Informed Source Separation for Multiple Instruments of Similar Timbre

Jakue López Armendáriz

MASTER THESIS UPF / 2013
Master in Sound and Music Computing

Master thesis supervisor:
Dr. Jordi Janer
Department of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona
Acknowledgements

I would first like to thank Prof. Xavier Serra for giving me the opportunity to join the SMC Master’s program at the Pompeu Fabra University (UPF) and to all the people from the Music Technology Group (MTG), specially the members and SMC students which have helped and influenced the development of this work.

I owe my deepest gratitude to my supervisor, Dr. Jordi Janer, whose expertise, understanding and patience made possible the writing of this thesis. I really appreciate his knowledge and his assistance during this time.

I would like to thank my parents for the support they provide through my entire life. Mila esker Aita eta Ama.

I would also like to thank my girlfriend, because she does not only give unlimited help and support, but she also brings optimism and happiness into my life. Nagyon köszönöm Orsi.

To all of you.
Thanks.
Abstract

This Master’s thesis focuses on the challenging task of separating the musical audio sources with instruments of similar timbre. We address the case in which external pitch information to assist the separation process is available. This information is provided to the source / filter model, which is embedded in a Non-Negative Matrix Factorization (NMF) framework that processes the audio input spectrogram. Different state of the art literature methods are inspected and extended. As an extension to these, two new separation methods are proposed, the Multi-Excitation and Single Filter Instantaneous Mixture Model and the Multi-Excitation and Multi-Filter Instantaneous Mixture Model. The use of dedicated source and filter decomposition for each instrument is proposed. In addition, we introduce the use of timbre models in the separation process. Timbre models are previously trained on isolated instrument recordings. The methods are compared with the BSS Eval and PEASS evaluation toolkits over an existing dataset. Promising results obtained in the conducted experiments, which shows that this is a path to be further investigated.
Contents

Index of figures viii
Index of tables xi

1 INTRODUCTION 1

2 STATE OF THE ART 3
2.1 Source Separation 3
  2.1.1 Models and Methods 5
2.2 Musical Audio Source Separation 9
  2.2.1 Instantaneous Mixture Model 10
  2.2.2 Multi-Excitation Model 14
2.3 Bowed String Instrument Modeling 17
  2.3.1 Physics and Models of Violin Family Instruments 17
  2.3.2 Source / Filter Modeling 19

3 METHODOLOGY 21
3.1 First approach: Modifications over the IMM 21
  3.1.1 Source Model 21
3.2 Proposed Extensions: Combining IMM and Multi-Excitation Model 24
  3.2.1 Multi-Excitation and Single Filter IMM 24
  3.2.2 Multi-Excitation and Multi-Filter IMM 25
3.3 Supervised Timbre Models 27

4 EVALUATION AND RESULTS 30
4.1 Evaluation Metrics 30
  4.1.1 The BSS EVAL Toolbox 30
  4.1.2 The PEASS Toolkit 31
4.2 Description of the Dataset 32
4.3 Evaluation Methodology and Results 33
  4.3.1 Comparison of IMM and the Proposed Models 33
4.3.2 Multi-Excitation and Multi-Filter Model Parameter Optimization

4.3.3 Multi-Excitation and Multi-Filter Model with Supervised Timbre Model Extension

5 CONCLUSIONS AND FUTURE WORK

5.1 Contributions

5.2 Conclusions

5.3 Future work

5.3.1 Source / Filter Model Improvements

5.3.2 Trained Filter Models Improvements

A APPENDIX A

B APPENDIX B
## List of Figures

2.1 Source / filter + accompaniment diagram of the IMM .......................... 11
2.2 Solo/Accompaniment algorithm outline [Durrieu et al., 2009a] .......... 13
2.3 Resonance curve of a typical violin body defined as the mechanical input admittance at the bridge [Fletcher, 1999] ..................... 19

3.1 Example of an $W_{F_0}$ basis generated with the glottal source model . 22
3.2 Source / filter + accompaniment diagram of the modified IMM .......... 23
3.3 Multi-Excitation and Single Filter IMM diagram ................................. 25
3.4 Multi-Excitation and Multi-Filter IMM diagram ................................. 26
3.5 Timber model $W_\phi$ bases decomposed into an smooth filter dictionary bases $W_\Gamma$ and its gains $H_\Gamma$ ........................................ 27
3.6 Timbre matrices example over a cello excerpt. Smooth filter dictionary bases $W_\Gamma$ (top), its activations $H_\Gamma$ (middle) and the obtained bases $W_\phi$ .................................................. 28
3.7 Timbre matrices example over a violin excerpt. Smooth filter dictionary bases $W_\Gamma$ (top), its activations $H_\Gamma$ (middle) and the obtained bases $W_\phi$ .................................................. 29
3.8 Multi-Excitation and Multi-Filter IMM diagram with supervised filters loaded .......................................................... 29

4.1 Overall Perceptual Score (OPS) results by model for the Schubert audio excerpt ................................................................. 37
4.2 Overall Perceptual Score (OPS) results by model for the Mozart audio excerpt ................................................................. 37
4.3 Pitch estimations of a 8 second violin excerpt (Schubert recording). Estimations with Yin (top), MEF IMM (middle) and MEF IMM loading supervised timbre models (bottom) .................. 49
4.4 Pitch estimations over a 8 second cello excerpt (Schubert recording). Estimations with Yin (top), MEF IMM (middle) and MEF IMM loading supervised timbre models (bottom) .................. 49
4.5 Pitch estimations over a 8 second viola excerpt (Mozart recording). Estimations with Yin (top), MEF IMM (middle) and MEF IMM loading supervised timbre models (bottom). . . . . . . . . . . 50

4.6 Pitch estimations over a 8 second clarinet excerpt (Mozart recording). Estimations with Yin (top), MEF IMM (middle) and MEF IMM loading supervised timbre models (bottom). . . . . . . . . . . 50
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>SDR score results by model for the Schubert audio excerpt</td>
<td>34</td>
</tr>
<tr>
<td>4.2</td>
<td>SIR score results by model for the Schubert audio excerpt</td>
<td>34</td>
</tr>
<tr>
<td>4.3</td>
<td>SAR score results by model for the Schubert audio excerpt</td>
<td>35</td>
</tr>
<tr>
<td>4.4</td>
<td>SDR score results by model for the Mozart audio excerpt</td>
<td>35</td>
</tr>
<tr>
<td>4.5</td>
<td>SIR score results by model for the Mozart audio excerpt</td>
<td>35</td>
</tr>
<tr>
<td>4.6</td>
<td>SAR score results by model for the Mozart audio excerpt</td>
<td>36</td>
</tr>
<tr>
<td>4.7</td>
<td>PEASS toolkit results by model for the Schubert audio excerpt</td>
<td>37</td>
</tr>
<tr>
<td>4.8</td>
<td>PEASS toolkit results by model for the Mozart audio excerpt</td>
<td>38</td>
</tr>
<tr>
<td>4.9</td>
<td>SDR score results by iteration over 90 $P$ smooth filters and 20 $K$ bases</td>
<td>39</td>
</tr>
<tr>
<td>4.10</td>
<td>SDR score results by $P$ smooth filters over 30 iterations and 20 $K$ bases</td>
<td>40</td>
</tr>
<tr>
<td>4.11</td>
<td>SDR score results by $K$ bases over 90 $P$ smooth filters and 30 iterations</td>
<td>40</td>
</tr>
<tr>
<td>4.12</td>
<td>Best parameter configuration over different tested $K$ bases, $P$ smooth filters and iterations values</td>
<td>40</td>
</tr>
<tr>
<td>4.13</td>
<td>BSSEval results of the Schubert recording for the Multi-Excitation and Multi-Filter Model with estimated filters (MEF IMM), supervised filters (MEF IMM + SV FILT.) and unsupervised filters (MEF IMM + USV FILT.)</td>
<td>42</td>
</tr>
<tr>
<td>4.14</td>
<td>BSSEval results of the Mozart recording for the Multi-Excitation and Multi-Filter Model with estimated filters (MEF IMM), supervised filters (MEF IMM + SV FILT.) and unsupervised filters (MEF IMM + USV FILT.)</td>
<td>42</td>
</tr>
<tr>
<td>4.15</td>
<td>SDR score of the Schubert recording for the Multi-Excitation and Multi-Filter Model with estimated filters and supervised filters for different iteration values</td>
<td>44</td>
</tr>
<tr>
<td>4.16</td>
<td>SDR score of the Mozart recording for the Multi-Excitation and Multi-Filter Model with estimated filters and supervised filters for different iteration values</td>
<td>45</td>
</tr>
</tbody>
</table>
4.17 SDR score of the Schubert recording for the Multi-Excitation and Multi-Filter Model with estimated filters and supervised filters for different block size values ................................................................. 46
4.18 SDR score of the Mozart recording for the Multi-Excitation and Multi-Filter Model with estimated filters and supervised filters for different block size values ................................................................. 47
A.1 BSS Eval results for the Schubert excerpt over 10 iterations and different values for $P$ smooth filters and $K$ bases ................................................................. 56
A.2 BSS Eval results for the Schubert excerpt over 30 iterations and different values for $P$ smooth filters and $K$ bases ................................................................. 57
A.3 BSS Eval results for the Schubert excerpt over 50 iterations and different values for $P$ smooth filters and $K$ bases ................................................................. 58
A.4 BSS Eval results for the Schubert excerpt over 70 iterations and different values for $P$ smooth filters and $K$ bases ................................................................. 59
A.5 BSS Eval results for the Schubert excerpt over 90 iterations and different values for $P$ smooth filters and $K$ bases ................................................................. 60
B.1 SIR error of the Schubert recording for the Multi-Excitation and Multi-Filter Model with estimated filters and trained filters for different iteration values ................................................................. 62
B.2 SIR error of the Mozart recording for the Multi-Excitation and Multi-Filter Model with estimated filters and trained filters for different iteration values ................................................................. 63
B.3 SAR error of the Schubert recording for the Multi-Excitation and Multi-Filter Model with estimated filters, trained filters for different iteration values ................................................................. 64
B.4 SAR error of the Mozart recording for the Multi-Excitation and Multi-Filter Model with estimated filters, trained filters for different iteration values ................................................................. 65
B.5 SIR error of the Schubert recording for the Multi-Excitation and Multi-Filter Model with estimated filters and trained filters for different block size values ................................................................. 66
B.6 SIR error of the Mozart recording for the Multi-Excitation and Multi-Filter Model with estimated filters and trained filters for different block size values ................................................................. 67
B.7 SAR error of the Schubert recording for the Multi-Excitation and Multi-Filter Model with estimated filters and trained filters for different block size values ................................................................. 68
B.8 SAR error of the Mozart recording for the Multi-Excitation and Multi-Filter Model with estimated filters and trained filters for different block size values
Chapter 1

INTRODUCTION

Computational analysis of audio signals where multiple sources are present is a challenging problem. This task, which is defined as Source Separation (SS), could be solved by applying methods that separate the signals of individual sources from the original mixture. This well known technique to isolate the components from a mixture has been demonstrated to be achievable.

The source separation problem is ubiquitous in many different application areas as, for example: Audio Processing, Image processing, Chemometris or Bioinformatics. In audio signal processing, there are many tasks where sound source separation can be used, but the performance of the existing algorithms is still limited compared to the human auditory system. Human listeners are able to perceive and discriminate individual sources in complex mixtures. Based on the sound segregation ability in humans several algorithms have been proposed. Specially, when the signals are composed by sources of similar characteristics, i.e. instruments of similar timbre, the estimation of an individual source from the acoustic signal mixture is disturbed by other coexisting sounds.

One of the divisions for source separation methods is done considering the prior available information. The different approaches can be divided into blind and non-blind. Blind Source Separation (BSS) denotes the separation of completely unknown sources without using additional information. Non-blind or Informed Source Separation (ISS) denotes the separation of sources for which further information is available, either for the individual sources or for the mixture.

More precisely, the work will focus on the separation for multiple instruments of similar timbre, having a String Quartet as a use case. In a set of instruments like the one here addressed there is an absence of a main or lead instrument and the harmonicity between sources is high.
The attempt of this thesis is to improve the quality of the separation. For some techniques the required information for the separation of audio signals is the pitch of each source. In complex mixtures one way of obtaining this information is performing a multi-pitch estimation. However, this task has turned to be a challenging problem too.

For this reason, Informed Source Separation techniques are considered in here. As the goal is to provide a potentially higher quality result, the multi-pitch estimation step is avoided. The additional information that is used contains information of the melody or pitch for each of the audio signal sources, (i.e. an aligned pitch-contour signal). This approach could be considered similar to the use of symbolic information of the pitch (e.g., an aligned score), an approach that has already been addressed in the literature under the name of Score-Informed Source Separation. In contrary, for the case addressed in this work, this kind of informed SS could be named as Pitch-Informed Source Separation.

For facilitating this task various multimodal datasets are available. Besides the general mixture audio recordings, even if not all the information is utilized, more prior information is available. For example: pitch information (scores, pick-pup recordings), gestures (vide recordings) or spatial cues (binaural recordings).

Moreover, The Music Technology Group of the Universitat Pompeu Fabra, where this thesis has been carried out, is the main coordinator of the Performances as Highly Enriched aNd Interactive Concert eXperiences (PHENICX) project. The MTG provides the project with expertise coming from different research areas, being audio source separation one of them. For this reason, this work could be potentially useful as a first exploration for future applications in this project.
Chapter 2

STATE OF THE ART

In this chapter, the theoretical background and a review on the relevant work for this thesis is presented. First, the general concepts and methods for source separation are introduced. Second, a more specific view on musical audio source separation is exposed. Finally, some of the physical/acoustic principles and modeling approaches for bowed string instrument are introduced.

2.1 Source Separation

Source separation is the challenging problem of extracting individual signals from an observed mixture by computational means. It is usefully applied in many different fields and types of signals such as image, video, medical, financial or radio. This work focuses on the separation of audio signals, and more specifically of music audio signals.

The source separation task was first formulated in the mid 1980’s within a statistical framework by [Hérault et al., 1985]. In the early 1990’s, the introduction of Independent Component Analysis (ICA) by [Comon, 1994] and the appearance of other related techniques set the ground for the development of the topic.

More specifically, in the field of auditory perception, [Bregman, 1990] proposed a cognitive approach for describing hearing complex auditory environments, which increased the interest of the research community. This process, named as Auditory Scene Analysis (ASA), explains the steps required for the human auditory system to analyze mixtures of sounds and recover descriptions of individual sounds. The work provided the basis for the computational implementation of algorithms that approximate the sound separation capabilities of the human auditory system, concept known as Computational Auditory Scene Analysis (CASA). At the same time, both works on ICA and ASA set the beginning of two new approaches to acoustic separation: statistical/mathematical and biolog-
ically inspired approaches. [Wang and Brown, 2006] presented a detailed literature review on the different approaches that have been proposed to deal with the computational modeling of ASA.

The difficulty of a source separation problem is mainly determined by three factors. First, the proportion between the number of mixture channels and the number of original sources determines it. The separation is easier if the observed mixture has more channels, or the same number of channels, than there are sources to separate. Second, the nature, i.e., complexity, of the mixture. Third, the amount of information about the sources or the mixture available a priori.

We don’t find a generalized agreement in the literature when assigning labels, neither according to all the different existing methods, nor according to the different degrees of knowledge available. However, nomenclature overviews proposed by authors like [Vincent et al., 2003] or [Buried, 2009] are reviewed in this section.

According to the proportion between the number of mixture channels and the number of original sources, Burred names the cases as over-determined, even-determined and undetermined source separation. He defines the task as over-determined when the observed mixture has more channels than sources. Even-determined (or determined) source separation corresponds to the case when there are the same numbers of channels than sources to separate. In mixtures with less channels than sources, named as undetermined, additional difficulties, such as stronger assumptions, are often to be addressed.

Depending on the a priori information (i.e., given knowledge) available about the sources or the mixture there is a distinction between blind, semi-blind and non-blind source separation. The task is said to be blind if there is little or no knowledge available before the observation of the mixture. Particularly, Blind Source Separation (BSS) has become a standard nomenclature in the literature to denote the statistical methods that approach this case. However, it is to be mentioned, that there is no strictly blind system due to the fact that, at least, some general probabilistic assumptions like statistical independence and sparsity must be taken. ICA and sparsity-based methods (e.g., time-frequency masking) are included in here. When talking about Semi-blind Source Separation (SBSS), he refers to methods that include sinusoidal models and supervised methods where the source models are learned in advance. At last, Non-blind or, as it will be referred in this document, Informed Source Separation (ISS), corresponds to systems that, besides the mixture, they have detailed high level information about the sources or mixture as an input. This information can be, for example, the number and kind of sources. More specifically, for musical source separation, the a priori information can be of the kind of the score, a MIDI sequence, pitch-contour or any other kind of detailed information of the melody been played. Also, another case contained in here is where the information is extracted from the original isolated
tracks that are encoded and embedded in the mixture (e.g., using watermarking). This information is later used in the separations stage to provide a higher quality separation [Liutkus et al., 2012]. ISS will be the case addressed in this work.

In the more specific context of audio source separation, Vincent proposed a division between Audio Quality Oriented (AQO) and Significance Oriented (SO) applications. On the one hand, AQO has the goal of fully unmixing the individual sources present in the mixture with the highest possible quality. In this case, the output signals are intended to be listened to. Some AQO applications, as the one to be addressed in this document, look for a fully unmixing of the original mixture into separated sources that are aimed to be listened separately. This turns to be the most demanding application scenario, as the goal is to obtain a multitrack recording equivalent to the one created for the final mix of the mixture. Ideally, the separated tracks should have the same or similar quality than they would have had if recorded separately.

On the other hand, SO methods don’t require such a high quality separation, since the goal is to facilitate a high-level, semantic feature extraction from the sources. SO involves many tasks from the Music Information Retrieval field. For example, polyphonic transcription, or the task of automatically extracting the music score from the mixture, is a high demanding task included in SO application. It is worth mentioning that, obviously, AQO methods are also valid for the SO context too.

### 2.1.1 Models and Methods

As mentioned before, source separation can be defined as the task of estimating one or more of the original source signals \( s_m(n) \) from the observed mixture signal \( x(n) \). When several sources are present simultaneously, the mixture \( x(n) \) can be defined as

\[
x(n) = \sum_{m=1}^{M} s_m(n)
\]

where \( s_m \) is the \( m^{th} \) source signal and \( M \) is the number of sources.

Several probabilistic approaches to source separation have been proposed in the literature. The techniques about to be presented are very versatile and can be likely used in various applications with different purposes.

**Signal Representation**

It has to be mentioned that the separation is usually not computed directly from the original representation of the signal. In contrary, the signal is transformed to
another kind of representation to fulfill certain goals. [Schmidt, 2008] describes the following four considerations to signal representation.

First, to make explicit the desired characteristics of the signal. A representation different from the original, e.g., the Fourier transform, which emphasizes the meaningful characteristics of it could help in the task of separating the sources.

Second, to introduce invariances with the goal to decrease the adverse characteristics that are known to be not relevant to separate the signals, e.g., the use of power spectrum that ignores the phase, which introduces an invariance to phase shift.

Third, most of the approaches have the common goal of reducing the dimensionality of the data in order to reduce the computational cost. An example on one of the most common techniques for dimensionality reduction, Principal Component Analysis (PCA) is presented in 2.1.1.

Finally, to allow signal reconstruction. A distinction between reversible (lossless) and non-reversible (lossy) is made. In both cases the signals are separated in the representation domain. In the lossless case, the representation is inverted to obtain the separated signals in the original domain. However, in the lossy case, the separated signals must be reconstructed in the original signal domain. A filtering applied to the signal mixture is done to achieve this. A common practice is to use time varying Wiener filters [Hopgood and Rayner, 2003]. The selection of the signal representation is important, since it determines how accurate the signal reconstruction could be.

**Principal Component Analysis, Independent Component Analysis and Independent Subspace Analysis**

Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Independent Subspace Analysis (ISA) are very popular dimensionality reduction techniques suitable for source separation.

Karl Pearson first introduced PCA in 1901. The main goal of the technique is to convert a set of possibly correlated variables into a new set of uncorrelated variables, named principal components. To do so, it tries to find the components with the largest possible variance. The variables in the new space are uncorrelated (i.e., orthogonal) and a linear combination of the original variables. The new set of variables has the same dimension as the original, however, a selection of a smaller set of principal components is performed to reduce dimensionality. Even though some information is discarded with the selection, PCA tries to minimize this error. As a signal decomposition technique the link with source separation is patent. Some other possible applications of PCA include image processing, data visualization and data compression. The work by [Jolliffe, 2002] provides a deep review on the specifics of the technique and applications.
As an extension to PCA, Independent Component Analysis tries to estimate the source data from other observed data, e.g. noise, by assuming that the sources are statistically independent. The main difference between ICA and PCA is that the former goes further than just looking for a second order independence and provides solutions that are not just orthogonal. Thanks to this assumption ICA can be applied to arbitrary time-series signals. However, it requires at least as many mixture observation signals as sources. This technique has been often applied to speech signal separation, where this independence assumption is truly fulfilled [Comon and Jutten, 2010]. When applying it to more complex signals like polyphonic music, the statistical independence assumption is unsuitable. Also, most commercial music recordings are composed of more than two sources, while the number of sensor mixtures is often limited to one, for monophonic, or two, for stereophonic recordings.

In addition, we also find in the literature another source separation method related to PCA and ICA, Independent Subspace Analysis (ISA). As an extension to ICA, ISA [Casey, 2000] identifies multiple independent spaces from a given data, with the advantage that relaxes the constraint of requiring at least as many mixture observation signals as sources.

It is important to remark that signal decomposition is very related to source separation. Actually, some of the approaches used for source separation, such as ICA, have also been successfully applied to signal decomposition.

**Non-negative Matrix Factorization**

Non-negative matrix factorization (NMF) has been widely used technique in source separation. It was initially proposed by [Paatero and Tapper, 1994] and has been later on extended by [Lee and Seung, 1999] and [Lee and Seung, 2006] as an unsupervised learning method.

It is distinguished from other methods by its use of non-negativity constrains that lead to a parts-based representation, since they only allow additive, not subtractive combinations. The main benefit of this factorization technique is its ability to decompose signals into objects that have a meaningful interpretation. In the case of musical audio signals, when applying it to the spectrogram, the resulting objects can correspond, for example, to individual pitches of each instrument. This kind of representation makes the analysis of complex signals significantly easier.

NMF is based in the decomposition of a matrix $X$ of size $N \times M$ which is approximated by the form $X \approx WH$, or

$$X_{i\mu} \approx (WH)_{i\mu} = \sum_{i=1}^{r} W_{i\alpha} H_{\alpha\mu} \quad \text{s.t.} \quad W, H \geq 0$$

(2.2)
where the $r$ columns of $W$ are the basis vectors of the mixture. Each column in $H$ is called an activation coefficient (i.e., gain). Each of the sources on the matrix mixture $X$ is represented with a linear combination of basis vectors. The dimensions of the matrix factors $W$ and $H$ are $N \times r$ and $r \times M$, respectively. In order to consider the product $WH$ a compressed form of the data in $X$, the rank $r$ of the factorization has to be chosen so that $(N + M)r < MN$.

NMF is related to previously explained techniques like PCA and ICA, that can all be written as a matrix factorization of the form $X \approx WH$. The main differences between these methods and NMF are the constrains placed on the factorizing matrices $W$ and $H$. In PCA, columns of $W$ and rows of $H$ have to be orthogonal (i.e., uncorrelated). In ICA, rows of $H$ are maximally statistically independent. As mentioned above, in NMF $W$ and $H$ are constrained to be non-negative.

More precisely, the goal is to find $W$ and $H$ that minimize the divergence between the data $X$, and the approximation, $WH$. In its basic form, NMF is usually sought through the minimization problem

$$\min_{W,H} \mathcal{D}(X|W,H) \quad W,H \geq 0,$$

where $\mathcal{D}$ is a cost function or divergence that measures the quality of the approximation defined by

$$\mathcal{D}(X|W,H) = \sum_{f=1}^{N} \sum_{n=1}^{M} d([X]_{fn}|[WH]_{fn}),$$

The most commonly used basic cost functions are the Euclidean distance,

$$d_{EUC}(x|y) = \frac{1}{2}(x - y)^2$$

and the Kullback-Leibler (KL) divergence

$$d_{KL}(x|y) = x \log \frac{x}{y} - x + y$$

[Lee and Seung, 2001] proposed two different multiplicative algorithms for NMF. The first algorithm can be shown to minimize the conventional least squares error (i.e., from the Euclidean distance) while the other minimizes the generalized KL divergence. They proved that the convergence is considerably faster with the use of multiplicative rules similar to the Expectation Maximization (EM) algorithm.

More recent approaches proposed by [Févotte et al., 2009] and [Lefevre et al., 2011] are based on the Itakura-Saito (IS) divergence, which is denoted by
As opposed to other cost functions like Euclidean distance or KL-divergence, the use of IS-divergence provides several advantages, e.g., is scale invariant and has a faster convergence. The IS-NMF is a model based in superimposed Gaussian components and is equivalent to maximum likelihood estimation of variance parameters. As the other cost functions, IS-NMF can also be performed using a gradient multiplicative algorithm whose convergence is observed in practice, though not proven.

\[ d_{IS}(x|y) = \frac{x}{y} - \log \frac{x}{y} - 1 \] (2.7)

2.2 Musical Audio Source Separation

In the following subsection the specific case of musical audio source separation is addressed. First, a general introduction to audio spectrogram factorization is presented. After this, two different models for instrument mixture separation are explained. The first one, an Instantaneous Mixture Model proposed by [Durrieu et al., 2009a] aims to separate the main instrument from a stereophonic mixture. The second one, a Multi-Excitation Model that makes use of instrument-dependent models to separate the different sources of multiple instrument mixtures presented by [Carabias-Orti et al., 2011] and [Rodriguez-Serrano et al., 2012].

For the specific case of musical audio source separation, we can express equation 2.1 as follows. When several sources are present simultaneously, the acoustic waveform \( x(n) \) of the observed time domain-signal is the superposition of the source signals \( s_m(n) \). The signal \( x(n) \) can be defined as

\[ x(n) = \sum_{m=1}^{M} s_m(n), \quad n = 1, ..., N \] (2.8)

where \( s_m \) is the \( m^{th} \) source signal at time \( n \), and \( M \) is the number of sources present in the mixture.

However, when NMF (see Section 2.1.1) is applied to audio signals, the non-negative data \( X \) is usually taken as the magnitude or power spectrogram on the signal. In this case, the basis functions \( W \) are the magnitude or power spectra and they are activated over time by the activation coefficients or amplitudes contained in \( H \). This decomposition is a good approach to separate the structure of audio signals, as it represents the combinations of spectral features that correspond to contributions of different sound sources over time.
2.2.1 Instantaneous Mixture Model

A system for separating the main instrument in stereophonic mixtures was presented by [Durrieu et al., 2009a]. This algorithm is an extension of a previous work by the same author, where the same methodology was used for monophonic mixtures [Durrieu et al., 2009b].

The algorithm introduces an approach for stereophonic source separation with the goal of separating a polyphonic musical mixture into two main sources: a main instrument track (predominant or solo melody) and an accompaniment track. The solo part is modeled using a source / filter model. The accompaniment follows a general instantaneous mixture of several components within a NMF framework.

Mixture Model

The musical audio signal $x$ is composed of two main contributions. The first one, $v$ (for voice), represents the main instrument and the second one, $m$ (for music), the accompaniment. The mixture is assumed to be instantaneous, where $x = v + m$. The spectrogram or Short Time Fourier Transform (STFT) matrices are

$$X = V + M$$

(2.9)

The fact that the model is applied to a stereophonic signal is solved by considering it as a panning effect. The original sources are considered monophonic and distributed into each of the two channels to simulate their spatial positions. The solo $V$ is further assumed to have only one spatial position. Each of the several components $J$ of the accompaniment is assumed to have its own spatial position. All the spatial positions are also considered to be static. The STFTs, at frequency $f$ and frame $n$ are defined by:

$$\begin{align*}
X_{R,f,n} &= \alpha_{R}V_{R,f,n} + \sum_{j=1}^{J} \beta_{Rj}M_{Rj,f,n} \\
X_{L,f,n} &= \alpha_{R}V_{L,f,n} + \sum_{j=1}^{J} \beta_{Lj}M_{Lj,f,n}
\end{align*}$$

(2.10)

where $V_{R}, V_{L}, M_{Rj}$ and $M_{Lj}$ are supposed realizations of mutually and individually independent random variables across frequency and time.

Source / Filter Model of the Main Instrument

The solo part is assumed to be composed by a monophonic and harmonic instrument. The model is well adapted for describing speech or singing voice, however, due to its generality, it may be also applied for other musical instruments.

The source of the model is closely related to the pitch or melody of the main voice as it follows a KLGLOT88 glottal source model [Klatt and Klatt, 1990]. The filter acts as an envelope that reshapes the source comb to approximate the
timbre. Thus, it is more related to the timbre characteristics. More specifically, the variance of the solo voice $S_{V,fn}$ is inspired by a speech processing source / filter model. It is parameterized as follows:

$$S_{V,fn} = S_{\phi,fn} S_{F,fn}$$  \hfill (2.11)

been $S_{\phi,fn}$ the source contribution and $S_{F,fn}$ the filter contribution of the variance. The entries on their respective matrices $S_V$, $S_{\phi}$ and $S_{F_0}$ are also defined.

First, only voiced (V) components of the source were considered. A voiced / unvoiced (VU) extension of the model, that seems to lead to better results, was introduced afterwards. The source variance $S_{F_0}$ is modeled as a non-negative linear combination of all the allowed fundamental frequencies. Having the spectra $W_{F_0}$ and activation coefficients $H_{F_0}$, it can be defined as:

$$S_{F_0} = W_{F_0} H_{F_0}$$  \hfill (2.12)

For the filter, a dictionary $W_{\phi}$ and its activation coefficients $H_{\phi}$ are defined such that $S_{\phi} = W_{\phi} H_{\phi}$. An additional smoothness constraint on its frequency responses was also introduced. These frequencies are also modeled as a non-negative combination of a smooth filter dictionary $W_{\Gamma}$ and its activations $H_{\Gamma}$.

$$S_{\phi} = W_{\phi} H_{\phi} = W_{\Gamma} H_{\Gamma} H_{\phi}$$  \hfill (2.13)

By considering the full source and filter model, the source variance matrix $S_V$ can be defined as

$$S_V = S_{\phi} S_{F_0} = (W_{\Gamma} H_{\Gamma} H_{\phi}) \cdot (W_{F_0} H_{F_0})$$  \hfill (2.14)

where "·" is the Hadamart product, a pointwise multiplication between the matrices.

Since the source part can be seen as the instantaneous mixture of all the possible notes, with amplitudes activated by the coefficients $H_{F_0}$, this solo voice model, ant the general model, is referred as the Instantaneous Mixture Model (IMM). A general diagram of the model can be seen in figure 2.1.
**Accompaniment Model**

The accompaniment is modeled as an instantaneous mixture of $J$ components. For each of them the variance is modeled as a centered Gaussian.

$$S_{M,j,fn} = w_{fj} h_{jn}$$  \hspace{1cm} (2.15)

For each channel $C \in \{R, L\}$, the global variance of the accompaniment can be computed as

$$S_{M,C,fn} = \sum_{j=1}^{J} \beta_{Cj}^2 S_{M,j,fn} = [W_M B_C H_M]_{fn},$$  \hspace{1cm} (2.16)

where $W_M$ is a dictionary matrix, $B_C$ equals $\text{diag}(\beta_{Cj}^2)$ and $H_M$ is the amplitude coefficient matrix.

**Parameter Estimation**

Maximum Likelihood (ML) is used to estimate the parameters. Under the Gaussian assumption, ML is equivalent of minimizing the Itakura-Saito divergence (see 2.1.1) between the power spectrum $|X|^2$ of the observed STFT and the parameterized variance $S_X$, defined as $S_X \in \{S_XR, S_XL\}$. The variances for the left and right channel are given by the equations 2.10, 2.14 and 2.16:

\[
\begin{align*}
S_{xR,fn} &= \alpha_R^2 [(W_T H_T H_\phi) \cdot (W_{F0} H_{F0})]_{fn} + [W_M B_R H_M]_{fn} \\
S_{xL,fn} &= \alpha_L^2 [(W_T H_T H_\phi) \cdot (W_{F0} H_{F0})]_{fn} + [W_M B_L H_M]_{fn}
\end{align*}
\]  \hspace{1cm} (2.17)

The parameters $\Theta = \{H_T, H_\phi, H_{F0}, W_M, H_M, \alpha_R, \alpha_L, B_R, B_L\}$ are estimated by multiplicative update rules. The source dictionary $W_{F0}$ is fixed and generated with the previously mentioned glottal model. The smooth elements of the filter $W_T$ are set by using overlapping Hann functions that cover the whole frequency spectrum.

The algorithm proposed to estimate the parameters consists of four main steps: One, 1st parameter Estimation Round (ER); two, melody tracking; three, 2nd parameter ER; and last 3rd parameter ER (see figure 2.2). Each of the steps in the process is explained as:

First, the set of parameters $\Theta_0$ are randomly initialized.

Second, a smooth path for the fundamental frequencies is computed. Using a Viterbi algorithm, the corresponding $H_{F0}$ activation coefficients are extracted.

Third, the parameters are again randomly initialized. The new parameter $\tilde{H}_{F0}$ is an exception, in contrast, it is obtained by setting to zero the coefficients of $H_{F0}$ lying outside a scope of a quarter tone around the tracked melody (step two).
After this second ER round, a first solo/accompaniment separation result V-I MM is obtained (only Voiced parts of the solo are considered).

At last, in the 3rd ER, the initial parameters are the ones estimated in the 2nd ER, except for $W_{F_0}$, where an unvoiced basis vector (uniform value for all the frequencies) is added. By fixing the filter dictionary $W$, it is assumed that the unvoiced parts of the solo instrument are generated by the same filters that generate the voiced parts. By doing this the separation VU-I MM is obtained (voiced and unvoiced parts of the solo are now considered). The multiplicative update rules for the parameters can be revised in [Durrieu et al., 2009a].

**Source Separation using Wiener Filters**

Wiener filters are used to produce an estimate of a target random process by filtering another random process through the filter. These filters provide the minimum mean square error (MMSE) between the estimated random process and the desired process. Thanks to the independence assumption, the frequency response of the MMSE estimator obtained with the Wiener filter can be defined as:

$$G_V(f) = \frac{S_V(f)}{S_V(f) + S_M(f)}$$  \hspace{1cm} (2.18)
For the final separation, with the set of parameters $\Theta$ obtained by the algorithm, $S_V$ and $S_M$ are estimated for each frame. Then, following equation 2.18, the corresponding Wiener filter is computed. Finally, applying an overlap-add procedure, $\hat{v}$ (i.e., solo track) and $\hat{m}$ (i.e., accompaniment track) are reconstructed.

### 2.2.2 Multi-Excitation Model

The work presented by [Carabias-Orti et al., 2011] and [Rodriguez-Serrano et al., 2012] proposed an approach to model the excitations of musical instruments. These excitations represent vibrating objects, while the filter represents the resonance structure of the instrument that colors the produced sound. The work focuses on modeling the excitations as the weighted sum of harmonic basis functions, whose parameters are tied across different pitches of an instrument. An NMF-based framework is used to learn the model parameters. As explained in section 2.1.1, NMF approaches try to decompose the audio spectrogram of a signal (see in equation 2.8) as a linear combination of spectral basis functions as

$$\hat{x}_t(f) = \sum_{n=1}^{N} g_{n,t} b_n(f)$$

where $g_{n,t}$ is the gain of the basis function $n$ at frame $t$, and $b_n(f), n = 1, ..., N$ are the bases.

### Previous modeling approaches and limitations

As explained in the previous sections, several types of methods have been proposed in the literature for estimating this kind of decomposition. Further, it is possible to constrain the most commonly applied NMF model with some priors such as harmonicity, temporal continuity and sparsity.

For instruments in which the partials of each pitch have a smooth distribution, a basic Harmonic Comb (HC) model has turned to be a solution. The elements in the basis $b_{n,j}(f)$ are approximated by this harmonic shape. The magnitude of the STFT for the HC model is estimated as

$$\hat{x}_t(f) = \sum_{j=1}^{J} \sum_{n=1}^{N} g_{n,t,j} \sum_{m=1}^{M} a_{n,m,j} G(f - m f_0(n))$$

where $m = 1, ..., M$ is the number of harmonics, $a_{n,m,j}$ the amplitude for the $m$-th partial of the pitch $n$ and instrument $j$, $f_0(n)$ the fundamental frequency of pitch $n$, $G(f)$ the magnitude spectrum of the window function and $J$ the number
of instruments. The parameters that have to be estimated by the NMF algorithm are the time activation coefficients or gains $g_{n,t}$ and the pitch amplitudes $a_{n,m,j}$.

A fundamental problem of this model is that the spectra of certain instruments, like woodwinds or bowed-string instruments, have a specific structure of non-smoothed patterns that varies as a function of the pitch and excitation intensity. For this reason, modeling their spectra as the sum of harmonic elementary functions, or using harmonic excitations that are flat as the function of frequency, turns not to be sufficient. In order to model the whole pitch range of these instruments the use of more advanced techniques is required.

Excitation or source/filter models, like the ones presented in section 2.2.1 or proposed by [Virtanen and Klapuri, 2006a], try to represent all the possible pitch and instrument combinations, while keeping a reduced amount of parameters. The excitation models the time-varying pitch produced by a vibrating element (e.g., a piano string) while the filter models the unique resonant structure of the instrument (e.g., the piano soundboard), which colors the radiated sound.

However, it has been shown that the total admittance of two connected systems, like as a string and a body, is more complex than the product of the admittances of the parts [Woodhouse, 2004]. Therefore, the sound production in the actual physical system cannot be exactly modeled as the product of an excitation and a filter. Though, practical applications of the excitation-filter model have turned out to be a sufficient approximation.

**Harmonic Multi-Excitation Model**

As an extension of the previously explained methods, [Carabias-Orti et al., 2011] introduced a new approach that models the excitation as the weighted sum of harmonic excitation basis functions, with parameters as a function of each instrument, and weights that vary as a function of pitch.

For the excitations, it is assumed that a certain instrument should provide an identifiable characteristic structure. For the filter, the conditions of the music scene might produce variations in it. Therefore, the excitations of each musical instrument and the filter are learned in a training stage. The filter parameters can be updated in testing to adapt the model to the conditions of the music scene. The excitations can be adapted to the variations that the differences between the physical properties of each instrument may introduce.

**Excitation**: The spectral basis functions of a generic excitation model can be represented as

$$b_{n,j}(f) = h_j(f)e_{n,j}(f). \quad (2.21)$$
A generic harmonic excitation is defined as

\[ e_{n,j}(f) = \sum_{m=1}^{M} a_{m,n,j} G(f - mf_0(n)) \]  \hspace{1cm} (2.22)

where \( a_{m,n,j} \) is the amplitude of the harmonic \( m \), pitch \( n \) and instrument \( j \). The author proposes the modeling of the amplitudes as a linear combination of \( I \) excitation basis vectors \( v_{i,m,j} \) as

\[ a_{m,n,j} = \sum_{j=1}^{J} w_{i,n,j} v_{i,m,j} \]  \hspace{1cm} (2.23)

where \( w_{i,n,j} \) is the weighted sum of the \( i \)th excitation basis vector for pitch \( i \) and instrument \( j \). In this case, the excitations are unique for each instrument and partial but they are shared across pitches. The weights, though, are unique for each instrument and pitch but shared between harmonics or partials. By substituting equation 2.23 into 2.21, we obtain that the harmonic excitation functions can be defined as

\[ e_{n,j}(f) = \sum_{m=1}^{M} \sum_{i=1}^{I} w_{i,n,j} v_{i,m,j} G(f - mf_0(n)) \]  \hspace{1cm} (2.24)

**Filter:** To obtain the spectral basis functions defined in 2.21, the harmonic excitation functions are multiplied by the instrument filter. The model for a magnitude spectrum of the whole signal \( x(t) \) is given by the sum of instruments and pitches as

\[ \hat{x}_r(f) = \sum_{n,j} g_{n,t,j} h_j(f) \sum_{m=1}^{M} \sum_{i=1}^{I} w_{i,n,j} v_{i,m,j} G(f - mf_0(n)) \]  \hspace{1cm} (2.25)

where \( n = 1, \ldots, N \) (with \( N \) the number of pitches) and \( j = 1, \ldots, J \) (with \( J \) the number of instruments), \( M \) the number of harmonics and \( I \) the number of considered excitations.

The use of a reduced number of excitation bases turns to reduce significantly the parameters of the model and benefits their learning. This Multi-Excitation model is able to produce non-smoothed bases that match better the shape of the original spectrum.
Parameter estimation

For estimating the parameters of the model an NMF algorithm is proposed. To minimize the cost function defined in equation 2.4, the Kullback-Leibler divergence (see equation 2.6) is used. The multiplicative update rules, which minimize the divergence, can be revised in [Carabias-Orti et al., 2011]. In practice, as learning the parameters from a polyphonic mixture is a difficult task, recordings with a full pitch range of each instrument (with isolated notes) are provided.

Taking advantage of the temporal gain continuity over gains, an inherent property in musical instruments, the number of note insertion errors caused by interferences is reduced. For enforcing this temporal continuity of the gains, a Gamma chain presented in [Virtanen et al., 2008] is used.

With the presented model, pitch-dependent excitations and instrument-dependent body responses can be estimated.

2.3 Bowed String Instrument Modeling

Bowed string instruments are a subcategory of string instruments that are played by a bow rubbing the strings. Of all the instruments of Western music, the family of bowed strings is perhaps the most important and most studied. The instruments produced by the Italian masters of the 17th century in Cremona-Amati, Stradivari and others are taken to define the style and quality that modern instruments aim to reproduce.

The bowed string instruments that will be addressed in this work are enclosed into the violin family and they consist of just four instruments: violin, viola, cello and double bass. More precisely, the case to be addressed consists of a string quartet, a musical ensemble of four string players, usually composed by two violin players, a violist and a cellist. The string quartet is one of the most prominent chamber ensembles in classical music.

2.3.1 Physics and Models of Violin Family Instruments

Four main basic parts of the violin family instruments can be distinguished: The strings, bridge, bow and body.

The strings of a violin are stretched across the bridge and nut of the violin so that the ends are essentially stationary, allowing for the creation of the mentioned standing waves. The fundamental frequency and harmonic overtones of the resulting sound depend on the material properties of the string, such as the tension, length, mass, elasticity and damping factor.
The bridge supports one end of the strings playing length and transfers vibration from the strings to the top of the violin. The most significant bridge motion is side-to-side rocking, coming from the transverse component of the strings’ vibration.

The bow is generally the one in charge of the excitation of string vibration. It consists of a flat ribbon of parallel hairs stretched between the ends of a stick. They are usually made of wood or synthetic materials, such as fiberglass or carbon-fiber composite. The length, weight, and balance point of modern bows are standardized. The three most prominent factors that the player can control are bow speed, downward force, and location of the point where it crosses string.

The body of a violin family instrument consists of two arched wooden plates as top and bottom of a box. An internal sound post helps transmit sound to the back of the violin and serves as structural support. It acts as a sound box to couple the vibration of strings to the surrounding air, making it audible. The construction characteristics of this soundbox have a huge influence in the overall sound quality of the instrument [Schelleng, 1974].

The core of the violin or any of its relatives (viola, cello, bass) is the bowed string. The string plays a major role in establishing the musical identity of this family of instruments. Conceptually, the string is the simplest of the components. However, the action of the string under the bow presents many unanswered questions. One of the first attempts to understand the phenomena when string was bowed was made by Hermann von Helmholtz. With the use of a vibration microscope, what we nowadays will name an oscilloscope, he acquired the basis for a mathematical description of the motion of the string as a whole [Helmholtz, 1860].

Two simple physical facts underlie the action of the bowed string. First, relies on the fact that the sliding friction is smaller than the static friction and the change from one to the other is almost discontinuously abrupt. Second, that the flexible string in tension has a succession of natural modes of vibration whose frequencies are almost exact whole-number multiples of the lowest frequency. Without outside compulsion, the string is therefore by its very nature given to produce a quasi-periodic sound pressure wave.

According to the models, a linear harmonic model has guided a good deal of contemporary violin research [Fletcher, 1999]. The complex frequency envelope that describes the behavior of the violin body is usually represented by a simple transfer function. The easiest measurement to interpret is the mechanical admittance (velocity to force ratio) at the bridge of the instrument where the string rests.

As seen in Figure 2.3, for the violin, the lowest peak is associated with the air mode coupled in phase with the lowest body mode. Higher resonances are associated with an out of phase coupling of these two modes and with the direct
Figure 2.3: Resonance curve of a typical violin body defined as the mechanical input admittance at the bridge [Fletcher, 1999]

coupling to higher body modes.

The spectral envelope of the curve generally shows a dip between 1 and 2 kHz, a broad peak around 3 kHz, and then a smooth decline at higher frequencies. This envelope effectively defines good classical violin quality. When input impedance curves of other bowed-string instruments are compared with that of the violin, it is seen that they are not simply larger and lower-pitched versions of the violin, but have their own musical signatures.

As seen, the vibrational behavior of the body of a bowed-string instrument is closely linear, however, the frictional mechanism that drives the bowed string is highly nonlinear.

One of the main characteristics from an acoustical point of view is the transverse force applied to the bridge by the vibrating string. This has the form of a sawtooth wave, having the bowing position no influence. Fourier analysis of this force waveform shows a spectrum containing all harmonics of the fundamental, with the amplitude of the \( n \)th harmonic varying as \( 1/n \), corresponding to a spectral envelope decrease of 6 dB per octave.

### 2.3.2 Source / Filter Modeling

Source / Filter models have also been successfully applied to represent musical instruments. The regularity of the above explained spectrum gives some hint of why a simple source plus linear resonator model works for bowed-string instruments. In the violin family instruments, in a simple way, the vibrating string(s) can be considered the source, which is normally excited by the bow, and the
sound box together with the bridge can be approached as the filter. The generalized model proposed by [Hahn et al., 2010] is also applied for the violin. The source-filter-model aims to represent the time-varying spectral characteristics of quasi-harmonic instruments. For this, it is composed of an excitation source, generating sinusoidal parameter trajectories, and a modeling resonance filter, whereas basic-splines are used to model continuous trajectories.

The work by [Virtanen and Klapuri, 2006b] proposes a method where the input spectrogram of polyphonic audio is modeled as a linear sum of basis functions with time-varying gains. Each of the basis is represented as a product of a source spectrum and the magnitude response of a filter. This modeling can also be found in the source separation literature for the specific case of performing musical instrument recognition [Heittola et al., 2009]. In this approach, the mixture signal is decomposed by NMF and modeled as a product of excitations and filters. The excitations are restricted to harmonic spectra and their fundamental frequencies are estimated in advance using a multipitch estimator, whereas the filters are restricted to have smooth frequency responses by modeling them as a sum of elementary functions. More precisely, the filter consists of elementary triangular bandpass magnitude responses, uniformly distributed on the Mel-frequency scale.

This previous works show that source-filter models outperform the standard NMF representations for singing voice signals and polyphonic audio. One of the advantages of this formulation is the reduction the number of free parameters needed to approximate the audio signals, which leads to a more reliable parameter estimation.
Chapter 3

METHODOLOGY

After understanding and acknowledging the advantages and limitations of current state of the art algorithms in source separation, our hypothesis is that it is possible to provide external information, such as the melodies or timbral characteristics of each of the instruments, in order to enhance their results. Such steps will be investigated following the methodology presented in this chapter.

First, the initial approach and the modifications of the previously explained Instantaneous Mixture Model (IMM) are explained. Second, the Multi-excitation Model and the proposed modifications are exposed. In here, both the modifications to the source and the filter representations are introduced.

3.1 First approach: Modifications over the IMM

Having the previously presented Instantaneous Mixture Model as a starting point (see Section 2.2.1), some modifications to match our requirements have been introduced. The original model is designed to separate a single predominant or lead instrument, being explicitly focused on the singing voice. In contrast, we mainly want to perform the separation of bowed-string instruments.

3.1.1 Source Model

Excitation bases

The source variance of the model is parameterized as a combination of its bases $W_{F_0}$ and the time-varying activation coefficients $H_{F_0}$ as expressed in equation 2.14. Since it is mainly focused on separating the singing voice, the bases of the source are defined by a previously exposed KLGLOT88 glottal source model. Specifically, it is modeled as a non-negative linear combination of the spectral combs of all the NF0 possible (allowed) fundamental frequencies. An example of
Figure 3.1: Example of an $W_{F_0}$ basis generated with the glottal source model

A generated spectral comb base of the glottal source can be seen in figure 3.1. As we want the excitation bases to be able to fit other instruments, and not only the singing voice, the excitations are modified to follow a flat pattern. That is, a potentially more flexible and less restrictive solution is chosen. The new source excitation bases are described by impulse trains that cover the allowed $F_0$ frequency range. The excitations are restricted to harmonic spectra and their fundamental frequencies are obtained as follows.

**Excitation gains**

In order to obtain the $H_{F_0}$ gains or activation coefficients, the model estimates the main melody over the mixture. As it will be explained later, our proposed model aims to separate as many instrument as are present in the mixture, not only the predominant or lead one. Following the procedure of the original model involves the estimation of all the fundamental frequencies of the various instruments over the mixture. In order to skip the difficulties that actual multipitch estimation algorithms entail, the pitch or fundamental frequency estimation of each of the instruments present in the mixture is provided to the system. That is, as a first modification, the fundamental frequency estimation step is avoided. Instead, it is obtained from prior information and later provided to the system. As we have exposed, this approach is known in the literature as Informed Source Separation, since more information than the audio mixture is provided to the system.
To obtain this prior knowledge, different information like the musical score, isolated pick-up recordings or even the multitrack recordings have to be available. Musical score following and audio alignment is a broad topic of research and is not intended to be covered in this work. In order to more simply obtain the required data to inform the separation and focus on developing other aspects of the model, the pitch information is directly estimated from individual instrument recordings. The nature of this recordings can be different. Thanks to the availability of multimodal datasets, they are recently more and more accessible. On the one hand, by means of pick-up piezoelectric microphones located in the body of the instruments, the clean vibrations that the sound produces are directly captured. Pitch estimation over this signals is simpler to perform than from the complex mixture, where more instruments are present too. The obtained signals provide a clean representation of the melody being played. On the other hand, when multitrack recordings are available, the pitch estimation can be directly performed from this recordings. A pitch estimation algorithm will perform better in this isolated monophonic recordings than in the complex polyphonic mixture. As opposed to the musical score, the fundamental frequency obtained from this signals is more complete, since it contains information that it’s not usually present in the former, such as vibratos. Besides, the melody lines are obtained directly from the aligned audio, so there is no need of performing any other alignment process.

More precisely, we use Yin [De Cheveigné and Kawahara, 2002], a state-of-the-art predominant f0 estimation algorithm, over this isolated recordings to obtain the prior melodic information. Finally, the source / filter model is provided with the fundamental frequency of the lead instrument (see figure 3.2). Thanks to the obtained pitch information, the $H_{F_0}$ gains of the source are fixed. In this manner, the gains are not longer estimated in the separation process.

Summing up, the excitations are restricted to harmonic spectra and their fundamental frequencies are obtained in a prior step to the separation. When combin-
ing both $W_{F_0}$ bases and $H_{F_0}$ activations the source variance $S_{F_0}$ is obtained. This source can be interpreted as the different pitch values (notes) of the instrument to be separated.

### 3.2 Proposed Extensions: Combining IMM and Multi-Excitation Model

As mentioned before, one of the goals of the system is to be able to separate as many instruments as present in the audio mixture. However, the original proposal of the IMM is limited to separate only the predominant instrument. To allow the full separation of the mixture, the IMM has been extended with two modifications inspired on the Multi-Excitation model (see Section 2.2.2). First, each of the instruments source is modeled individually. Second, each of the instruments has a source and a filter model assigned to it. The extensions to the model are explained below.

#### 3.2.1 Multi-Excitation and Single Filter IMM

Besides keeping the sources flat, a modification that adds one excitation for each of the instruments present in the mixture is introduced. With this extension, the model assigns now as many sources as instruments are present in the mixture. More specifically, the model generates a source for each of the different fundamental frequency or melody lines provided. The filter sub-model is shared among all the sources.

In the case where prior information to estimate the pitch of one or more of the instruments is not available, the separation of this instrument is not performed. The decomposition of this instrument or instruments is not done with the source / filter model, but it is represented by the general NMF accompaniment sub-model. In other words, if the pitches of all the four instruments of a string quartet are estimated in a previous step and provided to the system, all of the four instruments are separated. However, lets assume that from a mixture of four instruments only the pitch information of two of them, e.g. the cello and the viola, is provided. Each of this instruments is assigned with its own source model and the other two, in this case both violins, are fitted into the accompaniment sub-model.

The source variance previously defined by equation 2.12 can be redefined for this extension as

$$S_{F_0} = \sum_{i}^{N_I} S_{F_0}^i = \sum_{i}^{N_I} W_{F_0}^i H_{F_0}^i$$

(3.1)
where $S_{Fi}$ is now the source variance of instrument $i$. Each of the instruments provided pitch $i$ is assigned to $H_{Fi}$. The bases $W_{Fi}$ are generated according to the estimated fundamental frequencies.

Having this modification in the source into account, the global variance can be redefined now as

$$\hat{V}_{mesfIMM} = \sum_{i} \left( (W_{Fi}^\dagger H_{Fi}) \cdot (W_{Fi}^\dagger H_{Fi}) \right) + W_M H_M$$

(3.2)

The update rules for the parameter estimation of this method remain the same as for the specified for the IMM (see Section 2.2.1). In contrast, for this new case the parameters to be updated are increased, since the $H_{Fi}$ matrix is different for each source.

According to the filter model, even if one excitation for each of the instruments is now defined, only one filter for all of them is assigned. Since this model is derived by combining the IMM and the Multi-Excitation models, it will be addressed as the Multi-Excitation and Single Filter IMM. A general diagram of the new proposed model can be seen in figure 3.3.

![Figure 3.3: Multi-Excitation and Single Filter IMM diagram](image)

### 3.2.2 Multi-Excitation and Multi-Filter IMM

As a second step, the filter model is extended to have multiple source and filter models. This proposed model assigns one specific filter to each instrument excitation.

Each of the instruments is modeled now with its own excitation and filter. For this reason, the filter bases and activations have to be estimated for each of the instruments too. The filter variance previously defined by equation 2.13 can be redefined for this extension as
\[ S_\phi = \sum_{i} S^i_\phi = \sum_{i} W^i_\phi H^i_\phi = \sum_{i} W_\Gamma H^i_\Gamma H^i_\phi \quad (3.3) \]

where \( S^i_\phi \) is now the source variance of instrument \( i \). The \( W_\Gamma, H^i_\Gamma \) and \( H^i_\phi \) matrices are different for each source.

With the filter sub-model extension, the global variance can be redefined now as

\[ \hat{V}_{mefIMM} = \sum_{i} ((W^i_{F_0} H^i_{F_0}) \cdot (W_\Gamma H^i_\Gamma H^i_\phi)) + W_M H_M \quad (3.4) \]

The update rules for the parameter estimation of this method remain the same as for the specified for the IMM (see Section 2.2.1). In contrast, as in the previous extension, the parameters to be updated are increased. As mentioned, the \( H^i_\Gamma, H^i_\phi \) and \( H^i_{F_0} \) matrices are different for each source sound. Each of the instruments provided pitch \( i \) is assigned to \( H^i_{F_0} \) and its bases \( W^i_{F_0} \) are generated accordingly. When a target source is estimated from the mixture spectrogram only the activations \( H^i_\Gamma, H^i_\phi \) and \( H^i_{F_0} \) of the corresponding pitch are used.

![Multi-Excitation and Multi-Filter IMM diagram](image)

**Figure 3.4: Multi-Excitation and Multi-Filter IMM diagram**

Summing up, one excitation and one filter for each of the instruments is now assigned. Since this model is derived by combining the IMM, Multi-Excitation model and a multiple filter model, it will be addressed as the Multi-Excitation and Multi-Filter IMM. An overview of the new proposed model can be seen in figure 3.4.
3.3 Supervised Timbre Models

In order to reduce the amount of parameters to be estimated and to facilitate the separation, a new method to obtain the filter bases from a previous training process is proposed. The prior training of different filter bases that represent the timbre characteristics of each instrument is here introduced.

The filter sub-model is specified in equation 2.13 and redefined for the proposed Multi-Excitation and Multi-Filter IMM in equation 3.3. A decomposition of the \( W_\phi \) bases in two other matrixes is introduced (see figure 3.5). The smooth filter dictionary \( W_i^\phi \) is originally defined in the IMM as part of a production model for instrument \( i \). Applying it to the Multiple-Excitation and Filter IMM, this bases are explained as a new combination of smooth filter bases \( W_\Gamma \) and its activations \( H_i^\Gamma \) for each instrument \( i \). The smoothness of the filters is more realistic than having unconstrained filters. Rather than a direct improvement in the source separation, this decomposition allows to learn the spectral shapes that are characteristic for a given instrument in this proposed supervised framework.

![Figure 3.5: Timber model \( W_\phi \) bases decomposed into an smooth filter dictionary bases \( W_\Gamma \) and its gains \( H_\Gamma \).](image)

The \( W_\Gamma \) smooth filters are fixed for all the instruments, they are composed by overlapping Hann functions covering the whole frequency range. Only the desired number of smooth filters has to be provided to the system, a parameter that can be considered an equivalent to the frequency resolution of the filter representation. When more filters are defined, the spectral shape or timbre can be more precisely represented.

To obtain each of the different instrument filter models a recording where only audio content of this specific instrument has to be provided. Over this data, the filter smooth dictionary bases \( W_\Gamma \) are generated, whereas the filter dictionary activations \( H_i^\Gamma \) are estimated and saved. This way, the combination of both bases and activations result in some bases (timbre bases) that can be seen as a model
containing the timbre characteristics of the instrument in the provided data. Examples of the different generated and estimated matrices can be seen in Figures 3.6 and 3.7.

Figure 3.6: Timbre matrices example over a cello excerpt. Smooth filter dictionary bases $W_{\Gamma}$ (top), its activations $H_{\Gamma}$ (middle) and the obtained bases $W_{\phi}$

The obtained timbre model is now used in the separation process. When separating the sources of a mixture where a certain instrument $i$ is present, the previously created model of this instrument $i$ can be loaded into the system. The $W_{\Gamma}$ bases are again generated. The $H_{\Gamma}$ activations are loaded from the $i$ instrument timbre model. Both matrices create the smooth filter dictionary $W_{\phi}$ that is kept fixed during the separation. However, the $H_{\phi}$ activations of this dictionary, since they are still depend of the content of the mixture, have to be still estimated.

According to the update rules, the parameter estimation of this method remains the same as for the proposed Multi-Excitation and Filter IMM (see Section 3.2.2). However, as the $H_{\Gamma}$ are loaded from the previously supervised timbre models they don’t have to be estimated and are kept fixed through the process.

In other words, one excitation and one filter for each of the instruments is assigned. In the filter part, a previously supervised timbre model is provided for
Figure 3.7: Timbre matrices example over a violin excerpt. Smooth filter dictionary bases $W_\Gamma$ (top), its activations $H_\Gamma$ (middle) and the obtained bases $W_\phi$ for each instrument. Figure 3.8 shows a general diagram of the Multi-Excitation and Multi-Filter IMM where the supervised filters are loaded.

Figure 3.8: Multi-Excitation and Multi-Filter IMM diagram with supervised filters loaded
Chapter 4

EVALUATION AND RESULTS

4.1 Evaluation Metrics

An intuitive way to assess the quality of a source separation algorithm could be to listen to the extracted sounds and to evaluate them with our own subjective criteria. However, in order to obtain more general results and to allow a fair comparison between different methods, we need to compute some objective evaluation measures. Different approaches have been proposed to calculate these evaluation metrics, all based on the comparison of the extracted sources with the original ones. The advantage of these evaluation techniques is to be completely impartial and to provide interpretable results. Their inconvenient side is that they require the original separated sources of the signals we intend to separate. Even if multitrack recordings are more and more available nowadays, the data usable for experiments is still dramatically reduced.

4.1.1 The BSS EVAL Toolbox

The BSS EVAL toolbox [C. Févotte et al., 2005], which stands for Blind Source Separation Evaluation, allows to compute objective performance measures. This metrics are the ones created and proposed by the Signal Separation Evaluation Campaign (SiSEC)\(^1\), a scientific evaluation contest where speech and music datasets are evaluated and standardized. This metrics evaluate the performance of the separation i.e. how well the estimated sources \(\hat{s}_j\) are approximated to the given true sources \(s_j\) [Vincent et al., 2007]. The estimated sources are decomposed as

\[
\hat{s}_j = s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}
\]

\(^1http://sisec.wiki.irisa.fr/\)
The \( s_{\text{target}} = q(s_j) \) is a modified version of \( s_j \) with an allowed distortion \( q \) that represents the part of \( \hat{s}_j \) coming from the target source \( s_j \). The \( \varepsilon_{\text{interf}}, \varepsilon_{\text{noise}}, \varepsilon_{\text{artif}} \) are the interferences, noise and artifacts error terms that represent the parts of \( s_j \) coming from the unwanted sources, from sensor noises and from other artifacts, respectively. The performance metrics are defined as energy ratios (in dB). Three main metrics are defined.

- **Source to Distortion Ratio:**

  \[
  SDR = 10 \log_{10} \frac{\| s_{\text{target}} \|^2}{\| \hat{s}_j - s_{\text{target}} \|^2} \quad (4.2)
  \]

- **Source to Interference Ratio**, which considers sounds from other sources:

  \[
  SIR = 10 \log_{10} \frac{\| s_{\text{target}} \|^2}{\| \varepsilon_{\text{interf}} \|^2} \quad (4.3)
  \]

- **Source to Artifacts Ratio**, related to the artifacts produced by the separation:

  \[
  SAR = 10 \log_{10} \frac{\| \hat{s}_j - \varepsilon_{\text{artif}} \|^2}{\| \varepsilon_{\text{artif}} \|^2} \quad (4.4)
  \]

The three performance measurement tools defined in this toolbox are inspired in the traditional measure of Signal to Noise Ratio (SNR), thus, the interpretation of the metrics is not complex.

### 4.1.2 The PEASS Toolkit

A further possibility is to use the Perceptual Evaluation methods for Audio Source Separation (PEASS) toolkit as proposed by [Emiya et al., 2011]. This software allows the computation of objective measures to assess the perceived quality of estimated source signals, based on perceptual similarity measures obtained with the PEMO-Q auditory model [Huber and Kollmeier, 2006]. For this approach, the estimated sources \( \hat{s}_j \) are decomposed in a similar manner to 4.1 as

\[
\hat{s}_j - s_j = s_{\text{target}} + \varepsilon_{\text{interf}} + \varepsilon_{\text{artif}} \quad (4.5)
\]

In here, the terms \( \varepsilon_{\text{interf}}, \varepsilon_{\text{noise}}, \varepsilon_{\text{artif}} \) the target distortion, the interference and the artifacts components, respectively.

In order to provide, in addition to the classic energy ratios, perceptually motivated results the following new performance metrics are introduced:
Overall Perceptual Score (OPS)

Target-related Perceptual Score (TPS)

Interference-related Perceptual Score (APS)

Artifacts-related Perceptual Score (APS)

Using the perceptual similarity measure (PSM) provided by the PEMO-Q auditory model, the scores are obtained by measuring the salience of each distortion component individually. To do so, the estimated sources are compared with themselves and the considered distortions are subtracted. The following salience features are obtained

\[ q_{overall}^j = \text{PSM}(\hat{s}_j, s_j) \]  \hspace{1cm} (4.6)

\[ q_{target}^j = \text{PSM}(\hat{s}_j, \hat{s}_j - s_{target}) \]  \hspace{1cm} (4.7)

\[ q_{interf}^j = \text{PSM}(\hat{s}_j, \hat{s}_j - e_{inter}) \]  \hspace{1cm} (4.8)

\[ q_{artif}^j = \text{PSM}(\hat{s}_j, \hat{s}_j - e_{art}) \]  \hspace{1cm} (4.9)

A nonlinear mapping is applied to these salience features, combine these features into a single scalar measure for each grading task and to adapt the feature scale to the subjective grading scale. In addition to the classic energy ratios (in dB), this method allows to provide the perceptual scores presented above (in %).

4.2 Description of the Dataset

In order to perform a Since the available datasets with score-aligned tracks or multitrack recordings are very few, [Fritsch, 2012] introduced a new dataset named TRIOS. The dataset is formed by five short extracts from chamber music trio pieces. Besides the mixtures, each separated instrumental track is provided. Since our interest is mainly to focus on bowed-string instruments, only two of the five recordings are used:

- a trio for violin, cello and piano by Franz Schubert (D.929, op.100)
- a trio for viola, clarinet and piano by Wolfgang Amadeus Mozart (K.498)

The database is accessible in the C4DM Research Data Repository. More details about the dataset can be found in [Fritsch, 2012].

\[ \text{http://c4dm.eecs.qmul.ac.uk/rdr/handle/123456789/27} \]
4.3 Evaluation Methodology and Results

This section addresses the evaluation of the different methods and configurations proposed in section 3. To do so, three main tests are performed. First, the IMM is compared with the two new proposed methods, the Multi-Excitation and Single Filter IMM and the Multi-Excitation and Multi-Filter IMM. Third, the model is more extensively evaluated with and without the extension of the proposal of the supervised timbre models. Finally, the influence over the pitch estimation when this supervised timbre models are used is evaluated.

4.3.1 Comparison of IMM and the Proposed Models

First, the Instantaneous Mixture Model presented by [Durrieu et al., 2009a] and the two proposed model extensions (see Section 3.2) are tested over the same data. Then, the obtained results are objectively evaluated with the above presented BSS Eval and PEASS metrics.

Description of the Data

To evaluate the performance of the different models, two 16 second audio excerpts of the F. Schubert and W. A. Mozart recordings from the TRIOS dataset are employed (see Section 4.3.1). The instrument to be separated are the violin and cello, for the Schubert recording, and the clarinet and viola, for the Mozart one.

Model Comparison Experimental Setup

Our proposed methods presented in Section 3.2 are applied to the audio excerpt. Then, the separation results are compared with those obtained from the original IMM method presented in Section 2.2.1. For this specific model comparison, the violin and the cello instruments, due to their high timbre similarity, are selected to be separated. Since the original goal is to mainly study similar timbre and bowed-string instruments, the piano is not that much of our interest and will be modeled by the accompaniment part of the model.

The pitch annotation of each track is carried out automatically using Yin [De Cheveigné and Kawahara, 2002], a monophonic pitch estimation method on the individual recordings. Afterwards, the output provided by the algorithm is manually corrected to ensure an optimal estimation.

For the experiment we use the default parameters that the IMM specifies. The spectrogram is calculated with a 2048 sample (46.4 ms) window and with a hop-size of 512 samples (11.6 ms) at a sampling rate of 44.1 KHz. The number of elements in the smooth filters dictionary $W_{\Gamma}$ is set to 30, the number of spectral
shapes for the filter part is set to 10, the number of elements for the accompaniment is 40 and the number of iterations 30.

It has to be mentioned that as two instruments are being separated, the original IMM has to be performed twice. First, providing the melody of the violin and second, providing the one of the cello. In contrast, since the proposed models are able to separate as many instrument as wanted, there is no need of performing the process twice.

Results

The results are objectively evaluated with the metrics specified in BSS Eval and PEASS toolkits.

First, the SDR, SIR and SAR scores for the Schubert audio excerpt are presented in Tables 4.1, 4.2 and 4.3, respectively.

![Table 4.1: SDR score results by model for the Schubert audio excerpt](image1)

![Table 4.2: SIR score results by model for the Schubert audio excerpt](image2)
Table 4.3: SAR score results by model for the Schubert audio excerpt

Second, the SDR, SIR and SAR scores for the Mozart audio excerpt are presented in Tables 4.1, 4.2 and 4.3, respectively.

Table 4.4: SDR score results by model for the Mozart audio excerpt

Table 4.5: SIR score results by model for the Mozart audio excerpt
Looking at the results provided by the BSS Eval toolbox, in average, both proposed models perform better than original the original IMM model. The average SDR score value is improved in 12 dB for the Schubert excerpt and in 6 dB for the Mozart excerpt.

When the instruments to be separated have a similar timbre, both of the proposed models show no big differences between each other, even if the Multi-Excitation and Multi-Filter IMM (MEF IMM) obtains better SDR results. It has to be mentioned that for the results obtained from the Mozart excerpt, the Multi-Excitation and Single Filter IMM performs worse than the MEF IMM. This may be cause by the fact that the timbre of the instruments is different and only one filter is modeling both of them.

The IMM model shows better SIR score values for the violin. However, in average the proposed MEF IMM improves the SIR score in 4 dB. Except for the case of the viola, in average, the SAR is also improved for both of the proposed methods.

In general, the violin performs better than the cello for the IMM method, while for both of the proposed models the results are similar. This decrease on the performance of the cello for the IMM may occur due to the possibility that the glottal model adapts better to the melody played by the violin than to the one of the cello.

So does the viola in respect to the clarinet, the IMM shows better results for the first one. For this Mozart excerpt, we can find differences between the proposed models. The Multi-Excitation and Multi-Filter IMM performs better in this case, which is a more flexible model than the MESF IMM.

In addition, the perceptual objective measures obtained with the PEASS toolkit for the Schubert audio excerpt are presented in Table 4.1 and Figure 4.7.
Figure 4.1: Overall Perceptual Score (OPS) results by model for the Schubert audio excerpt

![Graph showing overall perceptual scores by model for the Schubert audio excerpt.]

Table 4.7: PEASS toolkit results by model for the Schubert audio excerpt

<table>
<thead>
<tr>
<th>Model</th>
<th>IMM</th>
<th>MEsf IMM</th>
<th>MEF IMM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OPS</td>
<td>TPS</td>
<td>IPS</td>
</tr>
<tr>
<td>violin</td>
<td>21.46</td>
<td>9.34</td>
<td>53.59</td>
</tr>
<tr>
<td>cello</td>
<td>9.44</td>
<td>0.79</td>
<td>10.33</td>
</tr>
<tr>
<td>average</td>
<td>15.45</td>
<td>5.06</td>
<td>31.96</td>
</tr>
</tbody>
</table>

The perceptual objective measures obtained with the PEASS toolkit for the Mozart audio excerpt are exposed in Table 4.2 and Figure 4.8.

Figure 4.2: Overall Perceptual Score (OPS) results by model for the Mozart audio excerpt

![Graph showing overall perceptual scores by model for the Mozart audio excerpt.]

Table 4.8: PEASS toolkit results by model for the Mozart audio excerpt

<table>
<thead>
<tr>
<th>Model</th>
<th>IMM</th>
<th>MEsf IMM</th>
<th>MEF IMM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OPS</td>
<td>TPS</td>
<td>IPS</td>
</tr>
<tr>
<td>violin</td>
<td>28.19</td>
<td>32.77</td>
<td>47.74</td>
</tr>
<tr>
<td>cello</td>
<td>18.64</td>
<td>16.96</td>
<td>26.43</td>
</tr>
</tbody>
</table>

The perceptual objective measures obtained with the PEASS toolkit for the Mozart audio excerpt are exposed in Table 4.2 and Figure 4.8.
Table 4.8: PEASS toolkit results by model for the Mozart audio excerpt

As Figures 4.1 and 4.2 depict, the Overall Perceptual Score (OPS) obtained by the IMM is improved by both of the proposed MESF IMM and MEF IMM. As it can be seen from the results of both recordings, when the instruments have a more similar timbre (violin and cello, Schubert), the improvement over the OPS is not as high as when the instruments have a more different timbre (clarinet and viola, Mozart).

### 4.3.2 Multi-Excitation and Multi-Filter Model Parameter Optimization

During the many experiments that have been run through the design of the proposed method, it has been observed that the separation results may vary depending on the various parameters of the general decomposition process. The parameters that describe the decomposition, i.e. the number of bases of the filter dictionary, seem to be influential over the results.

Thus, before comparing the model with the supervised timbre models extension, a parameter optimization grid-search is performed. The goal of this search is no other than to obtain the best possible combination of parameters to represent the filter sub-model.

**Description of the Data**

The previously presented TRIOS dataset (see Section 4.3.1) is also used to evaluate this model and the timbre filters.

For the filter sub-model parameter optimization, the same 16 second extract employed in the previous model comparison test of the F. Schubert recording is used.

**Parameter Optimization Experimental Setup**

In order to obtain the best parameter configuration, a grid-search is performed over the audio excerpt. The violin and the cello instruments are the ones to be separated. The pitch is extracted over the recording as explained in Section 4.3.1.

<table>
<thead>
<tr>
<th>Model</th>
<th>IMM</th>
<th></th>
<th></th>
<th></th>
<th>MESF IMM</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>MEF IMM</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OPS</td>
<td>TPS</td>
<td>IPS</td>
<td>APS</td>
<td>OPS</td>
<td>TPS</td>
<td>IPS</td>
<td>APS</td>
<td>OPS</td>
<td>TPS</td>
<td>IPS</td>
<td>APS</td>
</tr>
<tr>
<td>clarinet</td>
<td>10.18</td>
<td>0.48</td>
<td>44.18</td>
<td>3.48</td>
<td>34.58</td>
<td>47.81</td>
<td>42.74</td>
<td>47.85</td>
<td>33.25</td>
<td>49.19</td>
<td>42.02</td>
<td>46.95</td>
</tr>
<tr>
<td>viola</td>
<td>10.55</td>
<td>0.82</td>
<td>9.30</td>
<td>13.68</td>
<td>12.25</td>
<td>1.82</td>
<td>7.90</td>
<td>16.74</td>
<td>11.42</td>
<td>1.32</td>
<td>2.40</td>
<td>11.86</td>
</tr>
<tr>
<td>average</td>
<td>10.37</td>
<td>0.65</td>
<td>26.74</td>
<td>8.58</td>
<td>23.42</td>
<td>24.81</td>
<td>25.32</td>
<td>32.29</td>
<td>22.33</td>
<td>25.25</td>
<td>22.21</td>
<td>29.40</td>
</tr>
</tbody>
</table>
The piano is not modeled and is approximated by the accompaniment. The goal of this test is to find how the different parameters values are influential for the separation and which ones are the more convenient for our scenario. The parameters and values tested are the following:

- \( P \) smooth elementary filters of the \( W_{\Gamma} \) dictionary: 10, 30, 70, 90, 100 and 120.
- \( K \) filters to be considered (bases): 1, 2, 3, 10, 20 and 30.
- Iterations: 10, 30, 50, 70 and 90.

Results

The results are evaluated with the objective metrics specified in BSS Eval toolbox. The SDR score is the measure that has been taken as the main reference to judge the best results.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Violin</th>
<th>Cello</th>
<th>Accompaniment</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>5.5673</td>
<td>6.2863</td>
<td>7.6324</td>
<td>6.4953</td>
</tr>
<tr>
<td>30</td>
<td>7.3023</td>
<td>7.3423</td>
<td>8.6341</td>
<td>7.7596</td>
</tr>
<tr>
<td>50</td>
<td>6.4457</td>
<td>5.7395</td>
<td>6.8046</td>
<td>6.3299</td>
</tr>
<tr>
<td>70</td>
<td>7.1337</td>
<td>5.0415</td>
<td>6.3999</td>
<td>6.1917</td>
</tr>
<tr>
<td>90</td>
<td>7.1651</td>
<td>5.7294</td>
<td>6.0596</td>
<td>6.3180</td>
</tr>
</tbody>
</table>

Table 4.9: SDR score results by iteration over 90 \( P \) smooth filters and 20 \( K \) bases
As shown in Tables 4.9, 4.10, 4.11 and 4.12 the best SDR results are obtained for 90 $P$ smooth elementary filters of the $W_Γ$ dictionary, 20 $K$ number of filters and 30 iterations. When running the estimation over more iterations, the estimation error keeps decreasing. However, since the SDR decreases, there must be leakage from one source to another, having a worse separation quality.

The complete table of results can be found in Appendix A.
4.3.3 Multi-Excitation and Multi-Filter Model with Supervised Timbre Model Extension

The second test consist of comparing the performance of the proposed Multi-Excitation and Multi-Filter model when the filters are estimated over the mixture and the performance when the supervised timbre models are loaded to the filters (see Sections 3.3 and 3.2.2, respectively).

Description of the Data

The previously presented TRIOS dataset is also used to evaluate this model and the timbre filters (see Section 4.3.1).

For the comparison between the proposed Multi-Excitation and Multi-Filter model and the extension with the supervised timbre models, both F. Schubert and W. A. Mozart recordings are used. The entire duration of the recordings is now considered for this test.

For the different pitch estimations test, two new mixes of instruments are created. The violin and the cello from F. Schubert recording are combined in a new mixture. The clarinet and viola from the W. A. Mozart recording are joined in a new mixture too. The piano is not included in any of the new mixtures.

Supervised Timbre Model Performance Experimental Setup

The performance of the Supervised Timbre Models is evaluated in three different experiments.

- Performance of the Timbre Models:

The first part of this test consist of comparing the performance of the proposed Multi-Excitation and Multi-Filter model when the filters are estimated over the mixture and when the supervised timbre models are loaded into the model. An extra case where the filters are unsupervised, this is, randomly initialized, and not updated is also introduced. This case will theoretically provide a bottom line or worst-case scenario for the filter sub-model.

In order not to train and test the filters with exactly the same data a N-Fold Cross Validation is performed. The number of folds is determined by the duration of the recordings and the folds themselves. In other words, it is a compromise between not letting the amount of information provided to the system be an influence over the performance of the model, and having enough different folds to train the timbre models and test it over the recording. Thus, each fold is selected to have a minimum amount of data of approximately 8.5 seconds. The influence of this parameter, which is referred in this work as block size, is tested afterwards.
For the above mentioned F. Schubert recording from the TRIOS dataset, the audio is divided into 6 folds. Each of the test audio extracts has an approximate duration of 8.5 seconds. Each of the training audio extracts, in contrary, an approximate of 45 seconds.

For the W. A. Mozart recording, since the length of the recording is shorter than the former, the audio is divided into 4 folds. Each of the test audio extracts has an approximate duration of 8.5 seconds. Each of the training audio extracts, in contrary, an approximate of 26 seconds.

The configuration for this test is obtained from the results of the parameter optimization test explained in Section 4.3.2. This is, the number of $P$ smooth elementary filters of the $W_T$ dictionary is set to 90, the $K$ number of filters to be considered is 20 and the number of iterations is 30.

Results

The results are evaluated with the objective metrics specified in BSS Eval toolbox.

<table>
<thead>
<tr>
<th>Model</th>
<th>MEF IMM</th>
<th>MEF IMM + SV FILT.</th>
<th>MEF IMM + USV FILT.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
</tr>
<tr>
<td>violin</td>
<td>8.06</td>
<td>16.53</td>
<td>9.29</td>
</tr>
<tr>
<td>cello</td>
<td>6.05</td>
<td>12.18</td>
<td>8.29</td>
</tr>
<tr>
<td>accomp.</td>
<td>6.54</td>
<td>9.05</td>
<td>11.35</td>
</tr>
<tr>
<td>average</td>
<td>6.89</td>
<td>12.58</td>
<td>9.64</td>
</tr>
</tbody>
</table>

Table 4.13: BSSEval results of the Schubert recording for the Multi-Excitation and Multi-Filter Model with estimated filters (MEF IMM), supervised filters (MEF IMM + SV FILT.) and unsupervised filters (MEF IMM + USV FILT.)

<table>
<thead>
<tr>
<th>Model</th>
<th>MEF IMM</th>
<th>MEF IMM + SV FILT.</th>
<th>MEF IMM + USV FILT.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
</tr>
<tr>
<td>viola</td>
<td>-4.12</td>
<td>0.79</td>
<td>5.04</td>
</tr>
<tr>
<td>accomp.</td>
<td>6.81</td>
<td>12.24</td>
<td>8.70</td>
</tr>
<tr>
<td>average</td>
<td>5.11</td>
<td>10.33</td>
<td>9.33</td>
</tr>
</tbody>
</table>

Table 4.14: BSSEval results of the Mozart recording for the Multi-Excitation and Multi-Filter Model with estimated filters (MEF IMM), supervised filters (MEF IMM + SV FILT.) and unsupervised filters (MEF IMM + USV FILT.)

As Table 4.13 and Table 4.14 depict, the MEF IMM performs better than the MEF IMM with the loading of the supervised timbre models case, for both Schubert and Mozart recordings. This can be a consequence of the timbre of the different instruments being properly approximated directly from the mixture audio. It is
also possible that the supervised filters are not designed in the most optimal way, an aspect it is later discussed. The results obtained for the Multi-Excitation and Multi-Filter Model with loaded unsupervised filters case show the worst possible case values.

- Influence of the Supervised Filter Model Extension over the amount of data and iterations:

The second part of this test consists of inspecting the influence of the above mentioned block size and the number of iterations in the separation process. This both factors have a direct influence on the latency of the system. The main goal is to compare the differences between the proposed Multi-Excitation and Multi-Filter model, with and without the use of supervised timbre models, when the amount of data provided is reduced or the number of iterations are reduced.

The following block size and iteration values are evaluated for the two cases. First, when the Multi-Excitation and Multi-Filter model filters are estimated over the mixture. Second, when the supervised timbre models are loaded into the model.

- Block size (in frames): 10, 20, 50, 100, 150, 200, 300, 400, 500, 600 and 700.
- Iterations: 3, 5, 7, 10, 15, 20, 25, 30, 50, 70, 90, 100 and 120.

The frame value that specifies the block size is equivalent to 11 ms, as the used hop size is 512 samples and the sampling rate 44.1 KHz. When the amount of data provided is the value to inspect, the number of iterations is set to 30. For the case when the number of iterations is the value to inspect, all the data of each fold is provided to the system.

Results

The results are evaluated with the objective metrics specified in BSS Eval toolbox. In this section only the SDR score values are presented. The SIR and SAR score values of both recordings can be found in Appendix B.

For the different iteration values, the MEF IMM obtains better SDR values than the MEF IMM with the loading of the supervised timbre models, for both Schubert and Mozart recordings (see Table 4.15 and Table 4.16). The use of the filters seems not to improve the separation with lower iteration values. The best average result is obtained with the MEF IMM and 50 iterations.
Table 4.15: SDR score of the Schubert recording for the Multi-Excitation and Multi-Filter Model with estimated filters and supervised filters for different iteration values

In the case of the amount of data (block size) provided to the system, the use of supervised timbre models over the MEF IMM shows no significant SDR score difference over the tested values. However, the average maximum value for the MEF IMM with loaded timbre models is obtained 100 frames earlier (at 400 frames) than the average maximum value for the MEF IMM with estimated filters (at 500 frames).
Table 4.16: SDR score of the Mozart recording for the Multi-Excitation and Multi-Filter Model with estimated filters and supervised filters for different iteration values
Table 4.17: SDR score of the Schubert recording for the Multi-Excitation and Multi-Filter Model with estimated filters and supervised filters for different block size values
Table 4.18: SDR score of the Mozart recording for the Multi-Excitation and Multi-Filter Model with estimated filters and supervised filters for different block size values.

<table>
<thead>
<tr>
<th>Block Size</th>
<th>Clarinet</th>
<th>Clarinet Fit</th>
<th>Viola</th>
<th>Viola Fit</th>
<th>Accompaniment</th>
<th>Accompaniment Fit</th>
<th>Average</th>
<th>Average Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block20</td>
<td>0.3688</td>
<td>-0.0004</td>
<td>-10.4991</td>
<td>-10.4959</td>
<td>-3.1623</td>
<td>-3.2640</td>
<td>-4.4976</td>
<td>-4.1868</td>
</tr>
<tr>
<td>Block50</td>
<td>3.2421</td>
<td>3.4713</td>
<td>-8.1680</td>
<td>-8.1020</td>
<td>-0.9507</td>
<td>-0.3001</td>
<td>-2.0989</td>
<td>-1.9709</td>
</tr>
<tr>
<td>Block100</td>
<td>10.3451</td>
<td>10.2934</td>
<td>-5.7717</td>
<td>-5.8424</td>
<td>2.7902</td>
<td>2.5845</td>
<td>2.4545</td>
<td>2.3632</td>
</tr>
<tr>
<td>Block400</td>
<td>12.2273</td>
<td>12.2193</td>
<td>-3.6992</td>
<td>-3.3508</td>
<td>5.4105</td>
<td>5.4238</td>
<td>4.7564</td>
<td>4.8076</td>
</tr>
<tr>
<td>Block500</td>
<td>12.5450</td>
<td>12.7960</td>
<td>-2.6992</td>
<td>-1.3000</td>
<td>6.3192</td>
<td>6.3424</td>
<td>4.7283</td>
<td>4.9760</td>
</tr>
</tbody>
</table>

- **SDR (dB)**
- **Block Size Values**
- Influence of the Supervised Filter Model Extension for Blind Pitch Estimation:

The third and last test of this section tries to measure the influence of the supervised filter models over a blind pitch estimation process. Until now, the pitch information of each instrument to be separated has been provided for all the experiments. In this case, the goal is to compare how the timbre models affect the pitch estimation process over the mixture. This is, the pitch is not provided to the system but it is estimated over the mixture. Two cases of F0 estimation are inspected: when the timbre models are loaded and when they are not (when filters are estimated).

The parameter configuration for this test is obtained from the results of the parameter optimization test (see Section 4.3.2).

Results

As the goal is not to compare the separation performance but the pitch estimation, the results can not be evaluated with the objective metrics. Instead, the different estimated F0 matrices are compared.

For the MEF IMM with estimated filters, in order to know which pitch corresponds to each instrument the estimations have to be compared to the annotation reference. As the model performs a free separation, the output of the instruments is in principle unknown. In our case, since the amount of instruments is reduced an a reference of each instruments pitch is available, we are able to label each output.

Tables 4.3, 4.4, 4.5 and 4.6 depict that the MEF IMM when the supervised timbre models are loaded performs a better pitch estimation than the MEF IMM when the filters are estimated. The melody lines appear more defined, with less missing segments and errors, especially for the clarinet excerpt. This may be caused by the fact that the harmonic series of the clarinet has missing partials, this is, because of its cylindrical resonator, only the odd-numbered harmonics are present. This characteristic may be easily modeled when the timbre model is supervised and not estimated from the mixture.
Figure 4.3: Pitch estimations of a 8 second violin excerpt (Schubert recording). Estimations with Yin (top), MEF IMM (middle) and MEF IMM loading supervised timbre models (bottom).

Figure 4.4: Pitch estimations over a 8 second cello excerpt (Schubert recording). Estimations with Yin (top), MEF IMM (middle) and MEF IMM loading supervised timbre models (bottom).
Figure 4.5: Pitch estimations over a 8 second viola excerpt (Mozart recording). Estimations with Yin (top), MEF IMM (middle) and MEF IMM loading supervised timbre models (bottom).

Figure 4.6: Pitch estimations over a 8 second clarinet excerpt (Mozart recording). Estimations with Yin (top), MEF IMM (middle) and MEF IMM loading supervised timbre models (bottom).
Chapter 5

CONCLUSIONS AND FUTURE WORK

The present work in this master thesis has focused on the study, combination and improvement of different source separation models. The main motivation was to find whether the use of instrument dependent source and filter model representations and the use of trained timbre models could help to improve the separation process. An existing source separation model, the IMM, has been used as a starting point. Overcoming the limitations of the model, modifications over the source / filter representation have been introduced. More precisely, for the filter sub-model, the use of trained timbre models has been studied. The different results show that the proposed extension of the source / filter model can improve the results of the separation and that the followed methodology is promising, with many possibilities for further research and potential applications.

5.1 Contributions

The following list contains the main contributions of this work:

- The literature review done in the related areas of research have determined that there is a large amount of research done in source separation and his different methods and strategies, but the application and performance of this methods is still novel for instruments of similar timbre.

- It has been demonstrated how the proposed extensions of modeling the data using multiple sources and filters allow to improve the quality of the results.

- The relative improvement of the performance of the proposed model over the IMM model is between 6 dB and 12 dB of SDR, depending on the content of the audio mixture.
• The use of individual instrument data to obtain Trained Timbre Models, which allows to improve the quality of the pitch estimation.

5.2 Conclusions

Considering the motivations and goals presented in the introduction and having into account the above presented contributions, some conclusions can be obtained.

In this thesis, an efficient method for score-informed source separation has been presented. It specifically uses a source / filter decomposition in a NMF framework, where each one of the instruments present in the mixture is modeled. This proposed Multi-Excitation and Filter IMM method produces acceptable audio results, and appears to give better performance results than the compared method from the literature. This improvement is obtained thanks to the ability of the method to estimate the representation of each of the instruments with its own source and filter sub-models. Its strength is that the informed pitch estimation provides a good reference of the harmonic parts of each instrument over the decomposition process. However, a weak point is that all the residual or inharmonic parts are not modeled and are fitted into the estimated accompaniment track.

Depending on the application of the method, this fact can be an inconvenient to a certain extent. If the goal is to obtain a total separation of the sources with listening purposes, i.e. ‘de-soloing’, the separation of only the harmonic part of the instrument may not be acceptable. In contrast, if the objective is to apply the separated sources for remixing, i.e., equalizing or changing the dynamics, the method could be totally valid, since the in-harmonic parts have not such a big effect in this case.

Regarding the use of trained timbre models, the results show no significant improvement over the estimation of the timbre directly from the audio mixture. The proposed Multi-Excitation and Filter IMM seems to be able to efficiently estimate the filter activations directly from the mixture. Further research has to be performed in the design of the filter to properly adapt it to the specific timbre characteristics of the different instruments. However, leaving apart the pitch-informed separation, the use of this trained timbre models appears to provide a better pitch estimation than the one obtained without any timbre information of the instruments. This may show us a new path to follow in further research.
5.3 Future work

Some positive results have been obtained with the proposed models, but the different steps over the process have shown that there is still a lot of room for improvement. In order to obtain more accurate results and a higher quality separation there are many different aspects that can be worked on.

5.3.1 Source / Filter Model Improvements

As it has been mentioned in the conclusions, the proposed model only takes into account the harmonic part of the sources for the decomposition. The extension of the source / filter model with a transient detection, i.e. a non-harmonic signal modeling, will potentially improve the results of the separation. Adding some non-harmonic bases to the NMF decomposition framework is definitely an aspect to work on.

In addition, the proposed model employs a very simple approach for the modeling of the source, where the bases are defined by a flat harmonic-comb excitation. Adapting this source excitation bases to each of the different instrument excitations is also a facet to improve. For example, the introduction of different excitation slopes over the harmonic-comb could be a simple and effective start point.

5.3.2 Trained Filter Models Improvements

The timbre models on the proposed extension are trained by learning the activations over some smooth elementary filters (Hann functions) that represent timbral characteristics of the different instruments. The use of this representation appears to be valid for the estimation of this characteristics over the mixture, but may be improved with the use of other elementary filters. Instead of using Hann filters, other function combinations should be tested to verify the best possible timbre modeling. For example, the use of wide Gaussian functions to capture the energy zones and narrower ones, with a sparsity constrain, to define the resonances of the spectrum of the instruments can be an option to inspect.

In accordance with the modeling of the non-harmonic parts, inspecting the residual obtained from the actual timbre model training could provide a good reference to know which requirements are needed to model this transitions.

As the use of this trained timbre models over the pitch estimation depicts a promising result, the use of this information for multi-pitch estimation over the mixture audio should be a field to keep on working. The use of pitch-informed source separation is a good alternative to avoid the problems that state of the art
multi-pitch estimation algorithms entail, but prior information of the sources is not always available. In this sense, with only having the information of the instruments present in the mixture, another supervised source separation methodology, where timbre information of the instruments is provided, could be followed.
Appendix A

APPENDIX A
Table A.1: BSS Eval results for the Schubert excerpt over 10 iterations and different values for $P$ smooth filters and $K$ bases.

<table>
<thead>
<tr>
<th></th>
<th>K1</th>
<th>SDR</th>
<th>SIR</th>
<th>SAR</th>
<th>K2</th>
<th>SDR</th>
<th>SIR</th>
<th>SAR</th>
<th>K3</th>
<th>SDR</th>
<th>SIR</th>
<th>SAR</th>
<th>K10</th>
<th>SDR</th>
<th>SIR</th>
<th>SAR</th>
<th>K20</th>
<th>SDR</th>
<th>SIR</th>
<th>SAR</th>
<th>K30</th>
<th>SDR</th>
<th>SIR</th>
<th>SAR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P10</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>P30</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>P70</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>P90</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BSS Eval results for the Schubert excerpt of the parameter optimization test explained in 4.3.2.
<table>
<thead>
<tr>
<th></th>
<th>K1</th>
<th>K2</th>
<th>K3</th>
<th>K10</th>
<th>K20</th>
<th>K30</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
<td>SDR</td>
</tr>
</tbody>
</table>

Table A.2: BSS Eval results for the Schubert excerpt over 30 iterations and different values for P smooth filters and K bases.
<table>
<thead>
<tr>
<th></th>
<th>K1</th>
<th>K2</th>
<th>K3</th>
<th>K10</th>
<th>K20</th>
<th>K30</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDR SAR</td>
<td>SIR SAR</td>
<td>SIR SAR</td>
<td>SDR SAR</td>
<td>SIR SAR</td>
<td>SIR SAR</td>
</tr>
<tr>
<td>P10</td>
<td>5.5027</td>
<td>13.652</td>
<td>5.8515</td>
<td>5.4336</td>
<td>13.129</td>
<td>6.4491</td>
</tr>
<tr>
<td></td>
<td>5.5111</td>
<td>12.46</td>
<td>6.7301</td>
<td>5.3689</td>
<td>15.95</td>
<td>5.8754</td>
</tr>
<tr>
<td></td>
<td>5.3689</td>
<td>15.95</td>
<td>5.8754</td>
<td>5.4876</td>
<td>15.79</td>
<td>6.0258</td>
</tr>
<tr>
<td></td>
<td>5.4876</td>
<td>15.79</td>
<td>6.0258</td>
<td>5.4684</td>
<td>14.28</td>
<td>6.2394</td>
</tr>
<tr>
<td></td>
<td>5.4684</td>
<td>14.28</td>
<td>6.2394</td>
<td>5.7016</td>
<td>15.83</td>
<td>6.2567</td>
</tr>
<tr>
<td></td>
<td>5.7016</td>
<td>15.83</td>
<td>6.2567</td>
<td>6.2567</td>
<td>15.83</td>
<td>6.2567</td>
</tr>
<tr>
<td></td>
<td>5.3689</td>
<td>15.95</td>
<td>5.8754</td>
<td>5.4876</td>
<td>15.79</td>
<td>6.0258</td>
</tr>
<tr>
<td></td>
<td>5.4876</td>
<td>15.79</td>
<td>6.0258</td>
<td>5.4684</td>
<td>14.28</td>
<td>6.2394</td>
</tr>
<tr>
<td></td>
<td>5.4684</td>
<td>14.28</td>
<td>6.2394</td>
<td>5.7016</td>
<td>15.83</td>
<td>6.2567</td>
</tr>
<tr>
<td></td>
<td>5.7016</td>
<td>15.83</td>
<td>6.2567</td>
<td>6.2567</td>
<td>15.83</td>
<td>6.2567</td>
</tr>
<tr>
<td>P30</td>
<td>5.5027</td>
<td>13.652</td>
<td>5.8515</td>
<td>5.4336</td>
<td>13.129</td>
<td>6.4491</td>
</tr>
<tr>
<td></td>
<td>5.5111</td>
<td>12.46</td>
<td>6.7301</td>
<td>5.3689</td>
<td>15.95</td>
<td>5.8754</td>
</tr>
<tr>
<td></td>
<td>5.3689</td>
<td>15.95</td>
<td>5.8754</td>
<td>5.4876</td>
<td>15.79</td>
<td>6.0258</td>
</tr>
<tr>
<td></td>
<td>5.4876</td>
<td>15.79</td>
<td>6.0258</td>
<td>5.4684</td>
<td>14.28</td>
<td>6.2394</td>
</tr>
<tr>
<td></td>
<td>5.4684</td>
<td>14.28</td>
<td>6.2394</td>
<td>5.7016</td>
<td>15.83</td>
<td>6.2567</td>
</tr>
<tr>
<td></td>
<td>5.7016</td>
<td>15.83</td>
<td>6.2567</td>
<td>6.2567</td>
<td>15.83</td>
<td>6.2567</td>
</tr>
<tr>
<td>P70</td>
<td>5.5027</td>
<td>13.652</td>
<td>5.8515</td>
<td>5.4336</td>
<td>13.129</td>
<td>6.4491</td>
</tr>
<tr>
<td></td>
<td>5.5111</td>
<td>12.46</td>
<td>6.7301</td>
<td>5.3689</td>
<td>15.95</td>
<td>5.8754</td>
</tr>
<tr>
<td></td>
<td>5.3689</td>
<td>15.95</td>
<td>5.8754</td>
<td>5.4876</td>
<td>15.79</td>
<td>6.0258</td>
</tr>
<tr>
<td></td>
<td>5.4876</td>
<td>15.79</td>
<td>6.0258</td>
<td>5.4684</td>
<td>14.28</td>
<td>6.2394</td>
</tr>
<tr>
<td></td>
<td>5.4684</td>
<td>14.28</td>
<td>6.2394</td>
<td>5.7016</td>
<td>15.83</td>
<td>6.2567</td>
</tr>
<tr>
<td></td>
<td>5.7016</td>
<td>15.83</td>
<td>6.2567</td>
<td>6.2567</td>
<td>15.83</td>
<td>6.2567</td>
</tr>
<tr>
<td>P90</td>
<td>5.5027</td>
<td>13.652</td>
<td>5.8515</td>
<td>5.4336</td>
<td>13.129</td>
<td>6.4491</td>
</tr>
<tr>
<td></td>
<td>5.5111</td>
<td>12.46</td>
<td>6.7301</td>
<td>5.3689</td>
<td>15.95</td>
<td>5.8754</td>
</tr>
<tr>
<td></td>
<td>5.3689</td>
<td>15.95</td>
<td>5.8754</td>
<td>5.4876</td>
<td>15.79</td>
<td>6.0258</td>
</tr>
<tr>
<td></td>
<td>5.4876</td>
<td>15.79</td>
<td>6.0258</td>
<td>5.4684</td>
<td>14.28</td>
<td>6.2394</td>
</tr>
<tr>
<td></td>
<td>5.4684</td>
<td>14.28</td>
<td>6.2394</td>
<td>5.7016</td>
<td>15.83</td>
<td>6.2567</td>
</tr>
<tr>
<td></td>
<td>5.7016</td>
<td>15.83</td>
<td>6.2567</td>
<td>6.2567</td>
<td>15.83</td>
<td>6.2567</td>
</tr>
<tr>
<td>P100</td>
<td>5.5027</td>
<td>13.652</td>
<td>5.8515</td>
<td>5.4336</td>
<td>13.129</td>
<td>6.4491</td>
</tr>
<tr>
<td></td>
<td>5.5111</td>
<td>12.46</td>
<td>6.7301</td>
<td>5.3689</td>
<td>15.95</td>
<td>5.8754</td>
</tr>
<tr>
<td></td>
<td>5.3689</td>
<td>15.95</td>
<td>5.8754</td>
<td>5.4876</td>
<td>15.79</td>
<td>6.0258</td>
</tr>
<tr>
<td></td>
<td>5.4876</td>
<td>15.79</td>
<td>6.0258</td>
<td>5.4684</td>
<td>14.28</td>
<td>6.2394</td>
</tr>
<tr>
<td></td>
<td>5.4684</td>
<td>14.28</td>
<td>6.2394</td>
<td>5.7016</td>
<td>15.83</td>
<td>6.2567</td>
</tr>
<tr>
<td></td>
<td>5.7016</td>
<td>15.83</td>
<td>6.2567</td>
<td>6.2567</td>
<td>15.83</td>
<td>6.2567</td>
</tr>
<tr>
<td>P120</td>
<td>5.5027</td>
<td>13.652</td>
<td>5.8515</td>
<td>5.4336</td>
<td>13.129</td>
<td>6.4491</td>
</tr>
<tr>
<td></td>
<td>5.5111</td>
<td>12.46</td>
<td>6.7301</td>
<td>5.3689</td>
<td>15.95</td>
<td>5.8754</td>
</tr>
<tr>
<td></td>
<td>5.3689</td>
<td>15.95</td>
<td>5.8754</td>
<td>5.4876</td>
<td>15.79</td>
<td>6.0258</td>
</tr>
<tr>
<td></td>
<td>5.4876</td>
<td>15.79</td>
<td>6.0258</td>
<td>5.4684</td>
<td>14.28</td>
<td>6.2394</td>
</tr>
<tr>
<td></td>
<td>5.4684</td>
<td>14.28</td>
<td>6.2394</td>
<td>5.7016</td>
<td>15.83</td>
<td>6.2567</td>
</tr>
<tr>
<td></td>
<td>5.7016</td>
<td>15.83</td>
<td>6.2567</td>
<td>6.2567</td>
<td>15.83</td>
<td>6.2567</td>
</tr>
</tbody>
</table>

Table A.3: BSS Eval results for the Schubert excerpt over 50 iterations and different values for P smooth filters and K bases
<table>
<thead>
<tr>
<th>K1</th>
<th>K2</th>
<th>K3</th>
<th>K10</th>
<th>K20</th>
<th>K30</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
</tr>
<tr>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
</tr>
<tr>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
</tr>
</tbody>
</table>

**P10**

| 5.1135 | 12.739 | 6.1625 | 5.0594 | 12.218 | 6.2404 |
| 4.8886 | 17.415 | 5.2165 | 5.1803 | 16.456 | 5.6138 |
| 3.947 | 4.7123 | 13.128 | 4.6727 | 5.8315 | 11.982 |
| 4.6437 | 11.8221 | 8.1690 | 5.0600 | 12.0722 | 8.0015 |
| 4.6782 | 12.828 | 5.6208 | 5.0264 | 12.68 | 6.0724 |
| 4.9266 | 6.1602 | 11.936 | 5.3154 | 6.6356 | 11.93 |
| 5.1701 | 11.8891 | 7.9780 | 5.3435 | 12.1069 | 8.2448 |
| 5.1802 | 12.511 | 6.9052 | 5.5307 | 12.67 | 6.6918 |
| 5.364 | 16.061 | 5.8569 | 6.1763 | 17.908 | 6.5476 |
| 5.0471 | 6.2804 | 12.038 | 5.3171 | 6.5213 | 12.35 |
| 5.1971 | 11.6175 | 8.0654 | 5.6747 | 12.3664 | 8.5299 |

**P90**

| 5.5358 | 13.924 | 6.3882 | 5.5715 | 12.533 | 7.8385 |
| 6.4572 | 17.348 | 6.9054 | 5.6627 | 17.254 | 6.056 |
| 5.6977 | 12.5058 | 8.5559 | 5.4023 | 11.9704 | 8.3630 |
| 5.2378 | 12.669 | 6.3321 | 5.6722 | 12.798 | 6.803 |
| 5.8148 | 16.062 | 6.352 | 5.8158 | 17.259 | 6.2199 |
| 5.6096 | 7.0994 | 11.753 | 5.4342 | 6.8917 | 11.739 |
| 5.5541 | 11.9435 | 8.1457 | 5.6407 | 12.3105 | 8.2631 |

**P100**

| 5.0551 | 12.311 | 6.208 | 6.0025 | 12.971 | 7.1907 |
| 5.1128 | 16.671 | 5.5198 | 5.662 | 17.942 | 5.966 |
| 5.5297 | 17.625 | 5.8008 | 6.2168 | 16.508 | 6.7393 |
| 4.7596 | 5.9837 | 11.833 | 4.7605 | 5.7407 | 12.733 |
| 4.9758 | 11.6552 | 7.8536 | 5.4750 | 12.2179 | 8.6399 |

**P120**

| 5.4156 | 12.2633 | 8.3217 | 6.3735 | 12.8814 | 8.7250 |

Table A.4: BSS Eval results for the Schubert excerpt over 70 iterations and different values for $P$ smooth filters and $K$ bases.
<table>
<thead>
<tr>
<th></th>
<th>K1</th>
<th>K2</th>
<th>K3</th>
<th>K10</th>
<th>K20</th>
<th>K30</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
</tr>
<tr>
<td>P10</td>
<td>5,1413</td>
<td>13,47</td>
<td>6,0266</td>
<td>5,3349</td>
<td>12,42</td>
<td>6,5224</td>
</tr>
<tr>
<td>P30</td>
<td>4,9797</td>
<td>12,659</td>
<td>6,0216</td>
<td>4,9431</td>
<td>12,515</td>
<td>6,0148</td>
</tr>
<tr>
<td>P70</td>
<td>4,8659</td>
<td>12,642</td>
<td>5,889</td>
<td>5,1185</td>
<td>12,561</td>
<td>6,7089</td>
</tr>
<tr>
<td>P90</td>
<td>5,3099</td>
<td>12,918</td>
<td>6,3536</td>
<td>4,4933</td>
<td>11,742</td>
<td>5,6813</td>
</tr>
<tr>
<td>P100</td>
<td>5,3691</td>
<td>12,0187</td>
<td>8,1030</td>
<td>5,0078</td>
<td>11,8592</td>
<td>7,9894</td>
</tr>
<tr>
<td>P120</td>
<td>5,0880</td>
<td>11,8548</td>
<td>7,7733</td>
<td>5,4787</td>
<td>12,3198</td>
<td>8,3097</td>
</tr>
</tbody>
</table>

Table A.5: BSS Eval results for the Schubert excerpt over 90 iterations and different values for P smooth filters and K bases
SIR and SAR error values of Schubert and Mozart recordings over different iteration values corresponding to the test presented in 4.3.3.

Table B.1: SIR error of the Schubert recording for the Multi-Excitation and Multi-Filter Model with estimated filters and trained filters for different iteration values

<table>
<thead>
<tr>
<th>Iter</th>
<th>3iter</th>
<th>5iter</th>
<th>7iter</th>
<th>10iter</th>
<th>15iter</th>
<th>20iter</th>
<th>25iter</th>
<th>30iter</th>
<th>50iter</th>
<th>70iter</th>
<th>90iter</th>
</tr>
</thead>
</table>
Table B.2: SIR error of the Mozart recording for the Multi-Excitation and Multi-Filter Model with estimated filters and trained filters for different iteration values
Table B.3: SAR error of the Schubert recording for the Multi-Excitation and Multi-Filter Model with estimated filters, trained filters for different iteration values
Table B.4: SAR error of the Mozart recording for the Multi-Excitation and Multi-Filter Model with estimated filters and trained filters for different iteration values
SIR and SAR error values of Schubert and Mozart recordings over different block size values corresponding to the test presented in 4.3.3.

Table B.5: SIR error of the Schubert recording for the Multi-Excitation and Multi-Filter Model with estimated filters and trained filters for different block size values
Table B.6: SIR error of the Mozart recording for the Multi-Excitation and Multi-Filter Model with estimated filters and trained filters for different block size values
Table B.7: SAR error of the Schubert recording for the Multi-Excitation and Multi-Filter Model with estimated filters and trained filters for different block size values
Table B.8: SAR error of the Mozart recording for the Multi-Excitation and Multi-Filter Model with estimated filters and trained filters for different block size values.
Bibliography


