assigns each source in the mixture an independent source-filter model. The contribution is a study of the performance differences in separation between the three methods. The results show that good results can be obtained with accurate pitch information. Another observation is that the improvement in separation results is more noticeable in music signals than in speech, probably due to harmonicity.

7.2 Future work

There is growing interest in low-latency and computationally inexpensive source separation methods in the context of music processing. We have shown that by making certain assumptions and imposing some restrictions we were able to significantly lower the latency and computational cost of existing methods. There is still a long way to go in terms of improving both the performance and the latency of music source separation. One possible future direction is to use prediction models, which are being widely studied in the field of music cognition (Hazan et al., 2009; Marchini and Purwins, 2010). These models could exploit the history of the music signal in order to improve the separation of the current frame.

The focus in the high-latency scenario is shifting towards exploiting user guidance or assistance to improve the separation. This research direction centers on minimizing human effort while maximizing the acquisition of information useful for the separation process. This research path is gaining traction with the arrival of new HCI techniques, such as multi-touch screens, depth cameras and Brain Computer Interaction (BCI) apparatus.

Furthermore, the fact that the role of the human is becoming more important in source separation scenarios will motivate even further the development of low-latency techniques. Rich interaction usually requires an

immediate response of the system with a preview of the expected result. In the spirit of *What You See Is What You Get* (WYSIWYG) editing systems in the graphics and text worlds we could imagine a *What You Hear Is What You Get* (WYHIWYG) source separation paradigm. We could also imagine the use of Brain Computer Interaction technologies to develop systems where the user guides the source separation process towards isolating the audio components on which he is focusing. These developments will require research in combined low-latency and user-guided source separation techniques.

More specifically many of the topics covered in this dissertation can be further developed and studied in order to improve audio source separation in both low-latency and high-latency scenarios. We here list several possible future directions of this work:

Theoretical Analysis of How NMF Compares To Tikhonov Regularization solutions Given the extensive work done in NMF and the promising results of applying Tikhonov regularization to spectral decomposition, pitch estimation and source separation, it would be useful to understand how these two solutions compare from a theoretical point of view. In Section 5.3 we compare the two factorization methods in terms of reconstruction error and separation capability of the predominant harmonic source. In this study only the standard NMF algorithm is considered, however an interesting future study could also consider NMF with a sparsity regularization term, since its objective function is more similar to that of Tikhonov regularization. Furthermore, a theoretical analysis could allow us to better understand how these results extrapolate to other basis matrices and signals.

Different Tikhonov Matrix Regularizations Throughout this dissertation we present several instantiations of the Tikhonov solution with different regularization terms. In its simplest form the Tikhonov regularization selects solutions with minimum norm. A more refined version regularizes the factorization using the norms of the components normalized by their energy. Finally, in a more advanced approach we propose using as regularization, the correlation between the pitch candidates, reducing coactivations of pitches with harmonic relations (e.g. octaves and fifths). There is still a lot of room for improvement in this direction. One idea is to perform training on the covariance matrix of known sources and use this as regularization. In a similar manner we could perform a weighting on the spectral

bins to give more importance to those where the target instrument has a higher probability of having energy. However, one must note that these regularizations must be independent of the signal in order to maintain the computational cost of factorizing each spectral frame low.

Parameter Study of the Filtered Harmonic Candidates Basis In Section 5.3 we propose a novel harmonic basis matrix composed of filtered pitch candidates that permits correct reconstruction of multiple different timbres. The tests are conducted with a single Mel-scale filterbank that tries to accommodate simultaneously the whole range of targeted instruments. However significant improvements could be made in terms of reconstruction and interpretability of the decomposition with filterbanks that are better adapted to the analyzed signals. In the field of singing voice one could use filterbanks that are specifically built to better reconstruct the formant structure of the singing voice. For the bass we could develop filterbanks that only cover the low and mid frequency range.

Evaluation and Improvements for the Low-latency Multipitch Estimation and Tracking The work we present in low-latency multi-pitch estimation is highly preliminary and only intends to demonstrate that Tikhonov regularization spectral decomposition can be useful for that specific task under low-latency constraints. The first subjective results are promising and an objective and comparative evaluation must be conducted. Given the high number of parameters involved this evaluation is not straight forward, and an extensive parameter grid search should be performed to achieve optimal results. In terms of the proposed pitch contour selection features, the octave error method should be tested with pitches maintaining other harmonic relations, such as fifths and thirds. This method could also be integrated into other multipitch estimation approaches. Furthermore the development of new features, such as pitch contour correlations, could lead to more accurate pitch estimation results.

Improvements to Low-latency Source Separation In the low-latency source separation side there is still much space for improvement and optimization. The proposed Wiener filtering separation method for the singing voice (Section 5.3) should be integrated with the low-latency predominant pitch estimation method that we propose and compared with existing techniques. Another quality improvement could be achieved by performing a more advanced selection of the pitch candidates when reconstructing the

target spectra. Currently the target pitch is selected by binary masking the gains resulting from the Tikhonov regularization decomposition. We could use a technique similar to that presented in Section 4.4, where the peaks are modeled using Gaussian functions, leading to a smoother masking of the gains and probably more accurate separation. Another direction for future research is the optimization of the Tikhonov regularization spectral decomposition when the basis matrix is large. In the case of singing voice separation, we use filtered harmonic components for a large range of pitches, this leads to a high number of basis components and therefore a large resolution matrix. Many coefficients of the basis are close to 0 and by allowing a certain amount of error many operations could be avoided while lowering the computational cost of the method.

Non-harmonic Components in Singing Voice Separation An obvious improvement in the estimation of the singing voice breathiness component is to estimate the gain automatically from the mixture. We foresee two main possible approaches to achieve this. One may exploit the regularity of the background music to find the gain using the difference between voiced and unvoiced frames. Another option is to estimate the gain making use of the fact that the breathiness is an amplitude modulated noise with a modulation rate equal to the pitch of the singing voice (Mehta and Quatieri, 2005). There are also many singing voice components that remain to be addressed. In Section 6.4 we propose a method to tackle unvoiced fricative consonants, which we also try to apply without success to other consonants such as plosives and trills. These consonants may require a special treatment because their timbral evolution in time is very diverse and instantaneous characterization of their spectrum may not be sufficient. More work on these and other phonemes will certainly lead to a more accurate separation of the singing voice.

Advances in Constraint-based Drums Separation The proposed regularization-based approach to drums separation (Section 6.5) does not provide any improvement over existing methods. The constraints-based method (Section 6.5) shows that accurate drum event positions can be used to achieve very good separation results. An obvious followup of this work is to automatically estimate the drum positions from the mixture. There is extensive research in the MIR community addressing such a task, and it could be used directly for the separation of drums. We show that with our specific configuration, the classification of the different drums does not

necessarily improve the separation, and may therefore not be necessary. However in the case where each of the percussive instruments may have especially adapted constraint parameters, this classification would probably lead to an improvement. Another future research direction might be to split the method into two steps, a first step dedicated to learning the drums basis from the mixture and a second step in which the basis is set constant and only the gains are learned. This would allow considering only drum positions with high confidence for the basis learning step which could lead to a better separation and might better tolerate errors in the drum position estimation.

7.3 Outcomes

Several methods and experiments presented in this dissertation have been published in the following scientific articles:

– R. Marxer, J. Janer, and J. Bonada. Low-Latency instrument separation in polyphonic audio using timbre models. *Latent Variable Analysis and Signal Separation*, pages 314–321, 2012
  This article presents a method for the removal of the singing voice based on time-frequency binary masks resulting from the combination of azimuth, phase difference and absolute frequency spectral bin classification and harmonic-derived masks as described in Section 5.2. For the harmonic-derived masks it uses the pitch likelihood estimation technique based on Tikhonov regularization and supervised timbre models from Section 4.3.
– R. Marxer and J. Janer. A Tikhonov regularization method for spectrum decomposition in low latency audio source separation. In *Proceedings IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP'2012)*, March 2012
  The comparison between the NMF and Tikhonov regularization spectral decomposition methods (see Section 5.3) for the singing voice separation task was presented in this article.
– J. Janer, R. Marxer, and K. Arimoto. Combining a harmonic-based NMF decomposition with transient analysis for instantaneous percussion separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 281–284. IEEE, 2012
  This work presents the use of NMF spectral decomposition in the task of drums separation. This study is similar to the one presented in Section 5.4, where Tikhonov regularization spectral decomposition is used

instead.

– J. Bosch, K. Kondo, R. Marxer, and J. Janer. Score-informed and timbre independent lead instrument separation in real-world scenarios. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 2417–2421, aug. 2012
  This article shows how score information can be integrated into the low-latency pitch estimation technique presented in Section 4.3.
– R. Marxer and J. Janer. Realtime Bass Separation using Harmonic-Percussion Decomposition. In *Proc. (DAFx) International Conference on Digital Audio Effects*, Dublin, Ireland, 2013a. (accepted)
  This paper covers the low-latency real-time bass separation method presented in Section 5.5.
– R. Marxer and J. Janer. Modelling and Separation of Singing Voice Breathiness in Polyphonic Mixtures. In *Proc. (DAFx) International Conference on Digital Audio Effects*, Dublin, Ireland, 2013b. (accepted)
  The singing voice breathiness component estimation and its integration into a source separation system found in Section 6.3 is presented in this study.
– J. Janer and R. Marxer. NMF-based Separation of Unvoiced Fricatives in Singing Voice Music Mixtures. In *Proc. (DAFx) International Conference on Digital Audio Effects*, Dublin, Ireland, 2013. (accepted)
  This paper presents the estimation of fricative components using semi-supervised NMF and its use in the task of singing voice isolation (see Section 6.4).
– R. Marxer and J. Janer. Use of regularizations and constraints in NMF-based drums monaural separation. In *Proc. (DAFx) International Conference on Digital Audio Effects*, Dublin, Ireland, 2013c. (accepted)
  This article presents the drums separation study from Section 6.5.

Additionally the following patents are related to the research presented in this thesis:

– Y. Umeyama, K. Kondo, Y. Takahashi, J. Bonada, J. Janer, and R. Marxer. Graphical Audio Signal Control, February 7 2012. US Patent App. 13/367,696
  This patent is related to the work described in Section 5.2, where binary masks based on spectral bin features are used in combination with human-controlled parameters.
– J. Bonada, J. Janer, R. Marxer, Y. Umeyama, K. Kondo, and F. Garcia. Technique for Estimating Particular Audio Component, May 3 2012b. US Patent 20,120,106,746
  This patent relates to the method of low-latency target source predominant pitch estimation and tracking using timbre models. Part of the

research leading to this patent is found in Section 4.3 of this dissertation.

– J. Bonada, J. Janer, R. Marxer, Y. Umeyama, and K. Kondo. Technique for Suppressing Particular Audio Component, May 3 2012a. US Patent 20,120,106,758

  This patent covers some of the separation methods that we propose in Section 5.3.

Furthermore Yamaha Corp. has applied for three more patents related to the research conducted and presented in this dissertation. At time of printing the details about these three patent applications cannot be disclosed.

Part I

Appendix

Appendix A

Other Signal Representations

Several methods have been proposed to overcome the problem of finding a good tradeoff between frequency and temporal localization in signal representations. In this Appendix we review some of the methods that are specially relevant to the analysis and processing of music signals.

Multiresolution Dressler (2006) proposes a fast multiresolution STFT that uses a different analysis window for each frequency band, larger windows for low frequency bands and smaller windows for high frequency ranges:

$$s[n] = \sum_{t=0}^{M-1} \sum_{\omega=0}^{N-1} c_{t,\omega}^{mstft} w_\omega[n - tH] e^{i2\pi \frac{\omega}{N} n} \tag{A.1}$$

where $w_\omega[n] \quad for \quad n \in [0, L-1]$ are the windowing functions of length L and H is the hop size of the windowing function, with different magnitudes depending on the frequency index ω. The expansion coefficients are simply computed as:

$$c_{t,\omega}^{mstft} \quad = \quad \frac{1}{N} \sum_{\omega=0}^{L-1} s[n - tH] w_\omega[n] e^{-i2\pi \frac{\omega}{N} n} \tag{A.2}$$

The multiresolution STFT has a much higher computational cost due to using different windowing for the different frequency bands. However an approximation has been proposed where the different windows are applied in the frequency domain by the use of a convolution. Since we normally use windows that have a large main lobe and very low side lobes, we can

approximate the DFT of the window by setting to zero all the coefficients that are far away from the center. This approximation significantly reduces the computational cost and has the additional advantage of being able to use the well-established FFT and perform one single transform of the signal per frame, on the other hand the FFT performed must be the size of the largest window in the multiresolution analysis.

$$c^{r}_{t,\omega} = \frac{1}{N}\sum_{\omega=0}^{L-1} s[n-tH]e^{-i2\pi\frac{\omega}{N}n} \tag{A.3}$$

$$c^{mstft}_{t,\omega} = c^{r}_{t,\omega} * W_{\omega} \tag{A.4}$$

where $W_{\omega}[\omega]$ is the DFT transform or the approximation of the DFT of the window used for bin ω.

Other multiresolution techniques (Zhou et al., 2009) may also reduce the computational cost in exchange for complexity, using recursive decimation and filtering followed by analysis.

While the multiresolution STFT allows modifying the frequency-temporal resolution depending on the frequency, the frequency scale of the coefficients still lies on a linear axis in Hertz. Numerous studies have proposed the use of different frequency scales (Mel, Bark, Constant-Q,...) that are more adapted to human auditory perception characteristics (Brown, 1991).

Matching Pursuit All of the previous signal representations are based on a previously known and fixed set of basis functions. For the waveform the basis functions are shifted diracs and for the DFT and STFT they are complex exponentials. Another family of representations called adaptive basis representations represents a signal by using a basis that is adapted to the specific signal being analyzed.

One of the most popular algorithms used to calculate such representations is the Matching Pursuit (MP) (Mallat and Zhang, 1993) technique. It is a greedy iterative solution that starts with the signal and at each step matches the residual of the previous step to all the functions in a dictionary $g_i[n] \in \boldsymbol{D}$ of possible basis functions in search of the maximally matching function. Then it removes the match from the residual leading to a new residual. This can be expressed as:

$\omega = 0$
$r_\omega[n] = s[n]$
while continue criterion **do**
  $g_{match}[n] = \arg\max_{g_i \in \boldsymbol{D}}(\|\langle r_\omega[n], g_i[n]\rangle\|)$
  $\boldsymbol{b}_\omega = g_{match}[n]$
  $c_\omega^{mp} = \langle r_\omega[n], g_{match}[n]\rangle$
  $r_{\omega+1}[n] = r_\omega[n] - c_\omega^{mp}\boldsymbol{b}_\omega$
  $\omega = \omega + 1$
end while

where the continue criterion can take several forms depending on the application (e.g. $\|r_\omega[n]\| >= threshold$). $\langle \boldsymbol{x}, \boldsymbol{y}\rangle$ denotes the inner product. The basis $\boldsymbol{B}$ of the MP representation is a subset of the dictionary $\boldsymbol{D}$. This can also be seen as $\boldsymbol{D}$ being the basis, with 0 coefficients for the elements in $\boldsymbol{D}$ which are not in $\boldsymbol{B}$. The computation of the inner product $\langle r_\omega[n], g_i[n]\rangle$ can be done recursively using previously calculated $\langle g_i[n], g_j[n]\rangle$.

An extension to this technique named Orthogonal Matching Pursuit (OMP) (Pati et al., 1993; Davis et al., 1994) leads to faster convergence when using non-orthogonal dictionaries D. The difference between OMP and MP is that at each step the coefficients c_ω^{omp} are updated in order to keep the residual orthogonal to the subspace spanned by $\boldsymbol{b}_i | \forall i \in 1 \ldots \omega$. The update of c_ω^{omp} can then be performed recursively using previously calculated inner products of the basis functions.

In all these cases there is no discussion about how to determine a good dictionary D. However some studies have targeted the problem of dictionary learning and of how these learned dictionaries relate to other known and established bases. In the context of audio Smith and Lewicki (2006) compared the basis learned using an MP technique to the gammatone filters that play an important role in the modeling of human and animal auditory perception.

Empirical Mode Decomposition (EMD) Another popular technique for deriving a time frequency representation of the signal using an adaptive basis is known as Empirical Mode Decomposition (EMD). EMD is a specific case of adaptive basis representation in which the basis is derived directly from the signal being analyzed. These methods are often known as data-driven or data-fusion transformations. The EMD method earned its popularity due to the fact that the Intrinsic Mode Function (IMF) com-

ponents into which the signals are decomposed are well suited for Hilbert Spectral Analysis (HSA). The HSA of the IMF leads to an easy estimation of the instantaneous frequencies of the individual components. The coupling of these two methods was proposed by Huang et al. (1996, 1998, 1999) and was named the Hilbert-Huang Transform (HHT). The HHT is especially interesting for the analysis of non-stationary and non-linear signals.

The EMD method consists in finding the set of IMFs that compose a given signal. An IMF is defined as a function with an equal (or different by 1) number of extrema and zero-crossings and with a symmetric extrema envelope. In other words, the mean between the envelope defined by the local maxima and the one defined by the local minima must be 0 or close to 0. An IMF component is similar to a harmonic in a DFT or STFT, however instead of having a constant amplitude and frequency, its amplitude and frequency can vary as a function of time.

The algorithm used to extract the IMF components can be described using an operation named *sifting* and defined as follows:

$\omega = 0$
$r_\omega[n] = s[n]$
while continue decomposition criterion **do**
  $i = 0$
  $g^i_\omega[n] = r_\omega[n]$
  while continue *sifting* criterion **do**
    Identify the local extrema of $g^i_\omega[n]$
    Estimate a curve $e_{up}[n]$ for the support of $g^i_\omega[n]$ connecting the local maxima
    Estimate a curve $e_{lo}[n]$ for the support of $g^i_\omega[n]$ connecting the local minima
    $g^{i+1}_\omega[n] = g^i_\omega[n] - \frac{(e_{up}[n]+e_{lo}[n])}{2}$
    $i = i + 1$
  end while
  $\boldsymbol{b}_\omega = g^{i+1}_\omega[n]$
  $r_{\omega+1}[n] = r_\omega[n] - \boldsymbol{b}_\omega$
  $\omega = \omega + 1$
end while

We note that given this representation, the coefficients c^{emd}_ω can be considered equal to 1. Similarly we can normalize the IMF components $\boldsymbol{b}_\omega$ and

then the normalization factor becomes c_ω^{emd}. The first criterion for continuing the *sifting*, proposed by Huang et al. (1998), is similar to a Cauchy convergence test, and consists in a threshold on the sum of differences:

$$SD_\omega = \frac{\sum_n \left|g_\omega^{i+1}[n] - g_\omega^i[n]\right|^2}{\sum_n \left|g_\omega^i[n]\right|^2} \tag{A.5}$$

Applying a Hilbert filter on the IMF components gives the following expression for the signal:

$$s(t) = \Re\left(\sum_\omega a_\omega[n] e^{j\Phi_\omega[n]}\right) \tag{A.6}$$

where $a_\omega[n]$ is the amplitude envelope of the IMF component ω and $\Phi_\omega[n]$ is the instantaneous phase.

From this expression it is easy to estimate the instantaneous frequency as the time derivative of the unwrapped instantaneous phase.

Given the empirical and algorithmic nature of this representation there has been significant research into explaining it's properties mathematically. Wu and Huang (2004) find empirically that EMD is in practice a dyadic filter. Flandrin et al. (2004) relate the method to adaptive constant-Q filter banks. Rilling et al. (2003) show different variations of the EMD standard algorithm. The authors compare different stopping criteria as well as an online version of the EMD algorithm, which performs the transformation using a limited amount of previous data. The results from the study also support the interpretation of the representation in terms of adaptive constant-Q filter banks.

Instantaneous frequency / Spectral reassignment Another type of representation which has long been useful in the field of audio and music analysis is the family of spectrum reassignment transforms (Auger and Flandrin, 1995). These methods derive from the interpretation of the STFT as a filterbank. Each output of a filter can be represented in polar coordinates as:

$$\boldsymbol{S}(\omega, t) = e^{a(\omega,t)+j\Phi(\omega,t)} \tag{A.7}$$

where $a(\omega, t)$ is the instantaneous amplitude and $\Phi(\omega, t)$ the instantaneous phase. Their values can be isolated and expressed as:

$$a(\omega, t) = \Re(log(\boldsymbol{S}(\omega, t))) \tag{A.8}$$

$$\Phi(\omega, t) = \Im(log(\boldsymbol{S}(\omega, t))) \tag{A.9}$$

From these expressions the instantaneous frequency can be estimated as:

$$\hat{\omega}(\omega, t) = \frac{\partial}{\partial t}\Phi(\omega, t) = \Im\left(\frac{\frac{\partial \boldsymbol{S}}{\partial t}}{\boldsymbol{S}}\right) \tag{A.10}$$

It is easy to show that:

$$\frac{\partial \boldsymbol{S}}{\partial t} = -\boldsymbol{S}_{w'} + j\omega\boldsymbol{S} \tag{A.11}$$

where $\boldsymbol{S}_{w'}$ is the DFT of our signal using the derivative of the window used for the computation of $\boldsymbol{S}$. Therefore the instantaneous frequency can be expressed as:

$$\hat{\omega}(\omega, t) = \omega - \Im\left(\frac{\boldsymbol{S}_{w'}}{\boldsymbol{S}}\right) \tag{A.12}$$

In addition to the instantaneous frequency, other Frequency Modulation (FM) parameters can be derived in the same manner by the use of higher order derivatives (Musevic and Bonada, 2010). We may also estimate the Amplitude Modulation (AM) parameters using the same technique. In that case the group delay $\hat{t}(\omega, t)$ can be computed as:

$$\hat{t}(\omega, t) = t + \Re\left(\frac{\boldsymbol{S}_{tw}}{\boldsymbol{S}}\right) \tag{A.13}$$

where $\boldsymbol{S}_{tw}$ is defined as the DFT of our signal using the window multiplied by the time function.

Once these parameters have been estimated we may use them in order to construct spectral representations better tailored for our needs. Several derived representations have been proposed over the years. Probably the most popular of them all is the reassigned spectrum. This representation consists in reassigning the magnitude or energy from each bin of the $\boldsymbol{S}$ to the frequency position corresponding to the instantaneous frequency. This reassignment procedure can be also be performed in the temporal domain using the estimation of the group delay $\hat{t}(\omega, t)$. Our reassigned spectrum can be then defined as:

$$\hat{\boldsymbol{S}}(\hat{\omega}(\omega, t), \hat{t}(\omega, t)) = \sum_{\forall \omega\, t} \|\boldsymbol{S}(\omega, t)\| \tag{A.14}$$

Abe et al. (1997) proposed another representation based on the instantaneous frequency called the IF spectrum which is a modified amplitude spectrum:

$$c_{\omega,t}^{ifspec} = \frac{1}{\Delta\hat{\omega}} \int_{\hat{\omega}<\hat{\omega}(\omega,t)<\hat{\omega}+\Delta\hat{\omega}} \|\boldsymbol{S}(\omega,t)\| d\omega \tag{A.15}$$

Dominance Nakatani and Irino (2004) proposed a measure of *degree of dominance* and a spectrum based on it called the *dominance spectrum*. The *degree of dominance* $D(\omega,t)$ represents the magnitude of a harmonic component relative to other components in each bin. This measure is computed as:

$$c_{\omega,t}^{dominance} = 10\log_{10}(1/B^2(\omega,t)) \tag{A.16}$$

$$B^2(\omega,t) = \frac{\int_{\omega-\Delta\omega/2}^{\omega+\Delta\omega/2} (\hat{\omega}(\omega',t)-\omega)^2 \boldsymbol{S}(\omega',t)^2 d\omega'}{\int_{\omega-\Delta\omega/2}^{\omega+\Delta\omega/2} \boldsymbol{S}(\omega',t)^2 d\omega'} \tag{A.17}$$

where $B^2(\omega,t)$ is a local average of the deviation of each bin's frequency from the instantaneous frequency weighted by the spectrum's energy. This measure is low for bins where the instantaneous frequency is close to the bin's frequency.

Bibliography

Each reference indicates the pages where it appears.

S. A. Abdallah and M. D. Plumbley. Polyphonic music transcription by non-negative sparse coding of power spectra. In *Proc. 5th Int'l Conf. on Music Information Retrieval (ISMIR)*, pages 10–14, 2004. 73

T. Abe, T. Kobayashi, and S. Imai. The IF spectrogram: A new spectral representation. *Proc. ASVA 97*, pages 423–430, 1997. 234

H. Akima. A new method of interpolation and smooth curve fitting based on local procedures. *JACM*, 17(4):589–602, 1970. 103, 174

S. Amari. Differential-geometrical methods in statistic. 1985. 62

F. Asano, H. Asoh, and T. Matsui. Sound source localization and signal separation for office robot "JiJo-2". pages 243–248. IEEE, 1999. ISBN 0-7803-5801-5. doi: 10.1109/MFI.1999.815997. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=815997. 6, 52

F. Asano, M. Goto, K. Itou, and H. Asoh. Real-time sound source localization and separation system and its application to automatic speech recognition. In *Seventh European Conference on Speech Communication and Technology*, number 8, pages 1–4. Citeseer, 2001. URL http://www.isca-speech.org/archive/eurospeech_2001/e01_1013.html. 8, 52

J.-J. Aucouturier, B. Defreville, and F. Pachet. The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *The Journal of the Acoustical Society of America*, 122(2):881–891, 2007. doi: 10.1121/1.2750160. URL http://link.aip.org/link/?JAS/122/881/1. 6

F. Auger and P. Flandrin. Improving the readability of time-frequency and

time-scale representations by the reassignment method. *Signal Processing, IEEE Transactions on*, 43(5):1068–1089, may 1995. ISSN 1053-587X. doi: 10.1109/78.382394. 57, 233

D. Barry. Drum source separation using percussive feature detection and spectral modulation. *IEE Irish Signals and Systems Conference*, pages 13–17(4), 2005. URL http://digital-library.theiet.org/content/conferences/10.1049/cp_20050280. 57, 189

Bell Anthony J. and Sejnowski Terrence J. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7(6):1129–1159, 1995. ISSN 0899-7667. doi: 10.1162/neco.1995.7.6.1129. doi: 10.1162/neco.1995.7.6.1129. 45

J. P. Bello and M. B. Sandler. Phase-Based Note Onset Detection For Music Signals. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,ICASSP*, 5(2):441–444, 2003. 57

J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler. A Tutorial on Onset Detection in Music Signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1–13, 2005. 56

A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *Signal Processing, IEEE Transactions on*, 45(2):434–444, feb 1997. ISSN 1053-587X. doi: 10.1109/78.554307. 44

L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):191–199, January 2006. ISSN 1558-7916. doi: 10.1109/TSA.2005.854110. 125, 145

E. Benetos and S. Dixon. Joint Multi-Pitch Detection Using Harmonic Envelope Estimation for Polyphonic Music Transcription. *Selected Topics in Signal Processing, IEEE Journal of*, 5(6):1111–1123, oct. 2011a. ISSN 1932-4553. doi: 10.1109/JSTSP.2011.2162394. 117

E. Benetos and S. Dixon. Multiple-instrument polyphonic music transcription using a convolutive probabilistic model. *8th Sound and Music Computing Conf*, pages 19–24, 2011b. 55

E. Benetos and S. Dixon. Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model. *The Journal of the Acoustical Society of America*, 133(3):1727–1741, 2013. doi: 10.1121/1.4790351. URL http://link.aip.org/link/?JAS/133/1727/1. 55

N. Bertin, R. Badeau, and E. Vincent. Enforcing Harmonicity and Smoothness in Bayesian Non-Negative Matrix Factorization Applied to Poly-

phonic Music Transcription. *IEEE Transactions on Audio, Speech & Language Processing*, 18(3):538–549, 2010. URL http://dx.doi.org/10.1109/TASL.2010.2041381. 65

J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. Mit Press, 1983. ISBN 9780262021906. URL http://books.google.es/books?id=rrC5QgAACAAJ. 6

J. Bonada, J. Janer, R. Marxer, Y. Umeyama, and K. Kondo. Technique for Suppressing Particular Audio Component, May 3 2012a. US Patent 20,120,106,758.

J. Bonada, J. Janer, R. Marxer, Y. Umeyama, K. Kondo, and F. Garcia. Technique for Estimating Particular Audio Component, May 3 2012b. US Patent 20,120,106,746.

J. Bosch, K. Kondo, R. Marxer, and J. Janer. Score-informed and timbre independent lead instrument separation in real-world scenarios. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 2417–2421, aug. 2012. 59, 171

A. S. Bregman. *Auditory Scene Analysis*. MIT Press, Cambridge, MA, 1990. 3

P. Brossier, J. Bello, and M. Plumbley. Fast Labelling of Notes in Music Signals. *Proceedings of the 5th International Symposium on Music Information Retrieval*, pages 331–336, 2004. 57

J. Brown. Calculation of a constant Q spectral transform. *J. of the Acoustical Soc. of America*, 89(1):425–434, 1991. 230

J. J. Burred. *From Sparse Models to Timbre Learning: New Methods for Musical Source Separation*. PhD thesis, Technical University of Berlin, Berlin, Germany, September 2008. URL http://recherche.ircam.fr/equipes/analyse-synthese/burred/phd/index.html. 7, 22, 28

J. J. Burred and P. Leveau. Geometric multichannel common signal separation with application to music and effects extraction from film soundtracks. In *ICASSP*, pages 201–204, 2011. 9

D. Calvetti and E. Somersalo. *Introduction to Bayesian Scientific Computing*. Surveys and Tutorials in the Applied Mathematical Sciences. Springer Science+Business Media, 2008. ISBN 9780387733944. 98

E. Candes and M. Wakin. An Introduction To Compressive Sampling. *Signal Processing Magazine, IEEE*, 25(2):21–30, March 2008. ISSN 1053-5888. doi: 10.1109/MSP.2007.914731. 28

J. J. Carabias-Orti, T. Virtanen, P. Vera-Candeas, N. Ruiz-Reyes, and F. J. Cañadas-Quesada. Musical Instrument Sound Multi-Excitation Model for

Non-Negative Spectrogram Factorization. *J. Sel. Topics Signal Processing*, 5(6):1144–1158, 2011. URL http://dx.doi.org/10.1109/JSTSP.2011.2159700. 56, 167, 169, 205, 207

J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, oct 1998. ISSN 0018-9219. doi: 10.1109/5.720250. 45

G. Carter. Coherence and time delay estimation. *Proceedings of the IEEE*, 75(2):236–255, 1987. 47

J. L. Castellanos, S. Gómez, and V. Guerra. The triangle method for finding the corner of the L-curve. *Appl. Numer. Math.*, 43(4):359–373, December 2002. ISSN 0168-9274. doi: 10.1016/S0168-9274(01)00179-9. 98

Celemony Software GmbH. DNA Direct Note Access, 2009. URL http://www.celemony.com/cms/index.php?id=dna. 7, 8, 9

C. C. Chang and C. J. Lin. *LIBSVM: a library for support vector machines.* 2001. URL http://www.csie.ntu.edu.tw. 104

Z. Chen and A. Cichocki. Nonnegative matrix factorization with temporal smoothness and/or spatial decorrelation constraints. In *Laboratory for Advanced Brain Signal Processing, RIKEN, Tech. Rep*, 2005. 68

Z. Chen, A. Cichocki, and T. M. Rutkowski. Constrained non-negative matrix factorization method for EEG analysis in early detection of Alzheimer's disease. In *In: IEEE International Conference on Acoustics, Speech, and Signal Processing,, ICASSP2006*, pages 893–896, 2006. 68

E. C. Cherry. Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, 25(5):975–979, 1953. doi: 10.1121/1.1907229. URL http://link.aip.org/link/?JAS/25/975/1. 5

A. Cichocki, R. Zdunek, and S. Amari. New algorithms for non-negative matrix factorization in applications to blind source separation. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 5, pages V–V. Ieee, 2006. 62, 64, 67

A. Cichocki, H. Lee, Y. Kim, and S. Choi. Non-negative matrix factorization with α-divergence. *Pattern Recognition Letters*, 29(9):1433–1440, 2008. 62, 67

P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994. 4, 45

M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus

for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120:2421, 2006. 208

G. Davis, S. Mallat, and Z. Zhang. Adaptive Time-Frequency Approximations with Matching Pursuits. Technical report, New York University, New York, NY, USA, 1994. 231

A. de Cheveigné. Speech f0 extraction based on Licklider's pitch perception model. In *International Congress of Phonetic Sciences*, volume 4, pages 218–221, 1991. 55

A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111:1917, 2002. 55, 119, 122, 136, 206, 208

G. Degottex. *Glottal source and vocal tract separation*. PhD thesis, UPMC-Ircam-UMR9912-STMS, France, 2010. 40

G. Degottex, A. Roebel, and X. Rodet. Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5128–5131, may 2011. doi: 10.1109/ICASSP.2011.5947511. 40, 172, 179

P. Desain and H. Honing. Computational models of beat induction: The rule-based approach. *Journal of New Research Music*, 28:29–42, 1999. 58

A. Dessein, A. Cont, and G. Lemaitre. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *Proc. of ISMIR International Symposium for Music Information Retrieval*, 2010. 8

I. Dhillon and S. Sra. Generalized nonnegative matrix approximations with Bregman divergences. *Advances in neural information processing systems*, 18:283, 2006. 62, 63, 76, 81

L. Di Persia, M. Yanagida, H. L. Rufiner, and D. H. Milone. Objective quality evaluation in blind source separation for speech recognition in a real room. *Signal Processing*, 87(8):1951–1965, 2007. URL `http://fich.unl.edu.ar/sinc/publications/2007/DYRM07`. 6

Dixon. *Beat induction and rhythm recognition*. 1997. 58

S. Dixon. Onset Detection Revisited. *Int. Conference on Digital Audio Effects (DAFx-06)*, 2006. 57

B. Doval, C. d'Alessandro, and N. Henrich. The voice source as a causal/anticausal linear filter. In *PROC. ISCA ITRW VOQUAL'03*, pages 15–19, 2003. 40

J. S. Downie, K. West, A. F. Ehmann, and E. Vincent. The 2005 Music Information retrieval Evaluation Exchange (MIREX 2005): Preliminary Overview. In *ISMIR*, pages 320–323, 2005. 107

K. Dressler. Extraction of the melody pitch contour from polyphonic audio. In *Proc. 6th International Conference on Music Information Retrieval*, 2005. 108

K. Dressler. Sinusoidal Extraction Using an Efficient Implementation of a Multi-Resolution FFT. *Proc. of the 9th Int. Conference on Digital Audio Effects*, pages 247–252, 2006. 229

Z. Duan, B. Pardo, and C. Zhang. Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(8):2121–2133, 2010. 122, 123

Z. Duan, G. J. Mysore, and P. Smaragdis. Online PLCA for Real-Time Semi-supervised Source Separation. In F. J. Theis, A. Cichocki, A. Yeredor, and M. Zibulevsky, editors, *LVA/ICA*, volume 7191 of *Lecture Notes in Computer Science*, pages 34–41. Springer, 2012. ISBN 978-3-642-28550-9. URL http://dx.doi.org/10.1007/978-3-642-28551-6_5. 8, 36

J.-L. Durrieu and J.-P. Thiran. Musical audio source separation based on user-selected f0 track. In *Proceedings of the 10th international conference on Latent Variable Analysis and Signal Separation*, LVA/ICA'12, pages 438–445, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-28550-9. doi: 10.1007/978-3-642-28551-6_54. URL http://dx.doi.org/10.1007/978-3-642-28551-6_54. 167, 169, 171, 205

J. Durrieu, G. Richard, and B. David. An iterative approach to monaural musical mixture de-soloing. In *IEEE Int. Conf. on Acoustics, Speech & Signal Processing (ICASSP)*, pages 105–108, 2009a. 136, 182

J. Durrieu, G. Richard, B. David, and C. Févotte. Source/Filter Model for Unsupervised Main Melody Extraction From Polyphonic Audio Signals. *IEEE Transactions on Audio, Speech & Language Processing*, 18(3):564–575, March 2010. ISSN 1063-6676. doi: 10.1109/TASL.2010.2041114. 40, 49, 79, 108, 117, 126, 128, 129, 132, 133

J.-L. Durrieu, B. David, and G. Richard. A Musically Motivated Mid-Level Representation for Pitch Estimation and Musical Audio Source Separation. *J. Sel. Topics Signal Processing*, 5(6):1180–1191, 2011. URL http://dx.doi.org/10.1109/JSTSP.2011.2158801. 56, 74, 79, 96, 169, 173, 175, 181, 206, 212

J. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David. Main instrument separation from stereophonic audio signals using a source/filter model. *Proc. European Signal Processing*, 2009b. 74, 79, 86, 96, 141, 167, 168, 169, 205

C. Duxbury, M. Sandler, and M. Davies. A hybrid approach to musical note onset detection. *Int. Conference on Digital Audio Effects (DAFx-02)*, pages 33–38, 2002. 56

C. Duxbury, J. Bello, M. Davies, and M. Sandler. Complex domain onset detection for musical signals. *Proceedings Digital Audio Effects Workshop (DAFx)*, 2003. URL http://citeseer.ist.psu.edu/duxbury03complex.html. 57

J. Eggert and E. Korner. Sparse coding and NMF. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, volume 4, pages 2529–2533. Ieee, 2004. 66

V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. Subjective and Objective Quality Assessment of Audio Source Separation. *IEEE Transactions on Audio, Speech & Language Processing*, 19(7):2046–2057, 2011. URL http://dx.doi.org/10.1109/TASL.2011.2109381. 90

M. Every. *Separation of Musical Sources and Structure from Single-Channel Polyphonic Recordings*. PhD thesis, University of York, UK, February 2006. URL http://www.ee.surrey.ac.uk/Personal/M.Every/EveryPhD06.pdf. 117

S. Ewert and M. Müller. Using score-informed constraints for NMF-based source separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 129–132. IEEE, 2012. 59, 74, 167, 169, 171, 205

S. Ewert and M. Müller. Score-Informed Voice Separation For Piano Recordings. In A. Klapuri and C. Leider, editors, *ISMIR*, pages 245–250. University of Miami, 2011. ISBN 978-0-615-54865-4. URL http://www.bibsonomy.org/bibtex/2ce688e18fcfa84ddd85b0273a8185c8d/dblp. 59, 74, 182, 190, 193

G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STLQPSR*, 4(4):1–13, 1985. 40

G. Fant. The LF-model revisited. Transformations and frequency domain analysis. *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, 2: 3, 1995. 40

T. Feng, S. Li, H.-Y. Shum, and H. Zhang. Local Non-Negative Matrix Factorization as a Visual Representation. In *Proceedings of the 2nd Interna-*

tional Conference on Development and Learning, ICDL '02, pages 178–, Washington, DC, USA, 2002. IEEE Computer Society. ISBN 0-7695-1459-6. URL http://dl.acm.org/citation.cfm?id=876897.881034. 65

C. Févotte, N. Bertin, and J. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Comput.*, 21(3):793–830, March 2009. ISSN 0899-7667. doi: 10.1162/neco.2008.04-08-771. 64, 65, 82, 83, 126, 133, 134

D. Fitzgerald. Upmixing from mono - A source separation approach. In *Digital Signal Processing (DSP), 2011 17th International Conference on*, pages 1–7, 2011. doi: 10.1109/ICDSP.2011.6004991. 7

D. FitzGerald, M. Cranitch, and E. Coyle. Extended nonnegative tensor factorisation models for musical sound source separation. *Computational Intelligence and Neuroscience*, 2008, 2008. 77

D. FitzGerald. *Automatic Drum Transcription and Source Separation*. PhD thesis, Dublin Institute of Technology, Dublin, Ireland, 2004. URL http://homepage.eircom.net;/derryfitzgerald/ThesisFitz.pdf. 41

D. FitzGerald, M. Cranitch, and E. Coyle. Non-negative tensor factorisation for sound source separation. In *PROCEEDINGS OF IRISH SIGNALS AND SYSTEMS CONFERENCE*, 2005. 151

D. FitzGerald, M. Cranitch, and E. Coyle. On the use of the beta divergence for musical source separation. In *Signals and Systems Conference (ISSC 2009), IET Irish*, pages 1–6, june 2009. doi: 10.1049/cp.2009.1711. 64

J. Flanagan. *Speech analysis; synthesis and perception*. Kommunikation und Kybernetik in Einzeldarstellungen. Springer-Verlag, 1972. ISBN 9780387055619. 40

P. Flandrin, G. Rilling, and P. Goncalves. Empirical Mode Decomposition as a Filter Bank. *IEEE Signal Processing Letters*, 11:112–114, February 2004. doi: 10.1109/LSP.2003.821662. 233

M. Foster. An Application of the Wiener-Kolmogorov Smoothing Theory to Matrix Inversion. *Journal of the Society for Industrial and Applied Mathematics*, 9(3):pp. 387–392, 1961. ISSN 03684245. URL http://www.jstor.org/stable/2099031. 97

F. Fuhrmann. *Automatic musical instrument recognition from polyphonic music audio signals*. PhD thesis, Universitat Pompeu Fabra, 2012. 33

F. Fuhrmann and P. Herrera. Quantifying the relevance of locally extracted information for musical instrument recognition from entire pieces of music. In *International Society for Music Information Retrieval Conference (ISMIR)*, Miami, Florida, USA, 24/10/2011 2011. 33

H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. Okuno. F0 Estimation Method for Singing Voice in Polyphonic Audio Signal Based on Statistical Vocal Model and Viterbi Search. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, page V, May 2006. doi: 10.1109/ICASSP.1661260. 125, 127

O. Fujimura and J. Lindqvist. Sweep-Tone Measurements of Vocal-Tract Characteristics. *The Journal of the Acoustical Society of America*, 49 (2B):541–558, 1971. doi: 10.1121/1.1912385. 40

H. Fujisaki and M. Ljungqvist. Proposal and evaluation of models for the glottal source waveform. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.*, volume 11, pages 1605–1608, 1986. doi: 10.1109/ICASSP.1986.1169239. 40

J. Ganseman, P. Scheunders, G. J. Mysore, and J. S. Abel. Source separation by score synthesis. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 462–465, 2010. 59

E. Gaussier and C. Goutte. Relation between PLSA and NMF and implications. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 601–602, New York, NY, USA, 2005. ACM. ISBN 1-59593-034-5. URL http://www.bibsonomy.org/bibtex/2669b75d12f9b188fe85a16b09489fb7c/folke. 84, 85

D. Giannoulis et al. A Database and Challenge for Acoustic Scene Classification and Event Detection. 2013. 6

O. Gillet and G. Richard. Automatic transcription of drum loops. *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, 4:iv–269–iv–272 vol.4, May 2004. ISSN 1520-6149. doi: 10.1109/ICASSP.2004.1326815. 41, 57

O. Gillet and G. Richard. Extraction and remixing of drum tracks from polyphonic music signals. In *Applications of Signal Processing to Audio and Acoustics. IEEE Workshop on*, pages 315–318, 2005. doi: 10.1109/ASPAA.2005.1540232. 189

O. Gillet and G. Richard. Transcription and separation of drum signals from polyphonic music. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(3):529–540, 2008. 57, 142

O. Gillet. *Transcription des signaux percussifs. Application ï¿½ l'analyse de scï¿½nes musicales audiovisuelles.* PhD thesis, Telecom ParisTech - ENST, Paris, France, June 2007. URL http://pastel.paristech.org/2805. 57

G. Golub. Numerical methods for solving linear least squares problems. *Numerische Mathematik*, 7(3):206–216, 1965. 97

E. Gómez, F. J. Cañadas Quesada, J. Salamon, J. Bonada, P. Vera Candea, and P. Cabañas Molero. Predominant Fundamental Frequency Estimation vs Singing Voice Separation for the Automatic Transcription of Accompanied Flamenco Singing. In *Proceedings of the 13th International Society for Music Information Retrieval Conference*, Porto, Portugal, October 8-12 2012. http://ismir2012.ismir.net/event/papers/601-ismir-2012.pdf. 40

S. Gorlow and J. D. Reiss. Model-Based Inversion of Dynamic Range Compression. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1434–1444, 2013. doi: 10.1109/TASL.2013.2253099. URL http://hal.archives-ouvertes.fr/hal-00728059. 24

M. Goto and S. Hayamizu. A Real-time Music Scene Description System: Detecting Melody and Bass Lines in Audio Signals. In *Speech Communication*, 1999. 55, 125

M. Goto and Y. Muraoka. A beat tracking system for acoustic signals of music. In *In Proc. of the Second ACM Intl. Conf. on Multimedia*, pages 365–372, 1994. 57

R. Gray, A. Buzo, A. G. Jr, and Y. Matsuyama. Distortion measures for speech processing. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):367–376, 1980. 63

J. M. Grey. Multidimensional Perceptual Scaling of Musical Timbre. *J. of the Acoustical Soc. of America*, 1977. 32

R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte. Proposals for performance measurement in source separation. In *Proc. (ICA) Int. Conf. on Independent Component Analysis and Blind Source Separation*, pages 763–768, Nara, 2003. 87

J. Han, G. J. Mysore, and B. Pardo. Audio Imputation Using the Non-negative Hidden Markov Model. In *Lecture Notes in Computer Science: Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 347–355, 2012. 6

P. C. Hansen, T. K. Jensen, and G. Rodriguez. An adaptive pruning algorithm for the discrete L-curve criterion. *Journal of Computational and Applied Mathematics*, 198(2):483–492, 2007. 98

P. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics, 1987.

ISBN 9780898714036. 98

P. Hansen. *Discrete Inverse Problems: Insight and Algorithms*. Fundamentals of algorithms. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2010. ISBN 9780898718836. 97

M. Haro, J. Serrà, P. Herrera, and A. Corral. Zipf's Law in Short-Time Timbral Codings of Speech, Music, and Environmental Sound Signals. *PLoS ONE*, 7:e33993, 03/2012 2012. URL http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0033993. 32

A. Hazan. BillaBoop: Real-Time Voice-Driven Drum Generator. In *Proceedings of Audio Engineering Society, 118th Convention*, 2005. 41

A. Hazan, R. Marxer, P. Brossier, H. Purwins, P. Herrera, and X. Serra. What/when causal expectation modelling applied to audio signals. *Connection Science*, 21:119–143, June 2009. 220

A. Hazan, P. Brossier, P. Holonowicz, P. Herrera, and H. Purwins. Expectation Along The Beat: A Use Case For Music Expectation Models. In *Proceedings of International Computer Music Conference 2007*, pages 228–236, Copenhagen, Denmark, 2007. 58

T. Heittola, A. Klapuri, and T. Virtanen. Musical Instrument Recognition in Polyphonic Audio Using Source-Filter Model for Sound Separation. In K. Hirata, G. Tzanetakis, and K. Yoshii, editors, *ISMIR*, pages 327–332. International Society for Music Information Retrieval, 2009. ISBN 978-0-9813537-0-8. URL http://www.bibsonomy.org/bibtex/241d8584ef951ac59256650108af17663/dblp. 74, 78, 96, 182

M. Helén and T. Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Proc. EUSIPCO*, volume 2005, 2005. 41, 142, 167, 189, 205

R. Hennequin, R. Badeau, and B. David. Time-dependent parametric and harmonic templates in non-negative matrix factorization. *Proc. of DAFx-10, Graz, Austria*, pages 109–112, 2010. 74, 182

R. Hennequin, B. David, and R. Badeau. Score informed audio source separation using a parametric model of non-negative spectrogram. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 45–48. IEEE, 2011. 59

P. Herrera, A. Yeterian, and F. Gouyon. Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques. In *Proceedings of Second International Conference on Music and Artificial Intelligence*, Edinburgh, Scotland, 2002. 41

J. Hidalgo. Low Latency Audio Source Separation for Speech Enhancement in Cochlear Implants. Master's thesis, 2012. URL http://www.mtg.upf.edu/node/2610. 8

A. E. Hoerl and R. W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, February 1970a. ISSN 00401706. doi: 10.2307/1267351. 97

A. E. Hoerl and R. W. Kennard. Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, 12(1):69–82, 1970b. 97

T. Hofmann. Probabilistic Latent Semantic Analysis. In K. B. Laskey and H. Prade, editors, *UAI*, pages 289–296. Morgan Kaufmann, 1999. URL http://www.bibsonomy.org/bibtex/2aaf90ba1e6d171233521cce51d7d5741/becker. 85

P. O. Hoyer. Non-negative sparse coding. *CoRR*, cs.NE/0202009, 2002. URL http://arxiv.org/abs/cs.NE/0202009. 66, 73

P. O. Hoyer and P. Dayan. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004. 28, 66

C.-L. Hsu and J.-S. Jang. On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(2):310–319, feb. 2010a. ISSN 1558-7916. doi: 10.1109/TASL.2009.2026503. 181

C. Hsu and J. Jang. On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(2):310–319, February 2010b. ISSN 1558-7916. doi: 10.1109/TASL.2009.2026503. 40, 92, 167, 205

G. Hu and D. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *Neural Networks, IEEE Transactions on*, 15 (5):1135–1150, 2004. 56

N. E. Huang, S. R. Long, and Z. Shen. The Mechanism for Frequency Downshift in Nonlinear Wave Evolution. In J. W. Hutchinson and T. Y. Wu, editors, , volume 32 of *Advances in Applied Mechanics*, pages 59–117C. Elsevier, 1996. doi: 10.1016/S0065-2156(08)70076-0. URL http://www.sciencedirect.com/science/article/pii/S0065215608700760. 232

N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathemat-*

ical, Physical and Engineering Sciences, 454(1971):903–995, 1998. doi: 10.1098/rspa.1998.0193. URL http://rspa.royalsocietypublishing.org/content/454/1971/903.abstract. 232, 233

N. E. Huang, Z. Shen, and S. R. Long. A new view of nonlinear water waves: the Hilbert spectrum. *Annual Review of Fluid Mechanics*, 31 (1):417–457, 1999. URL http://www.annualreviews.org/doi/abs/10.1146/annurev.fluid.31.1.417. 232

F. Itakura and S. Saito. Analysis synthesis telephony based on the maximum likelihood method. In *Proceedings of the International Congress on Acoustics*, pages 17–20, 1968. 63

J. Janer, R. Marxer, and K. Arimoto. Combining a harmonic-based NMF decomposition with transient analysis for instantaneous percussion separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 281–284. IEEE, 2012. 205

J. Janer and R. Marxer. NMF-based Separation of Unvoiced Fricatives in Singing Voice Music Mixtures. In *Proc. (DAFx) International Conference on Digital Audio Effects*, Dublin, Ireland, 2013. (accepted).

C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller. Real-Time Speech Separation by Semi-supervised Nonnegative Matrix Factorization. 7191:322–329. doi: 10.1007/978-3-642-28551-6_40. URL http://dx.doi.org/10.1007/978-3-642-28551-6_40. 8

D. H. Johnson and D. E. Dudgeon. *Array signal processing : concepts and techniques*. P T R Prentice Hall, Englewood Cliffs, NJ, 1993. ISBN 0130485136 9780130485137. 52

A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures. In *Proc. (ICASSP) International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 2985–2988 vol.5, 2000. 27, 46, 125, 127

C. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991. ISSN 0165-1684. doi: 10.1016/0165-1684(91)90079-X. URL http://www.sciencedirect.com/science/article/pii/016516849190079X. 4, 45

H. Kameoka. *Statistical Approach to Multipitch Analysis*. PhD thesis, University of Tokyo, Japan, March 2007. URL http://hil.t.u-tokyo.ac.jp;/kameoka/Kameoka2007DoctorThesis.pdf. 131

H. Kameoka, T. Nishimoto, and S. Sagayama. A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering. *IEEE Transactions on*

Audio, Speech & Language Processing, 15(3):982–994, 2007. 55, 131

H. Kameoka, M. Nakano, K. Ochiai, Y. Imoto, K. Kashino, and S. Sagayama. Constrained and Regularized Variants of Non-negative Matrix Factorization Incorporating Music-specific Constraints. In *Proc. of ICASSP*, pages 5365–5368, Mar. 2012. 58

M. Kim, J. Yoo, K. Kang, and S. Choi. Nonnegative Matrix Partial Co-Factorization for Spectral and Temporal Drum Source Separation. *J. Sel. Topics Signal Processing*, 5(6):1192–1204, 2011. URL http://dx.doi.org/10.1109/JSTSP.2011.2158803. 41

A. Klapuri. Sound Onset Detection by Applying Psychoacoustic Knowledge. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Phoenix, USA, 1999. 57

A. Klapuri, T. Virtanen, and T. Heittola. Sound source separation in monaural music signals using excitation-filter model and em algorithm. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5510–5513, 2010. 141

A. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Audio, Speech & Language Processing*, 11:804–816, 2003. doi: 10.1109/TSA.2003.815516. 55, 117, 205

A. Klapuri. *Signal Processing Methods for the Automatic Transcription of Music*. PhD thesis, Tampere University of Technology, Finnland, March 2004. URL http://www.cs.tut.fi/sgn/arg/klap/phd/klap_phd.pdf. 54

A. Klapuri. Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes. In *ISMIR*, pages 216–221, October 2006. 32, 100, 101

D. H. Klatt and L. C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2): 820–857, 1990. URL http://www.bibsonomy.org/bibtex/214fee05d2a6cd0eb1f88604fb8cec911/sourcefilter. 40, 79, 173

C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(4):320–327, 1976. 6, 47

R. Kompass. A Generalized Divergence Measure for Nonnegative Matrix Factorization. *Neural Comput.*, 19(3):780–791, March 2007. ISSN 0899-7667. doi: 10.1162/neco.2007.19.3.780. URL http://dx.doi.org/10.1162/neco.2007.19.3.780. 64

R. H. Lambert. Difficulty measures and figures of merit for source separation. In *Independent Component Analysis*, 1999. 86

D. Lang and N. de Freitas. Beat Tracking the Graphical Model Way. In *Advances in Neural Information Processing Systems (NIPS)*, volume 18, 2004. 58

D. Lee and H. Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2001. 60, 63, 66, 75, 81

A. Lefevre, F. Bach, and C. Févotte. Online algorithms for nonnegative matrix factorization with the Itakura-Saito divergence. In *WASPAA*, pages 313–316. IEEE, 2011. ISBN 978-1-4577-0692-9. URL http://dx.doi.org/10.1109/ASPAA.2011.6082314. 36

S. Z. Li, X. Hou, H. Zhang, and Q. Cheng. Learning Spatially Localized, Parts-Based Representation. In *CVPR (1)*, pages 207–212. IEEE Computer Society, 2001. ISBN 0-7695-1272-0. URL http://www.bibsonomy.org/bibtex/2cf04d90a83a3f5cacc8b85371244b074/dblp. 65

Y. Li and D. Wang. Separation of Singing Voice From Music Accompaniment for Monaural Recordings. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(4):1475–1487, May 2007. ISSN 1558-7916. doi: 10.1109/TASL.2006.889789. 40

Y. Li and D. Wang. Musical Sound Separation Using Pitch-Based Labeling and Binary Time-Frequency Masking. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 173–176, 2008. doi: 10.1109/ICASSP.2008.4517574. 47, 56

Y. Li, J. Woodruff, and D. Wang. Monaural Musical Sound Separation Based on Pitch and Common Amplitude Modulation. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(7):1361–1371, September 2009. ISSN 1558-7916. doi: 10.1109/TASL.2009.2020886. 56

J. Liljencrants. *Speech Synthesis with a Reflection-type Line Analog*. Trita-TÖM. Royal Institute of Technology, 1985. 40

I.-T. Lim and B. G. Lee. Lossless pole-zero modeling of speech signals. *Speech and Audio Processing, IEEE Transactions on*, 1(3):269–276, 1993. ISSN 1063-6676. doi: 10.1109/89.232610. 40

A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard. Informed source separation through spectrogram coding and data embedding. *Signal Processing*, September 2011. ISSN 0165-1684. doi: 10.1016/j.sigpro.2011.09.016. URL http://www.sciencedirect.com/science/article/pii/S0165168411003173. 35

A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard. Adaptive Filtering for MusicVoice Separation Exploiting the Repeating Musical Structure. In *Proceedings IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP'2012)*, March 2012. 58, 168

S. Macleod and Malcolm. Onset detection in musical audio signals. In *In Proc. Int. Computer Music Conference*, 2003. 57

S. Maeda. A digital simulation method of the vocal-tract system. *Speech Communication*, 1(3–4):199–229, 1982. ISSN 0167-6393. doi: 10.1016/0167-6393(82)90017-6. URL http://www.sciencedirect.com/science/article/pii/0167639382900176. 40

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online Learning for Matrix Factorization and Sparse Coding. *Journal of Machine Learning Research*, 11:19–60, 2010. URL http://doi.acm.org/10.1145/1756006.1756008. 36

S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993. URL http://dx.doi.org/10.1109/78.258082. 230

M. Marchini and H. Purwins. Unsupervised Generation of Percussion Sound Sequences from a Sound Example. In *Sound and Music Computing Conference*, 2010. 220

R. Marogna and F. Avanzini. Physically-based synthesis of nonlinear circular membranes. In *Proc. Int. Conf. Digital Audio Effects (DAFx-09)*, pages 373–379, 2009. 41

D. W. Marquardt. Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. *Technometrics*, 12(3):591–612, 1970. 97

A. Marti, M. Cobos, and J. J. Lopez. Automatic speech recognition in cocktail-party situations: A specific training for separated speech. *The Journal of the Acoustical Society of America*, 131(2):1529–1535, 2012. doi: 10.1121/1.3675001. URL http://link.aip.org/link/?JAS/131/1529/1. 6

R. Marxer. Signal decomposition by a joint pitch, timbre and wideband model. Technical report, Universitat Pompeu Fabra, 2011. 143

R. Marxer and J. Janer. A Tikhonov regularization method for spectrum decomposition in low latency audio source separation. In *Proceedings IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP'2012)*, March 2012. 209

R. Marxer and J. Janer. Realtime Bass Separation using Harmonic-

Percussion Decomposition. In *Proc. (DAFx) International Conference on Digital Audio Effects*, Dublin, Ireland, 2013a. (accepted).

R. Marxer and J. Janer. Modelling and Separation of Singing Voice Breathiness in Polyphonic Mixtures. In *Proc. (DAFx) International Conference on Digital Audio Effects*, Dublin, Ireland, 2013b. (accepted).

R. Marxer and J. Janer. Use of regularizations and constraints in NMF-based drums monaural separation. In *Proc. (DAFx) International Conference on Digital Audio Effects*, Dublin, Ireland, 2013c. (accepted).

R. Marxer, J. Janer, and J. Bonada. Low-latency Instrument Separation in Polyphonic Audio Mixtures Using Timbre Models. *Signal Processing (submitted)*, 2011. 36, 115, 116

R. Marxer, J. Janer, and J. Bonada. Low-Latency instrument separation in polyphonic audio using timbre models. *Latent Variable Analysis and Signal Separation*, pages 314–321, 2012. 205

P. Masri. *Computer modelling of sound for transformation and synthesis of musical signal*. PhD thesis, University of Bristol, UK, 1996. 56, 57

D. Mehta and T. F. Quatieri. Synthesis, analysis, and pitch modification of the breathy vowel. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1628–1639, 2005. 40, 172, 223

MIREX. *2005 MIREX Contest Results - Audio Melody Extraction*. 2005. URL http://www.music-ir.org/evaluation/mirex-results/audio-melody. 107, 108

M. Mørup and M. Schmidt. Sparse nonnegative tensor 2d deconvolution (sntf2d) for multichannel time-frequency analysis. *DTU Informatics and Mathematical Modeling Technical Report*, 2006. 76

S. Musevic and J. Bonada. Comparison of non-stationary sinusoid estimation methods using reassignment and derivatives. In *Sound and Music Computing Conference*, 2010. URL http://smcnetwork.org/files/proceedings/2010/14.pdf. 234

K. Nakadai, H. Nakajima, M. Murase, H. G. Okuno, Y. Hasegawa, and H. Tsujino. Real-time tracking of multiple sound sources by integration of in-room and robot-embedded microphone arrays. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 852–859. IEEE, 2006. 8

T. Nakatani and T. Irino. Robust and accurate fundamental frequency estimation based on dominant harmonic components. *The Journal of the Acoustical Society of America*, 116(6):3690–3700, 2004. doi: 10.1121/1.

1787522. URL http://link.aip.org/link/?JAS/116/3690/1. 235

X. Niu, A. Kain, and J. P. H. van Santen. Estimation of the acoustic properties of the nasal tract during the production of nasalized vowels. In *INTERSPEECH'05*, pages 1045–1048, 2005. 40

A. M. Noll. Cepstrum Pitch Determination. *The Journal of the Acoustical Society of America*, 41(2):293–309, 1967. doi: 10.1121/1.1910339. URL http://link.aip.org/link/?JAS/41/293/1. 33, 55

K. Nordstrom, G. Rutledge, and P. Driessen. Using voice conversion as a paradigm for analyzing breathy singing voices. In *Communications, Computers and signal Processing, 2005. PACRIM. 2005 IEEE Pacific Rim Conference on*, pages 428–431, aug. 2005. doi: 10.1109/PACRIM.2005.1517317. 172

K. Nordstrom, G. Tzanetakis, and P. Driessen. Transforming Perceived Vocal Effort and Breathiness Using Adaptive Pre-Emphasis Linear Prediction. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(6):1087–1096, aug. 2008. ISSN 1558-7916. doi: 10.1109/TASL.2008.2001105. 172

P. O'Grady and B. Pearlmutter. Convolutive non-negative matrix factorisation with a sparseness constraint. In *Machine Learning for Signal Processing, 2006. Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on*, pages 427–432. IEEE, 2006. 76

P. O'Grady, B. Pearlmutter, and S. Rickard. Survey of sparse and non-sparse methods in source separation. *International Journal of Imaging Systems and Technology*, 15(1):18–33, 2005. 16

M. Omologo and P. Svaizer. Use of the crosspower-spectrum phase in acoustic event location. *Speech and Audio Processing, IEEE Transactions on*, 5(3):288–292, 1997. 47

N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama. A real-time equalizer of harmonic and percussive components in music signals. In *ISMIR 2008: proceedings of the 9th International Conference of Music Information Retrieval*, page 139, 2008a. 4, 142, 148

N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama. Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram. In *Proc. EUSIPCO*, 2008b. 41, 167, 189, 195, 205

A. Ozerov, E. Vincent, and F. Bimbot. A general modular framework for audio source separation. In *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10)*, Saint-Malo, France,

September 2010. URL http://hal.inria.fr/inria-00553504/en. 80, 126, 132, 133, 141, 148, 151, 156, 167, 168, 169, 189, 195

P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994. URL http://dx.doi.org/10.1002/env.3170050203. 60

L. Parra and C. Spence. Convolutive blind separation of non-stationary sources. In *IEEE Transactions Speech and Audio Processing*, pages 320–327, 2000. 44

L. Parra, C. Spence, and B. De Vries. Convolutive blind source separation based on multiple decorrelation. In *Neural Networks for Signal Processing VIII, 1998. Proceedings of the 1998 IEEE Signal Processing Society Workshop*, pages 23–32, aug-2 sep 1998. doi: 10.1109/NNSP.1998.710626. 44

L. Parra and P. Sajda. Blind source separation via generalized eigenvalue decomposition. *J. Mach. Learn. Res.*, 4(7-8):1261–1269, October 2004. ISSN 1532-4435. doi: 10.1162/jmlr.2003.4.7-8.1261. URL http://dx.doi.org/10.1162/jmlr.2003.4.7-8.1261. 44, 45

M. Parvaix and L. Girin. Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding. *Audio, Speech, and Language Processing, IEEE Transactions on*, (99):1–1, 2011. 35

A. D. Pascual-Montano, J. M. Carazo, K. Kochi, D. Lehmann, and R. D. Pascual-Marqui. Nonsmooth Nonnegative Matrix Factorization (nsNMF). *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(3):403–415, 2006. URL http://www.bibsonomy.org/bibtex/20a3b41983507b4a744a49736a2b2fdb5/dblp. 68

Y. C. Pati, R. Rezaiifar, Y. C. P. R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition. In *Proceedings of the 27 th Annual Asilomar Conference on Signals, Systems, and Computers*, pages 40–44, 1993. 231

J. Paulus and T. Virtanen. Drum Transcription with Non-negative Spectrogram Factorisation. In *Proc. of the 13th European Signal Processing Conference*, Antalya, Turkey, September 2005. 41

A. Pedone, J. J. Burred, S. Maller, and P. Leveau. Phoneme-Level Text to Audio Synchronization on Speech Signals with Background Music. In *INTERSPEECH*, pages 433–436, 2011. 9

L. D. Persia, D. Milone, H. L. Rufiner, and M. Yanagida. Perceptual evaluation of blind source separation for robust speech recognition. *Signal Processing*, 88(10):2578–2583, 2008. ISSN 0165-1684. doi: 10.1016/j.sigpro.2008.04.006. URL http://www.sciencedirect.com/science/article/pii/S0165168408001230. 6

A. Pertusa and J. M. Inesta. Multiple fundamental frequency estimation using Gaussian smoothness. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 105–108. IEEE, 2008. 123

D.-T. Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of nonstationary sources. *Signal Processing, IEEE Transactions on*, 49 (9):1837–1848, sep 2001. ISSN 1053-587X. doi: 10.1109/78.942614. 45

D. L. Phillips. A Technique for the Numerical Solution of Certain Integral Equations of the First Kind. *J. ACM*, 9(1):84–97, January 1962. ISSN 0004-5411. doi: 10.1145/321105.321114. URL http://doi.acm.org/10.1145/321105.321114. 97

H. Purwins, M. Grachten, P. Herrera, A. Hazan, R. Marxer, and X. Serra. Computational models of music perception and cognition II: Domain-specific music processing. *Physics of Life Reviews*, 5(3):169–182, 2008a. ISSN 1571-0645. doi: 10.1016/j.plrev.2008.03.005. URL http://www.sciencedirect.com/science/article/pii/S157106450800016X. 31

H. Purwins, P. Herrera, M. Grachten, A. Hazan, R. Marxer, and X. Serra. Computational models of music perception and cognition I: The perceptual and cognitive processing chain. *Physics of Life Reviews*, 5 (3):151–168, 2008b. ISSN 1571-0645. doi: 10.1016/j.plrev.2008.03.004. URL http://www.sciencedirect.com/science/article/pii/S1571064508000158. 31

F. C. Quesada, N. R. Reyes, P. V. Candeas, J. Carabias, and S. Maldonado. A multiple-F0 estimation approach based on Gaussian spectral modelling for polyphonic music transcription. *Journal of New Music Research*, 39 (1):93–107, 2010. 56

L. Rabiner. On the use of autocorrelation analysis for pitch detection. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 25(1): 24–33, 1977. ISSN 0096-3518. doi: 10.1109/TASSP.1977.1162905. 55

S. A. Raczyński, N. Ono, and S. Sagayama. Extending Nonnagetive Matrix Factorization-a Discussion in the Context of Multiple Frequency Estimation of musical signals. *Proc. of 17th European Signal Processing Conference*, August 2008. 63, 70, 71

S. A. Raczyński, N. Ono, and S. Sagayama. Multipitch Analysis with Harmonic Nonnegative Matrix Approximation. In *in ISMIR 2007, 8th International Conference on Music Information Retrieval*, pages 381–386, 2007. 55, 132, 140

Z. Rafii and B. Pardo. A simple music/voice separation method based on the extraction of the repeating musical structure. In *ICASSP*, pages 221–224. IEEE, 2011. ISBN 978-1-4577-0539-7. URL http://dx.doi.org/10.1109/ICASSP.2011.5946380. 58, 167, 168

Z. Rafii and B. Pardo. Music/Voice Separation Using the Similarity Matrix. In *Proceedings of the 13th International Society for Music Information Retrieval Conference*, Porto, Portugal, October 8-12 2012. http://ismir2012.ismir.net/event/papers/583-ismir-2012.pdf. 40

B. Raj, R. Singh, and T. Virtanen. Phoneme-Dependent NMF for Speech Enhancement in Monaural Mixtures. In *INTERSPEECH*, pages 1217–1220. ISCA, 2011. URL http://www.bibsonomy.org/bibtex/23235b3838242bb942721b5542e015a57/dblp. 74, 182

D. M. Randel. Rhythm Section. In *The Harvard Concise Dictionary of Music and Musicians (Harvard University Press Reference Library)*, page 560. Belknap Press of Harvard University Press, 1999. ISBN 0674009789. 32

C. Raphael. A classifier-based approach to score-guided source separation of musical audio. *Computer Music Journal*, 32(1):51–59, 2008. 59

F. Reed, P. Feintuch, and N. Bershad. Time delay estimation using the LMS adaptive filter–static behavior. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 29(3):561–571, 1981. 47

H. Riemann. *Musikalische Syntaxis*. Leipzig, 1877. 32

J. D. Riley. Solving Systems of Linear Equations With a Positive Definite, Symmetric, but Possibly Ill-Conditioned Matrix. *Mathematical Tables and Other Aids to Computation*, 9(51):96–101, July 1955. ISSN 0891-6837. 97

G. Rilling, P. Flandrin, and P. Goncalves. On empirical mode decomposition and its algorithms. In *Proceedings of IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing NSIP-03*, Grado, Italy, 2003. URL http://hal.inria.fr/inria-00570628. 233

A. Röbel. Transient detection and preservation in the phase vocoder. In *Proc. Int. Computer Music Conference (ICMC)*, pages 247–250, 2003. 57, 145, 194

A. Röbel. Onset Detection in Polyphonic Signals by means of Transient Peak

Classification. *International Symposium for Music Information retrieval (ISMIR/MIREX'05)*, 2005. 57, 146

A. Röbel. *Onset Detection By Means Of Transient Peak Classification In Harmonic Bands*. MIREX, 2009. 57

F. J. Rodriguez-Serrano, J. J. Carabias-Orti, P. Vera-Candeas, T. Virtanen, and N. Ruiz-Reyes. Multiple Instrument Mixtures Source Separation Evaluation Using Instrument-Dependent NMF Models. volume 7191 of *LVA/ICA'12*, pages 380–387. Springer, 2012. ISBN 978-3-642-28550-9. URL http://dblp.uni-trier.de/db/conf/ica/ica2012.html#Rodriguez-SerranoCVVR12. 56, 167, 169, 205, 212

S. Rolewicz. *Metric linear spaces*. Mathematics and its applications: East European series. D. Reidel, 1985. ISBN 9789027714800. URL http://books.google.es/books?id=qVmHL2M5I9oC. 28

A. E. Rosenberg. Effect of glottal pulse shape on the quality of natural vowels. *The Journal of the Acoustical Society of America*, 49(2B):583–590, 1971. 40

T. D. Rossing. *Science of Sound, The (2nd Edition)*. Pearson Education, 1990. ISBN 0201157276. 30, 41

M. Ryynänen. Singing Transcription. In A. Klapuri and M. Davy, editors, *Signal Processing Methods for Music Transcription*, pages 361–390. Springer Science + Business Media LLC, 2006. ISBN 0-387-30667-6. 40, 126

S. Sagayama, K. Takahashi, H. Kameoka, and T. Nishimoto. Specmurt anasylis: A piano-roll-visualization of polyphonic music signal by deconvolution of log-frequency spectrum. In *in Proceedings of ISCA*, 2004. 55

S. Saito, H. Kameoka, K. Takahashi, T. Nishimoto, and S. Sagayama. Specmurt analysis of polyphonic music signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(3):639–650, 2008. 55

J. Salamon and E. Gomez. Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(6):1759–1770, aug. 2012. ISSN 1558-7916. doi: 10.1109/TASL.2012.2188515. 55, 110, 115, 116

E. D. Scheirer. Tempo and Beat Analysis of Acoustic Musical Signals. *J. of the Acoustical Soc. of America*, 103(1):588–601, 1998. 58

H. Schenker. *Der freie Satz*, volume 3 of *Neue musikalische Theorien und Phantasien*. Wien, 1935. 32

W. A. Schloss. *On the Automatic Transcription of Percussive Music: From*

Acoustic Signal to High Level Analysis. PhD thesis, Stanford University, CA, USA, May 1985. URL http://ccrma.stanford.edu/STANM/stanms/stanm27/stanm27.pdf. 41

M. N. Schmidt and H. Laurberg. Nonnegative matrix factorization with Gaussian process priors. *Intell. Neuroscience*, 8(1):1–10, 2008. URL http://portal.acm.org/citation.cfm?id=1384872&dl=GUIDE&coll=GUIDE&CFID=32586972&CFTOKEN=54678655. 82, 84

M. N. Schmidt and M. Mørup. Nonnegative Matrix Factor 2-D Deconvolution for Blind Single Channel Source Separation. In J. P. Rosca, D. Erdogmus, J. C. Príncipe, and S. Haykin, editors, *ICA*, volume 3889 of *Lecture Notes in Computer Science*, pages 700–707. Springer, 2006a. ISBN 3-540-32630-8. URL http://dx.doi.org/10.1007/11679363_87. 76

M. Schmidt and M. Mørup. Sparse non-negative matrix factor 2-d deconvolution for automatic transcription of polyphonic music. In *Proc. 6th International Symposium on Independent Component Analysis and Blind Signal Separation*, 2006b. 76

M. Schmidt and R. Olsson. Single-channel speech separation using sparse non-negative matrix factorization. In *Ninth International Conference on Spoken Language Processing*, 2006. 73, 181

R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34 (3):276–280, March 1986. ISSN 0096-1973. doi: 10.1109/TAP.1986.1143830. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1143830. 6, 52

D. Schobben and K. Torkkola. Evaluation of blind signal separation methods. *Proc Int Workshop on ICA and*, 5, 1999. URL http://www.oca.eu/Bijaoui/doc_ab/cardon/ica99.pdf. 86

M. R. Schroeder. Period Histogram and Product Spectrum: New Methods for Fundamental-Frequency Measurement. *The Journal of the Acoustical Society of America*, 43(4):829–834, 1968. doi: 10.1121/1.1910902. URL http://link.aip.org/link/?JAS/43/829/1. 55

B. Schuller, A. Lehmann, F. Weninger, F. Eyben, and G. Rigoll. Blind enhancement of the rhythmic and harmonic sections by NMF: Does it help? In *Proc. of the International Conference on Acoustics (NAG/DAGA 2009)*, pages 361–364, 2009. 41

F. Sha and L. K. Saul. Real-time pitch determination of one or more voices by nonnegative matrix factorization. In *in Advances in Neural Informa-*

tion Processing Systems 17, pages 1233–1240. MIT Press, 2005. 126

M. Shashanka, B. Raj, and P. Smaragdis. Probabilistic latent variable models as nonnegative factorizations. *Computational intelligence and neuroscience*, pages 947438+, January 2008. ISSN 1687-5265. doi: 10.1155/2008. 83

H. A. J. S. S. K.-M. Shim. Stereo Music Source Separation for 3-D Upmixing. In *Audio Engineering Society Convention 127*, 10 2009. URL http://www.aes.org/e-lib/browse.cfm?elib=15132. 7

R. Shrivastav and C. M. Sapienza. Objective measures of breathy voice quality obtained using an auditory model. *The Journal of the Acoustical Society of America*, 114(4):2217–2224, 2003. doi: 10.1121/1.1605414. URL http://link.aip.org/link/?JAS/114/2217/1. 177

P. Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. *Independent Component Analysis and Blind Signal Separation*, pages 494–499, 2004. 75, 76

P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 3(3):177–180, 2003. 4, 55, 61, 131

P. Smaragdis and B. Raj. Shift-Invariant Probabilistic Latent Component Analysis. Technical report, Mitsubishi Electric Research Laboratories, 2007. URL http://www.merl.com/publications/TR2007-009. 83

P. Smaragdis. Polyphonic pitch tracking by example. In *WASPAA*, pages 125–128. IEEE, 2011. ISBN 978-1-4577-0692-9. URL http://dx.doi.org/10.1109/ASPAA.2011.6082344. 55

P. Smaragdis and G. J. Mysore. "Separation by Humming": User Guided Sound Extraction from Monophonic Mixtures. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009. 167, 169, 171, 205

P. Smaragdis, B. Raj, and M. V. S. Shashanka. Supervised and Semi-supervised Separation of Sounds from Single-Channel Mixtures. In M. E. Davies, C. J. James, S. A. Abdallah, and M. D. Plumbley, editors, *ICA*, volume 4666 of *Lecture Notes in Computer Science*, pages 414–421. Springer, 2007. ISBN 978-3-540-74493-1. URL http://dx.doi.org/10.1007/978-3-540-74494-8_52. 73, 86

P. Smaragdis, M. Shashanka, and B. Raj. A Sparse Non-Parametric Approach for Single Channel Separation of Known Sounds. *Advances in Neural Information Processing Systems 22*, pages 1705–1713, 2009. 74

E. Smith and M. Lewicki. Efficient auditory coding. *Nature*, 439: 978–982, February 2006. URL http://www.bibsonomy.org/bibtex/271400ab57a408d344430d520092afa51/tb2332. 231

L. Smith. Listening to musical rhythms with progressive wavelets. In *TENCON '96. Proceedings. 1996 IEEE TENCON. Digital Signal Processing Applications*, volume 2, pages 508–513 vol.2, 1996. 57

S. Sra and I. S. Dhillon. Nonnegative matrix approximation: algorithms and applications. Technical report, Department Of Computer Sciences, University of Texas at Austin, 2006. 76, 81

K. N. Stevens. Airflow and Turbulence Noise for Fricative and Stop Consonants: Static Considerations. *The Journal of the Acoustical Society of America*, 50(4B):1180–1192, 1971. doi: 10.1121/1.1912751. URL http://link.aip.org/link/?JAS/50/1180/1. 40

K. N. Stevens. *Acoustic Phonetics*. MIT Press, Cambridge, MA, 1998. 38

S. S. Stevens, J. Volkmann, and E. B. Newman. A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937. doi: 10.1121/1.1915893. URL http://link.aip.org/link/?JAS/8/185/1. 19, 31

K. Tanghe, S. Degroeve, and B. D. Baets. An Algorithm For Detecting and labeling drum events in polyphonic music. In *Mirex Drum Recognition Contest (part of International Symposium on Music Information Retreival)*, 2005. 41

A. Tikhonov. Solution of incorrectly formulated problems and the regularization method. In *Soviet Math. Doklady*, volume 4, pages 1035–1038, 1963. 97, 135

I. Titze. *Principles of voice production*. Prentice Hall, 1994. ISBN 9780137178933. 38

I. Titze and F. Alipour. *The Myoelastic Aerodynamic Theory of Phonation*. National Center for Voice and Speech, 2006. ISBN 9780874141566. 38

L. Trautmann, S. Petrausch, and R. Rabenstein. Physical modeling of drums by transfer function methods. In *Proc. Int. Conf. on Acoustics, Speech & Signal Processing (ICASSP)*, pages 3385–3388. IEEE, 2001. 41

S. Twomey. On the Numerical Solution of Fredholm Integral Equations of the First Kind by the Inversion of the Linear System Produced by Quadrature. *J. ACM*, 10(1):97–101, January 1963. ISSN 0004-5411. doi: 10.1145/321150.321157. URL http://doi.acm.org/10.1145/321150.321157. 97

Y. Umeyama, K. Kondo, Y. Takahashi, J. Bonada, J. Janer, and R. Marxer.

Graphical Audio Signal Control, February 7 2012. US Patent App. 13/367,696.

J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat. Localization of simultaneous moving sound sources for mobile robot using a frequency- domain steered beamformer approach. In *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, volume 1, pages 1033–1038 Vol.1, april-1 may 2004. doi: 10.1109/ROBOT.2004.1307286. 48

J.-M. Valin, F. Michaud, and J. Rouat. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems*, 55(3):216–228, 2007. ISSN 0921-8890. doi: 10.1016/j.robot.2006.08.004. URL http://www.sciencedirect.com/science/article/pii/S0921889006001576. 8, 49

S. A. Van Duyne and J. O. Smith. Physical Modeling with the 2-D digital waveguide mesh. In *Proc. Int. Computer Music Conf.*, pages 40–47, Tokyo, Japan, 1993. 41

E. Vincent and M. Plumbley. A prototype system for object coding of musical audio. In *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, pages 239–242, 2005. doi: 10.1109/ASPAA.2005.1540214. 8

E. Vincent, R. Gribonval, and M. Plumbley. Oracle estimators for the benchmarking of source separation algorithms. *Signal Processing*, 87(8): 1933–1950, 2007a. 91, 92, 137, 148, 156, 209

E. Vincent, R. Gribonval, and M. Plumbley. *A toolbox to compute oracle estimators for source separation*. 2007b. URL http://bass-db.gforge.inria.fr/bss_oracle. 91, 92

E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca. First stereo audio source separation evaluation campaign: data, algorithms and results. In *Proceedings of the 7th international conference on Independent component analysis and signal separation*, pages 552–559, 2007c. ISBN 3540744932. 89, 129

E. Vincent, N. Bertin, and R. Badeau. Adaptive Harmonic Spectral Decomposition for Multiple Pitch Estimation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(3):528–537, March 2010. ISSN 1558-7916. doi: 10.1109/TASL.2009.2034186. 207

E. Vincent, C. Févotte, R. Gribonval, L. Benaroya, X. Rodet, A. Röbel, E. L. Carpentier, and F. Bimbot. A tentative typology of audio source separation tasks. In *4th Int. Symp. on Independent Component Analysis*

and Blind Signal Separation (ICA), pages 715–720, Nara, Japan, 2003. URL http://hal.inria.fr/inria-00544239. 5, 7, 87

E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech & Language Processing*, 14(4):1462–1469, 2006. 87, 137, 147, 156, 209

E. Vincent, N. Bertin, and R. Badeau. Harmonic and inharmonic Non-negative Matrix Factorization for Polyphonic Pitch transcription. In *ICASSP*, pages 109–112. IEEE, 2008. ISBN 1-4244-1484-9. URL http://dx.doi.org/10.1109/ICASSP.2008.4517558. 77, 96

E. Vincent, S. Araki, and P. Bofill. The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation. *Independent Component Analysis and Signal Separation*, pages 734–741, 2009. 90

M. Vinyes. MTG MASS database. http://www.mtg.upf.edu/static/mass/resources, 2008. 91

M. Vinyes, J. Bonada, and A. Loscos. Demixing Commercial Music Productions via Human-Assisted Time-Frequency Masking. In *Proceedings of Audio Engineering Society 120th Convention*. Morgan Kaufmann, 2006. 4, 47, 91, 125, 127

T. Virtanen and A. Klapuri. Analysis of polyphonic audio using source-filter model and non-negative matrix factorization. In *Advances in Models for Acoustic Processing, Neural Information Processing Systems Workshop*. Citeseer, 2006. 77, 96

T. Virtanen, A. Cemgil, and S. Godsill. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 1825–1828, 31 2008-april 4 2008a. doi: 10.1109/ICASSP.2008.4517987. 85

T. Virtanen, A. Mesaros, and M. Ryynänen. Combining Pitch-Based Inference and Non-Negative Spectrogram Factorization in Separating Vocals from Polyphonic Music. In *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Brisbane, Australia, September 2008b. 167, 168, 169, 205

T. Virtanen. Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, March 2007. ISSN 1558-7916. doi: 10.1109/TASL.2006.885253. 68, 69, 81, 132, 141, 167, 191, 192, 201

G. Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1990. ISBN 9780898712445. 98

D. Wang. Computational Auditory Scene Analysis. pages 71–85. L. Erlbaum Assoc., 1998. 3

E. Weinstein, M. Feder, and A. Oppenheim. Multi-channel signal separation by decorrelation. *Speech and Audio Processing, IEEE Transactions on*, 1 (4):405–413, oct 1993. ISSN 1063-6676. doi: 10.1109/89.242486. 44, 45

D. L. Wessel. Low Dimensional Control of Musical Timbre. In *Audio Engineering Society Convention 59*, 2 1978. URL http://www.aes.org/e-lib/browse.cfm?elib=3017. 32

K. W. Wilson, B. Raj, and P. Smaragdis. Regularized non-negative matrix factorization with temporal dependencies for speech denoising. In *INTERSPEECH*, pages 411–414. ISCA, 2008a. URL http://www.bibsonomy.org/bibtex/2a9df51adf590ccc2c949e46846dedd9a/dblp. 69, 85

K. Wilson, B. Raj, P. Smaragdis, and A. Divakaran. Speech denoising using nonnegative matrix factorization with priors. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4029–4032. IEEE, 2008b. 69

O. Winther and K. B. Petersen. Bayesian independent component analysis: Variational methods and non-negative decompositions. *Digital Signal Processing*, 17(5):858–872, 2007. URL http://dx.doi.org/10.1016/j.dsp.2007.01.003. 82

C. Wong, W. Szeto, and K. Wong. Automatic lyrics alignment for Cantonese popular music. *Multimedia Systems*, 12(4/5):307–323, March 2007. 181

J. Woodruff, B. Pardo, and R. Dannenberg. Remixing stereo music with score-informed source separation. In *Proc. ISMIR*, pages 314–319, 2006. 59

J. Wu, E. Vincent, S. Raczynski, T. Nishimoto, N. Ono, and S. Sagayama. Multipitch estimation by joint modeling of harmonic and transient sounds. In *IEEE Int. Conf. on Acoustics, Speech & Signal Processing (ICASSP)*, pages 25–28, Prague, Czech Republic, May 2011. doi: 10.1109/ICASSP.2011.5946319. 74, 131, 133, 143

Z. Wu and N. E. Huang. A study of the characteristics of white noise using the empirical mode decomposition method. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 460(2046):1597–1611, 2004. doi: 10.1098/rspa.2003.

1221. URL http://rspa.royalsocietypublishing.org/content/460/2046/1597.abstract. 233

C. Yeh, A. Roebel, and X. Rodet. Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals. *Trans. Audio, Speech and Lang. Proc.*, 18(6):1116–1126, August 2010. ISSN 1063-6676. doi: 10.1109/TASL.2009.2030006. 55, 126

O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *Signal Processing, IEEE Transactions on*, 52(7): 1830–1847, July 2004. ISSN 1053-587X. doi: 10.1109/TSP.2004.828896. 27, 47, 125, 127

J. Yoo, M. Kim, K. Kang, and S. Choi. Nonnegative matrix partial co-factorization for drum source separation. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 1942–1945, 2010. doi: 10.1109/ICASSP.2010.5495305. 189

K. Yoshii, M. Goto, and H. Okuno. INTER:D: a drum sound equalizer for controlling volume and timbre of drums. In *Integration of Knowledge, Semantics and Digital Media Technology, 2005. EWIMT 2005. The 2nd European Workshop on the (Ref. No. 2005/11099)*, pages 205–212, 2005a. 57, 189

K. Yoshii, M. Goto, and H. G. Okuno. AdaMast: A Drum Sound Recognizer based on Adaptation and Matching of Spectrogram Templates. In *Mirex Drum Recognition Contest (part of International Symposium on Music Information Retreival)*, 2005b. 41

J. R. Zapata and E. Gómez. Using voice suppression algorithms to improve beat tracking in the presence of highly predominant vocals. *The 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada.*, 2013. 7

R. Zhou, J. D. Reiss, M. Mattavelli, and G. Zoia. A Computationally Efficient Method for Polyphonic Pitch Estimation. *EURASIP J. Adv. Sig. Proc.*, 2009, 2009. 230

A. Ziehe and K.-R. Müller. TDSEP – an efficient algorithm for blind separation using time structure. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *Proceedings of the 8th International Conference on Artificial Neural Networks, ICANN'98*, Perspectives in Neural Computing, pages 675–680, Berlin, 1998. Springer Verlag. 44

A. Zils, F. Pachet, O. Delerue, and F. Gouyon. Automatic extraction of drum tracks from polyphonic music signals. In *Web Delivering of Music, 2002. WEDELMUSIC 2002. Proceedings. Second International Confer-*

ence on, pages 179–183, 2002. doi: 10.1109/WDM.2002.1176209. 57, 189

L. Zouari and G. Chollet. Efficient Gaussian Mixture for Speech Recognition. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 4, pages 294–297, 0-0 2006. doi: 10.1109/ICPR.2006.475. 114

E. Zwicker. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *The Journal of the Acoustical Society of America*, 33:248, 1961. 31