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

UNIVERSITAT POMPEU FABRA

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

List of Figures

List of Tables

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 [Burred, 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 ore 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

$$x(n) = \sum_{m=1}^{M} s_m(n) \tag{2.1}$$

where s_m is the mth 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

$$X_{i\mu} \approx (WH)_{i\mu} = \sum_{i=1}^{r} W_{ia} H_{a\mu} \quad \text{s.t.} \quad W, H \geq 0 \tag{2.2}$$

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 matrixes 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

$$\min \mathcal{D}(X|W, H) \quad W, H \geq 0, \tag{2.3}$$

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

$$\mathcal{D}(X|W, H) = \sum_{f=1}^{N} \sum_{n=1}^{M} d([X]_{fn}|[WH]_{fn}), \tag{2.4}$$

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

$$d_{EUC}(x|y) = \frac{1}{2}(x - y)^2 \tag{2.5}$$

and the Kullback-Leibler (KL) divergence

$$d_{KL}(x|y) = x \log \frac{x}{y} - x + y \tag{2.6}$$

[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

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

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.

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

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

where s_m is the mth 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

$$X = V + M \tag{2.9}$$

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:

$$\begin{cases} X_{R,fn} = \alpha_R V_{R,fn} + \sum_{j=1}^{J} \beta_{Rj} M_{Rj,fn} \\ X_{L,fn} = \alpha_R V_{L,fn} + \sum_{j=1}^{J} \beta_{Lj} M_{Lj,fn} \end{cases} \tag{2.10}$$

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

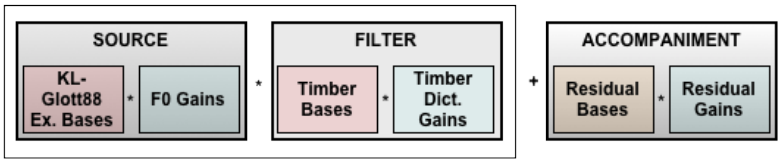


Figure 2.1: Source / filter + accompaniment diagram of the IMM

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:

$$S_{V,fn} = S_{\phi,fn} S_{F_0,fn} \tag{2.11}$$

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

$$S_{F_0} = W_{F_0} H_{F_0} \tag{2.12}$$

For the filter, a dictionary W_ϕ and its activation coefficients H_ϕ 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_Γ and its activations H_Γ.

$$S_\phi = W_\phi H_\phi = W_\Gamma H_\Gamma H_\phi \tag{2.13}$$

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

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

where "$\cdot$" 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.

$$S_{M,j,fn} = w_{fj} h_{jn} \tag{2.15}$$

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

$$S_{M_C,fn} = \sum_{j=1}^{J} \beta_{Cj}^2 S_{M,j,fn} = [W_M B_C H_M]_{fn}, \tag{2.16}$$

where W_M is a dictionary matrix, B_C equals diag(β_{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 mixture STFT and the parameterized variance S_X, defined as $S_X \in \{S_{X_R}, S_{X_L}\}$. The variances for the left and right channel are given by the equations 2.10, 2.14 and 2.16:

$$\begin{cases} S_{x_R,fn} = \alpha_R^2 [(W_\Gamma H_\Gamma H_\phi) \cdot (W_{F_0} H_{F_0})]_{fn} + [W_M B_R H_M]_{fn} \\ S_{x_L,fn} = \alpha_L^2 [(W_\Gamma H_\Gamma H_\phi) \cdot (W_{F_0} H_{F_0})]_{fn} + [W_M B_L H_M]_{fn} \end{cases} \tag{2.17}$$

The parameters $\Theta = \{H_\Gamma, H_\phi, H_{F_0}, W_M, H_M, \alpha_R, \alpha_L, B_R, B_L\}$ are estimated by multiplicative update rules. The source dictionary W_{F_0} is fixed and generated with the previously mentioned glottal model. The smooth elements of the filter W_Γ 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 Θ_0 are randomly initialized.

Second, a smooth path for the fundamental frequencies is computed. Using a Viterbi algorithm, the corresponding H_{F_0} activation coefficients are extracted.

Third, the parameters are again randomly initialized. The new parameter $\bar{H}_{F_0}$ is an exception, in contrast, it is obtained by setting to zero the coefficients of H_{F_0} lying outside a scope of a quarter tone around the tracked melody (step two).

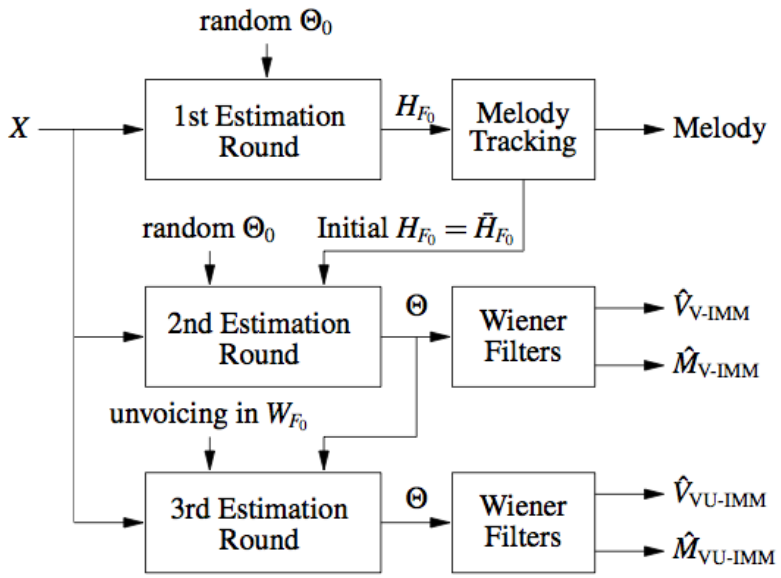


Figure 2.2: Solo/Accompaniment algorithm outline [Durrieu et al., 2009a]

After this second ER round, a first solo/accompaniment separation result V-IMM 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 al the frequencies) is added. By fixing the filter dictionary W_ϕ, 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-IMM 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:

$$G_V(f) = \frac{S_V(f)}{S_V(f) + S_M(f)} \tag{2.18}$$

For the final separation, with the set of parameters Θ 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

$$\hat{x}_t(f) = \sum_{n=1}^{N} g_{n,t} b_n(f) \tag{2.19}$$

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

$$\hat{x}_t(f) = \sum_{j=1}^{J} \sum_{n=1}^{N} g_{n,t,j} \sum_{m=1}^{M} a_{n,m,j} G(f - m f_0(n)) \tag{2.20}$$

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

$$b_{n,j}(f) = h_j(f)e_{n,j}(f). \tag{2.21}$$

A generic harmonic excitation is defined as

$$e_{n,j}(f) = \sum_{m=1}^{M} a_{m,n,j} G(f - mf_0(n)) \tag{2.22}$$

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

$$a_{m,n,j} = \sum_{j=1}^{J} w_{i,n,j} v_{i,m,j} \tag{2.23}$$

where $w_{i,n,j}$ is the weighted sum of the ith 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

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

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

$$\hat{x}_t(f) = \sum_{n,j} g_{n,t,j} h_j(f) \sum_{m=1}^{M} \sum_{i=1}^{I} w_{i,n,j} v_{i,m,j} G(f - mf_0(n)) \tag{2.25}$$

where $n = 1, ..., N$ (with N the number of pitches) and $j = 1, ..., J$ (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 cuasi-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

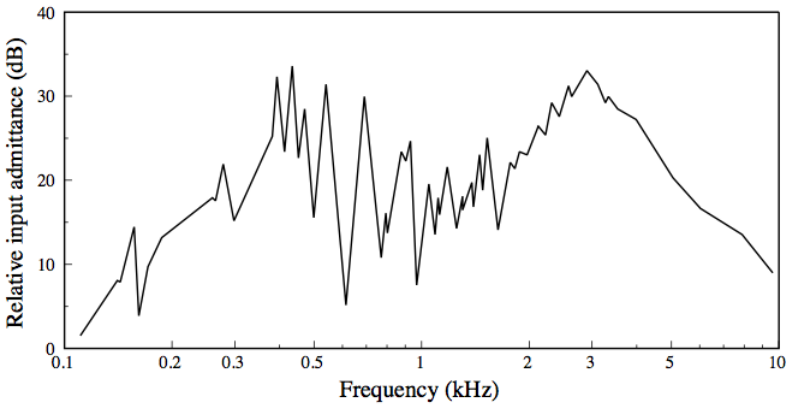


Figure 2.3: Resonance curve of a typical violin body defined as the mechanical input admittance at the bridge [Fletcher, 1999]

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 nth 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

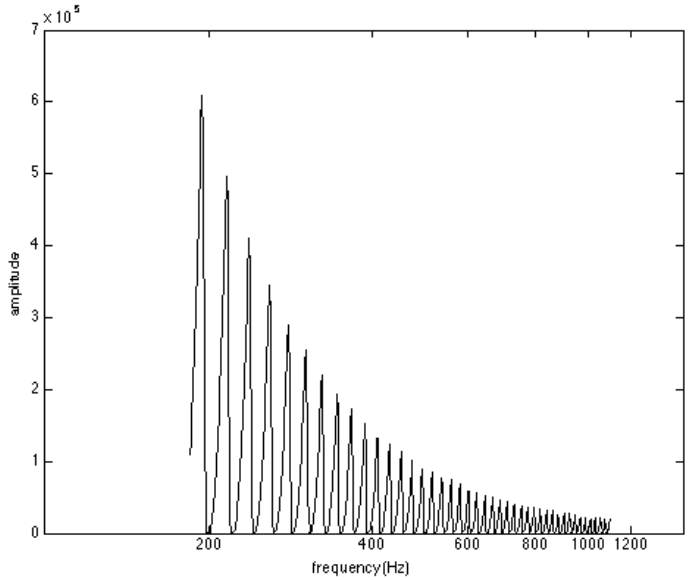


Figure 3.1: Example of an W_{F_0} basis generated with the glottal source model

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 F0 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 off 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.

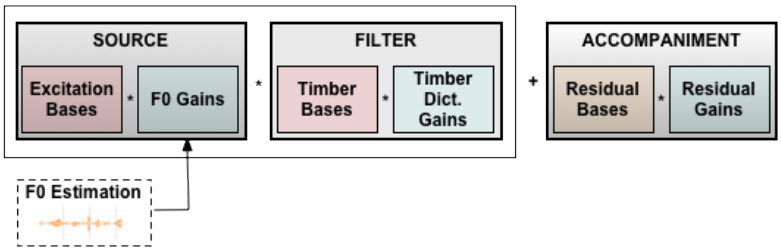


Figure 3.2: Source / filter + accompaniment diagram of the modified IMM

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 bellow.

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

$$S_{F_0} = \sum_{i}^{N_I} S^i_{F_0} = \sum_{i}^{N_I} W^i_{F_0} H^i_{F_0} \tag{3.1}$$

where $S^i_{F_0}$ is now the source variance of instrument i. Each of the instruments provided pitch i is assigned to $H^i_{F_0}$. The bases $W^i_{F_0}$ are generated according to the estimated fundamental frequencies.

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

$$\hat{V}_{mesfIMM} = \sum_{i}^{N_I}((W^i_{F_0}H^i_{F_0}) \cdot (W_\Gamma H_\Gamma H_\phi)) + W_M H_M \tag{3.2}$$

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 $H^i_{F_0}$ 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.

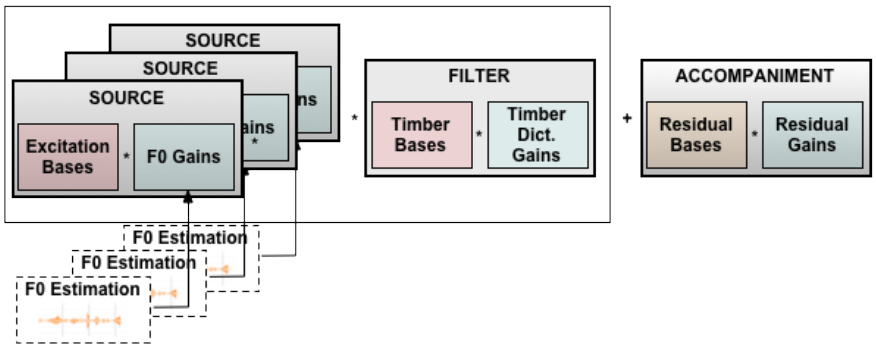


Figure 3.3: Multi-Excitation and Single Filter IMM diagram

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

$$S_\phi = \sum_i^{N_I} S_\phi^i = \sum_i^{N_I} W_\phi^i H_\phi^i = \sum_i^{N_I} W_\Gamma H_\Gamma^i H_\phi^i \tag{3.3}$$

where $S_{F_0}^i$ is now the source variance of instrument i. The W_Γ,H_Γ^i and H_ϕ^i matrices are different for each source.

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

$$\hat{V}_{mefIMM} = \sum_i^{N_I} ((W_{F_0}^i H_{F_0}^i) \cdot (W_\Gamma H_\Gamma^i H_\phi^i)) + W_M H_M \tag{3.4}$$

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_Γ^i, H_ϕ^i and $H_{F_0}^i$ matrices are different for each source sound. Each of the instruments provided pitch i is assigned to $H_{F_0}^i$ and its bases $W_{F_0}^i$ are generated accordingly. When a target source is estimated from the mixture spectrogram only the activations H_Γ^i, H_ϕ^i and $H_{F_0}^i$ of the corresponding pitch are used.

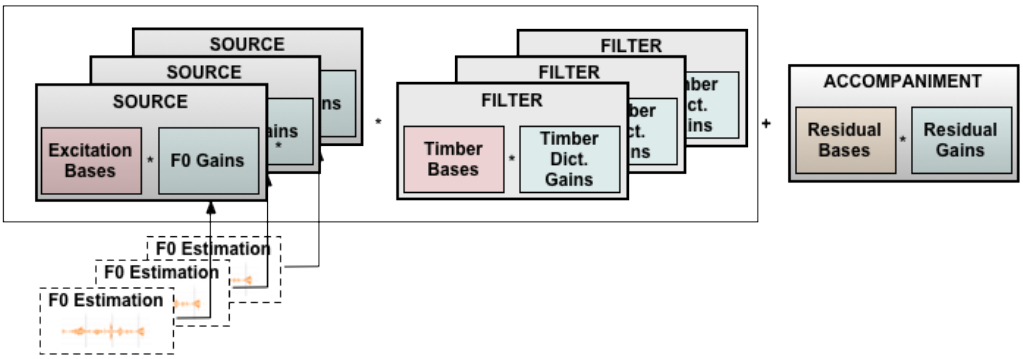


Figure 3.4: Multi-Excitation and Multi-Filter IMM diagram

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_ϕ bases in two other matrixes is introduced (see figure 3.5). The smooth filter dictionary W_ϕ^i 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_Γ and its activations H_Γ^i 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.

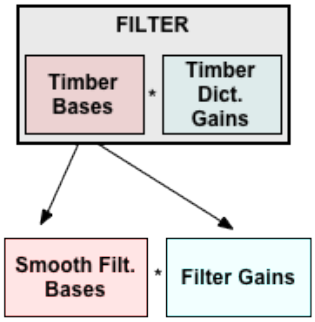


Figure 3.5: Timber model W_ϕ bases decomposed into an smooth filter dictionary bases W_Γ and its gains H_Γ.

The W_Γ 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_Γ are generated, whereas the filter dictionary activations H_Γ^i are estimated and saved. This way, the combination of both bases and activations result in some bases (timbre bases) that can bee 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.

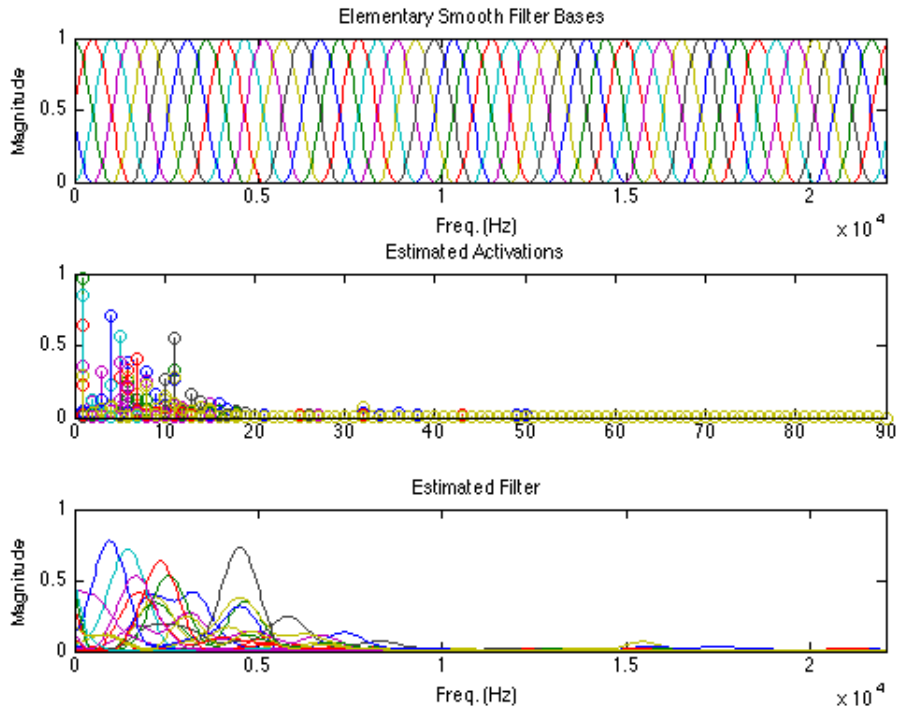


Figure 3.6: Timbre matrices example over a cello excerpt. Smooth filter dictionary bases W_Γ (top), its activations H_Γ (middle) and the obtained bases W_ϕ

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_Γ bases are again generated. The H_Γ^i activations are loaded from the i instrument timbre model. Both matrices create the smooth filter dictionary W_ϕ^i that is kept fixed during the separation. However, the H_ϕ 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_Γ^i 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

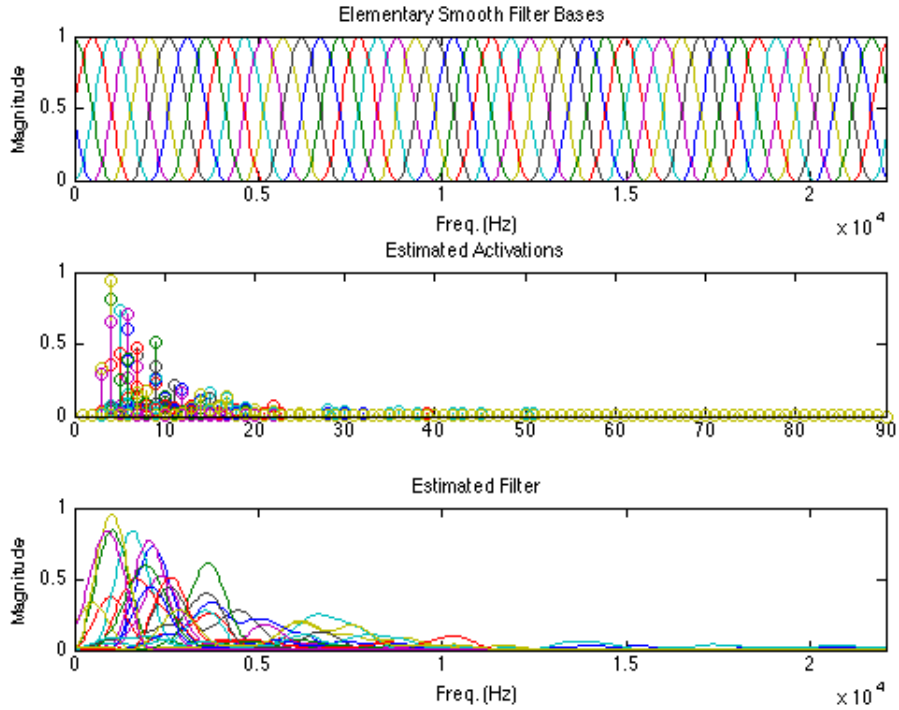


Figure 3.7: Timbre matrices example over a violin excerpt. Smooth filter dictionary bases W_Γ (top), its activations H_Γ (middle) and the obtained bases W_ϕ

each instrument. Figure 3.8 shows a general diagram of the Multi-Excitation and Multi-Filter IMM where the supervised filters are loaded.

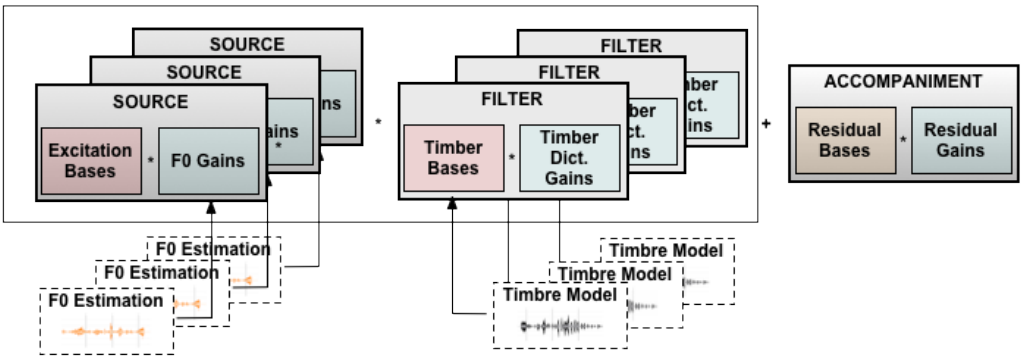


Figure 3.8: Multi-Excitation and Multi-Filter IMM diagram with supervised filters loaded

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 an 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 $\widehat{s_j}$ are approximated to the given true sources s_j [Vincent et al., 2007]. The estimated sources are decomposed as

$$\widehat{s_j} = s_{target} + e_{interf} + e_{noise} + e_{artif} \tag{4.1}$$

[1] http://sisec.wiki.irisa.fr/

The $s_{target} = q(s_j)$ is a modified version of s_j with an allowed distortion q that represents the part of $\widehat{s_j}$ coming from the target source s_j. The e_{interf}, e_{noise}, e_{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:

$$SDR = 10log_{10}\frac{\|s_{\text{target}}\|^2}{\|\widehat{s_j} - s_{\text{target}}\|^2} \tag{4.2}$$

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

$$SIR = 10log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2} \tag{4.3}$$

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

$$SAR = 10log_{10}\frac{\|\widehat{s_j} - e_{\text{artif}}\|^2}{\|e_{\text{artif}}\|^2} \tag{4.4}$$

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 $\widehat{s_j}$ are decomposed in a similar manner to 4.1 as

$$\widehat{s_j} - s_j = s_{\text{target}} + e_{\text{interf}} + e_{\text{artif}} \tag{4.5}$$

In here, the terms e_{interf}, e_{noise}, e_{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

$$q^j_{overall} = \text{PSM}(\widehat{s_j}, s_j) \tag{4.6}$$

$$q^j_{target} = \text{PSM}(\widehat{s_j}, \widehat{s_j} - s_{\text{target}}) \tag{4.7}$$

$$q^j_{interf} = \text{PSM}(\widehat{s_j}, \widehat{s_j} - e_{\text{interf}}) \tag{4.8}$$

$$q^j_{artif} = \text{PSM}(\widehat{s_j}, \widehat{s_j} - e_{\text{artif}}) \tag{4.9}$$

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[2]. More details about the dataset can be found in [Fritsch, 2012].

[2]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_Γ is set to 30, the number of spectral

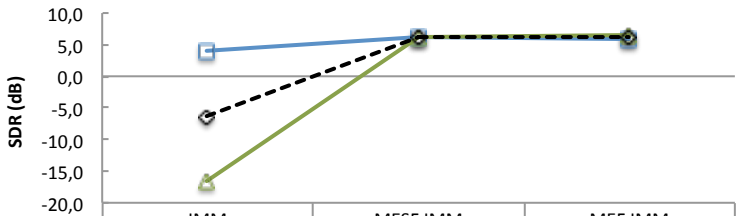
shapes for the filter part is set to 10, the number of elements for the accompaniment is 40 and the number of iterations 30.

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

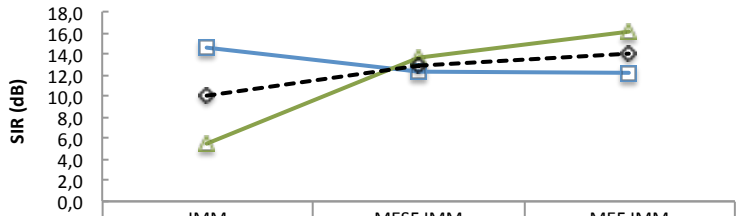
The results are objectively evaluated with the metrics specified in BSS Eval and PEASS toolkits.

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



	IMM	MESF IMM	MEF IMM
violin	4,0278	6,2247	5,7874
cello	-16,6330	6,0827	6,6009
average	-6,3026	6,1537	6,1942

Table 4.1: SDR score results by model for the Schubert audio excerpt



	IMM	MESF IMM	MEF IMM
violin	14,5640	12,3090	12,1960
cello	5,4401	13,6190	16,0080
average	10,0021	12,9640	14,1020

Table 4.2: SIR score results by model for the Schubert audio excerpt

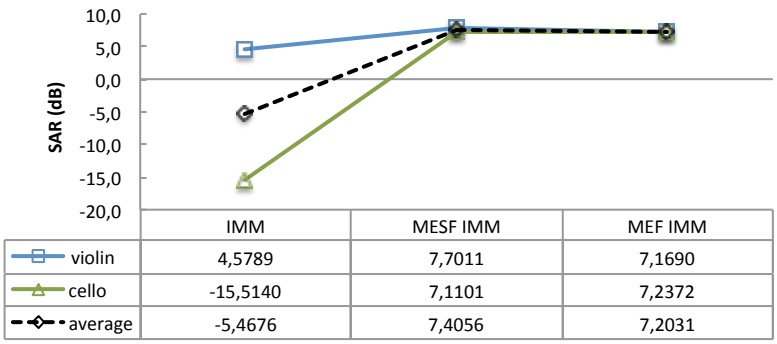


Table 4.3: SAR score results by model for the Schubert audio excerpt

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

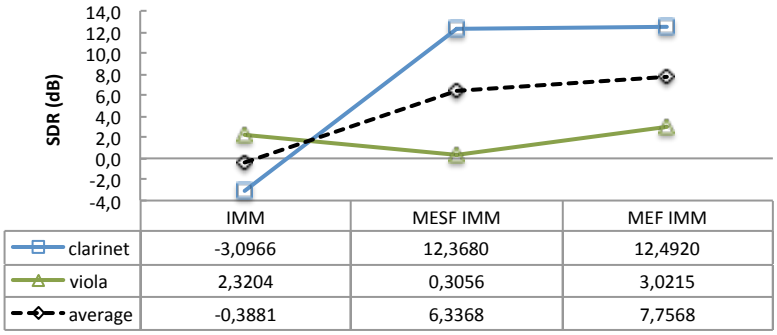


Table 4.4: SDR score results by model for the Mozart audio excerpt

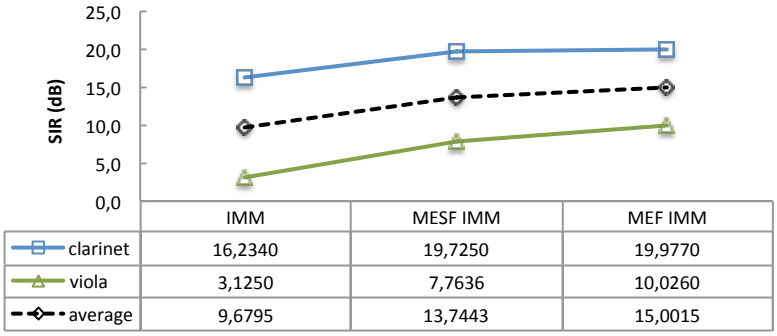
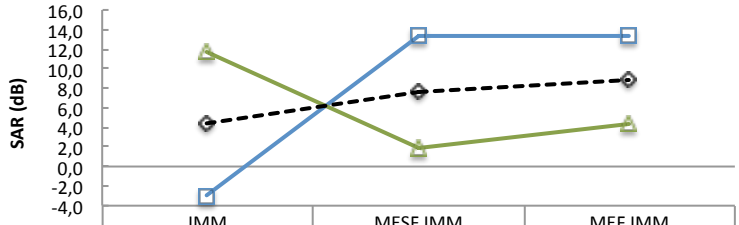


Table 4.5: SIR score results by model for the Mozart audio excerpt



	IMM	MESF IMM	MEF IMM
clarinet	-2,9435	13,2960	13,3890
viola	11,7610	1,8371	4,3983
average	4,4088	7,5666	8,8937

Table 4.6: SAR 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.

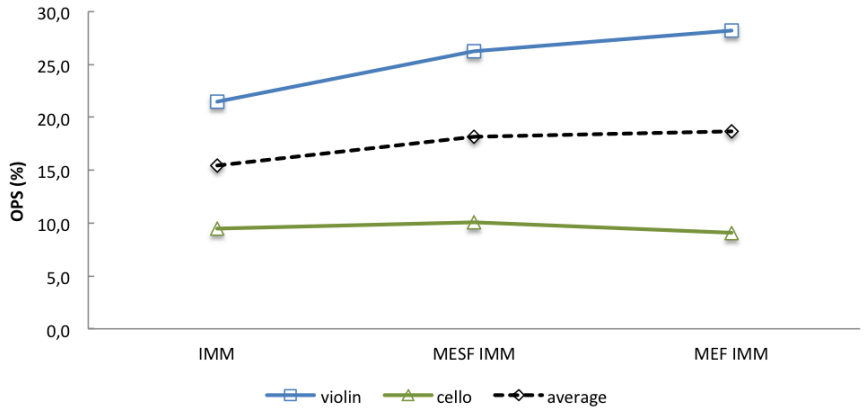


Figure 4.1: Overall Perceptual Score (OPS) results by model for the Schubert audio excerpt

Model	**IMM**				**MESF IMM**				**MEF IMM**			
	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS
violin	21,46	9,34	53,59	19,72	26,22	35,43	47,47	32,37	28,19	32,77	47,74	34,15
cello	9,44	0,79	10,33	15,17	10,12	1,16	6,53	23,10	9,09	1,16	5,13	24,19
average	15,45	5,06	31,96	17,45	18,17	18,29	27,00	27,74	18,64	16,96	26,43	29,17

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

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

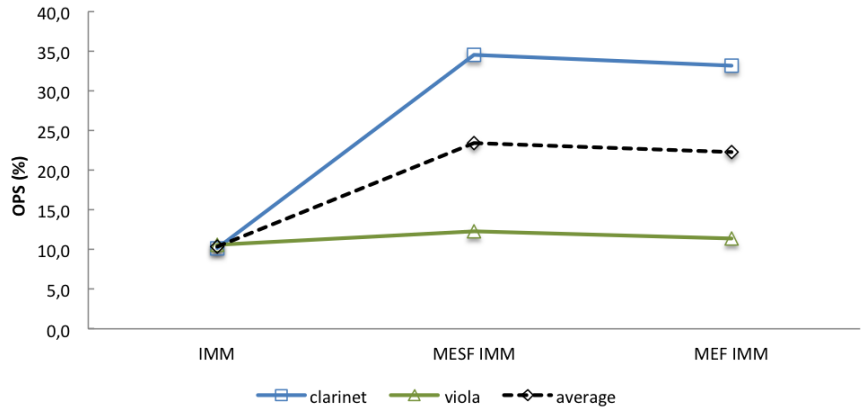


Figure 4.2: Overall Perceptual Score (OPS) results by model for the Mozart audio excerpt

Model	IMM				MESF IMM				MEF IMM			
	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS
clarinet	10,18	0,48	44,18	3,48	34,58	47,81	42,74	47,85	33,25	49,19	42,02	46,95
viola	10,55	0,82	9,30	13,68	12,25	1,82	7,90	16,74	11,42	1,32	2,40	11,86
average	10,37	0,65	26,74	8,58	23,42	24,81	25,32	32,29	22,33	25,25	22,21	29,40

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

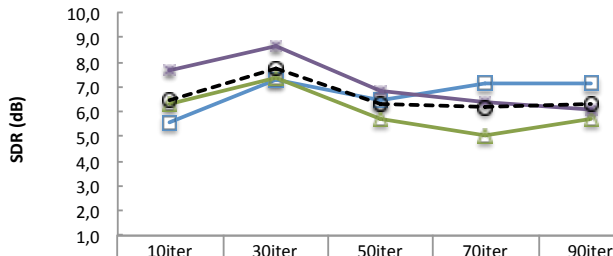
In order to obtain the best parameter configuration, a grid-search is performed over the audio excerpt. The violin and the cello instruments are the ones to be separated. The pitch is extracted over the recording as explained in Section 4.3.1.

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_Γ 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.

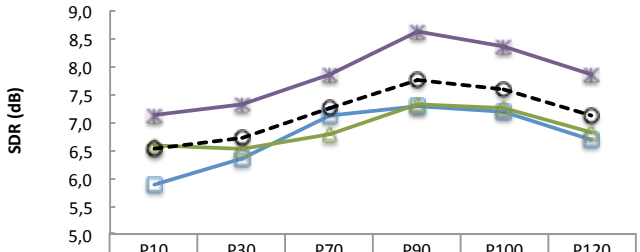
Results

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.



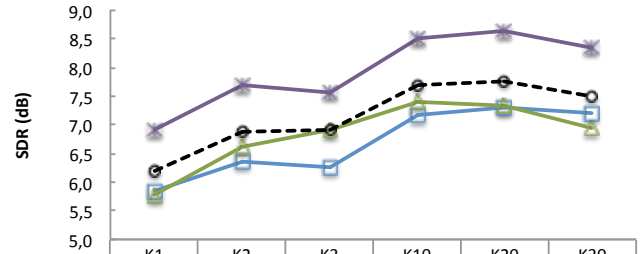
	10iter	30iter	50iter	70iter	90iter
violin	5,5673	7,3023	6,4457	7,1337	7,1651
cello	6,2863	7,3423	5,7395	5,0415	5,7294
accompaniment	7,6324	8,6341	6,8046	6,3999	6,0596
average	6,4953	7,7596	6,3299	6,1917	6,3180

Table 4.9: SDR score results by iteration over 90 P smooth filters and 20 K bases



	P10	P30	P70	P90	P100	P120
violin	5,8796	6,3414	7,1159	7,3023	7,1923	6,6956
cello	6,5791	6,5270	6,7922	7,3423	7,2596	6,8402
accompaniment	7,1346	7,3252	7,8744	8,6341	8,3647	7,8615
average	6,5311	6,7312	7,2608	7,7596	7,6055	7,1324

Table 4.10: SDR score results by P smooth filters over 30 iterations and 20 K bases



	K1	K2	K3	K10	K20	K30
violin	5,8473	6,3531	6,2651	7,1837	7,3023	7,1899
cello	5,7825	6,6088	6,9065	7,3907	7,3423	6,9473
accompaniment	6,9177	7,6831	7,5616	8,5001	8,6341	8,3546
average	6,1825	6,8817	6,9111	7,6915	7,7596	7,4973

Table 4.11: SDR score results by K bases over 90 P smooth filters and 30 iterations

# iterations	*P* smooth filters	*K* bases
30	90	20

Table 4.12: Best parameter configuration over different tested K bases, P smooth filters and iterations values

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_Γ 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.

Model	**MEF IMM**			**MEF IMM + SV FILT.**			**MEF IMM + USV FILT.**		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
violin	8,06	16,53	9,29	6,95	14,80	8,46	-16,10	5,18	6,53
cello	6,09	12,18	8,29	6,05	12,82	7,71	-20,99	9,06	-19,78
accomp.	6,54	9,05	11,35	6,52	9,97	10,53	-8,89	0,74	33,98
average	6,89	12,58	9,64	6,51	12,53	8,90	-15,33	5,00	6,91

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.)

Model	**MEF IMM**			**MEF IMM + SV FILT.**			**MEF IMM + USV FILT.**		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
clarinet	12,64	17,96	14,24	12,56	17,69	14,29	-13,16	9,38	-11,99
viola	-4,12	0,79	5,04	-6,45	0,35	1,00	-20,30	4,72	-18,77
accomp.	6,81	12,24	8,70	6,98	14,75	8,10	-4,49	-4,49	72,50
average	5,11	10,33	9,33	4,36	10,93	7,80	-12,65	3,20	13,91

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 or 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.

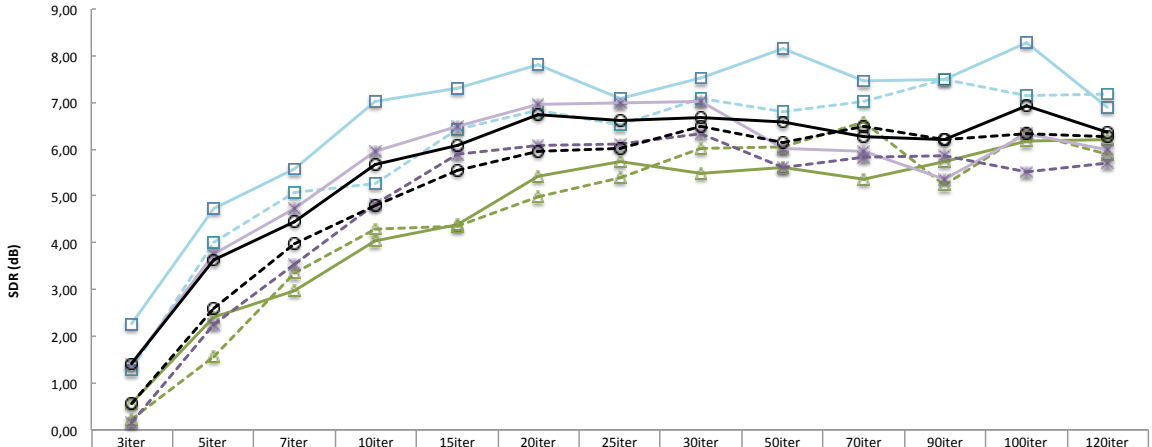
- *Block size* (in frames): 10, 20, 50, 100, 150, 200, 300, 400, 500, 600 and 700.
- Iterations: 3, 5, 7, 10, 15, 20, 25, 30, 50, 70, 90, 100 and 120.

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.

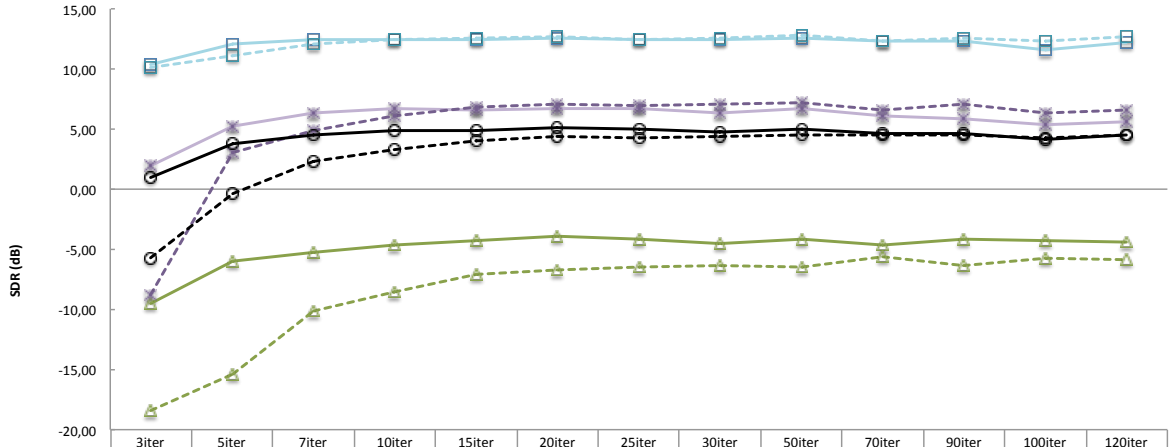
For the different iteration values, the MEF IMM obtains better SDR values than the MEF IMM with the loading of the supervised timbre models, for both Schubert and Mozart recordings (see Table 4.15 and Table 4.16). The use of the filters seems not to improve the separation with lower iteration values.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



	3iter	5iter	7iter	10iter	15iter	20iter	25iter	30iter	50iter	70iter	90iter	100iter	120iter
violin	2,2691	4,7209	5,5788	7,0149	7,3066	7,8159	7,1003	7,5169	8,1461	7,4512	7,5029	8,2824	6,9112
violin filt	1,2952	4,0130	5,0768	5,2544	6,4232	6,8400	6,5088	7,0918	6,8081	7,0127	7,4931	7,1349	7,1750
cello	0,5737	2,4178	2,9853	4,0317	4,4000	5,4342	5,7428	5,4768	5,6246	5,3663	5,7447	6,1886	6,1927
cello filt	0,2265	1,5518	3,3547	4,2823	4,3604	4,9861	5,4059	6,0343	6,0543	6,5842	5,2488	6,3672	5,9040
accomp.	1,4138	3,7554	4,7436	5,9690	6,4943	6,9473	6,9970	7,0374	6,0275	5,9652	5,3533	6,3352	5,9748
accomp. filt	0,1484	2,2631	3,5372	4,8129	5,8861	6,0892	6,1197	6,3475	5,5988	5,8258	5,8554	5,5047	5,7135
average	1,4189	3,6314	4,4359	5,6719	6,0670	6,7325	6,6133	6,6770	6,5994	6,2609	6,2003	6,9354	6,3595
average filt	0,5567	2,6093	3,9896	4,7832	5,5566	5,9718	6,0115	6,4912	6,1537	6,4743	6,1991	6,3356	6,2642

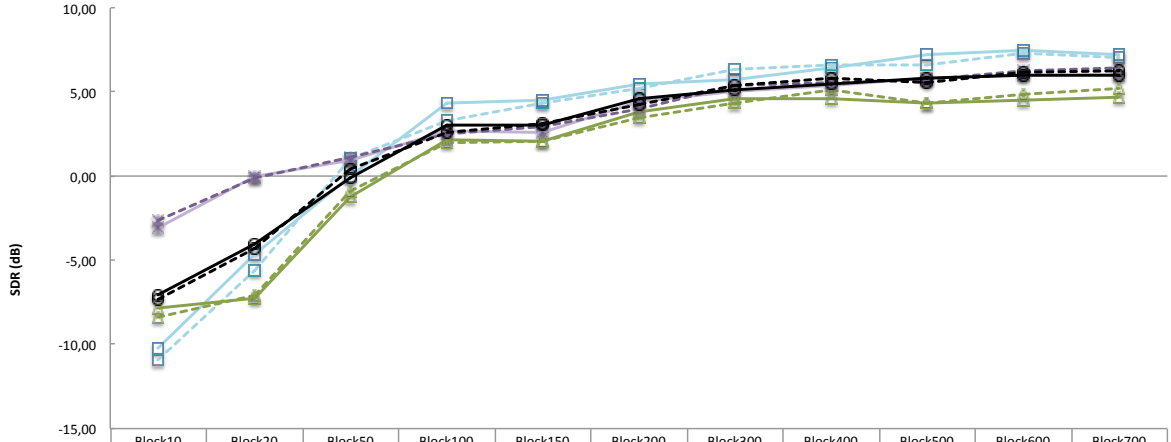
Table 4.15: SDR score of the Schubert recording for the Multi-Excitation and Multi-Filter Model with estimated filters and supervised filters for different iteration values

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).



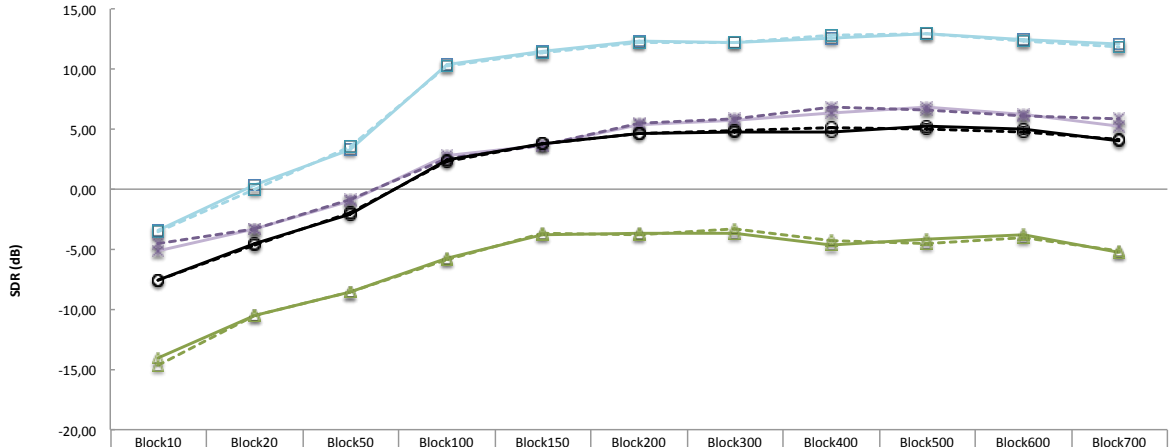
	3iter	5iter	7iter	10iter	15iter	20iter	25iter	30iter	50iter	70iter	90iter	100iter	120iter
clarinet	10,3668	12,0730	12,4338	12,4878	12,4583	12,5180	12,4293	12,4563	12,5838	12,2968	12,3173	11,5440	12,1496
clarinet filt	10,0654	11,1222	12,0480	12,4053	12,5495	12,6395	12,4545	12,5613	12,7480	12,3603	12,6125	12,3345	12,7105
viola	-9,5100	-6,0390	-5,3119	-4,5972	-4,2853	-3,9645	-4,2184	-4,5572	-4,2055	-4,6396	-4,1955	-4,3407	-4,3920
viola filt	-18,4153	-15,3822	-10,1179	-8,5653	-7,1452	-6,6670	-6,4702	-6,3863	-6,5010	-5,5901	-6,3626	-5,7597	-5,8499
accomp.	1,8951	5,2463	6,2942	6,6895	6,6219	6,6796	6,6517	6,3897	6,7612	6,1294	5,8373	5,3400	5,5909
accomp. filt	-8,7764	3,1007	4,8547	6,0541	6,7681	7,0261	6,9029	7,0216	7,1932	6,6003	7,0900	6,3378	6,5941
average	0,9173	3,7601	4,4720	4,8600	4,9316	5,0777	4,9542	4,7629	5,0465	4,5955	4,6530	4,1811	4,4495
average filt	-5,7087	-0,3864	2,2616	3,2980	4,0575	4,3329	4,2957	4,3988	4,4801	4,4568	4,4466	4,3042	4,4849

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



	Block10	Block20	Block50	Block100	Block150	Block200	Block300	Block400	Block500	Block600	Block700
violin	-10,2287	-4,6570	-0,0066	4,3277	4,5297	5,5032	5,7435	6,4503	7,2292	7,4694	7,1777
violin filt	-10,9171	-5,5691	1,0528	3,3034	4,3248	5,2279	6,3201	6,6022	6,5911	7,3197	7,0598
cello	-7,8403	-7,2995	-1,2415	2,1437	2,0376	3,7920	4,5793	4,6095	4,3446	4,5272	4,6920
cello filt	-8,3822	-7,1089	-0,8926	1,9672	2,0301	3,4726	4,3036	5,1345	4,3169	4,8930	5,1702
accomp.	-3,0970	-0,0528	0,9395	2,6564	2,6187	4,4376	5,0518	5,3844	5,7747	5,9763	5,9779
accomp. filt	-2,6817	-0,0934	1,1274	2,5132	2,9012	3,9873	5,3932	5,5848	5,6973	6,2528	6,4426
average	-7,0553	-4,0031	-0,1029	3,0426	3,0620	4,5776	5,1249	5,4814	5,7829	5,9909	5,9492
average filt	-7,3270	-4,2571	0,4292	2,5946	3,0854	4,2293	5,3390	5,7738	5,5351	6,1552	6,2242

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



	Block10	Block20	Block50	Block100	Block150	Block200	Block300	Block400	Block500	Block600	Block700
clarinet	-3,4834	0,3688	3,2421	10,3451	11,4070	12,2725	12,2273	12,5450	12,9585	12,4748	12,0393
clarinet filt	-3,5901	-0,0004	3,4713	10,2934	11,3115	12,2250	12,2193	12,7960	12,8740	12,2585	11,8713
viola	-14,0413	-10,4991	-8,5880	-5,7717	-3,8320	-3,6966	-3,6882	-4,6992	-4,1250	-3,7601	-5,2693
viola filt	-14,6677	-10,4959	-8,5020	-5,8424	-3,6479	-3,8004	-3,3508	-4,3000	-4,4941	-4,0799	-5,1411
accomp.	-5,1673	-3,3623	-0,9507	2,7902	3,6988	5,4105	5,6701	6,3392	6,7690	6,2134	5,2623
accomp. filt	-4,5553	-3,2640	-0,9001	2,5845	3,6802	5,4238	5,8244	6,7868	6,5480	6,0998	5,8181
average	-7,5640	-4,4976	-2,0989	2,4545	3,7579	4,6621	4,7364	4,7283	5,2009	4,9760	4,0107
average filt	-7,6044	-4,5868	-1,9769	2,3452	3,7813	4,6161	4,8976	5,0943	4,9760	4,7595	4,1828

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

- 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.

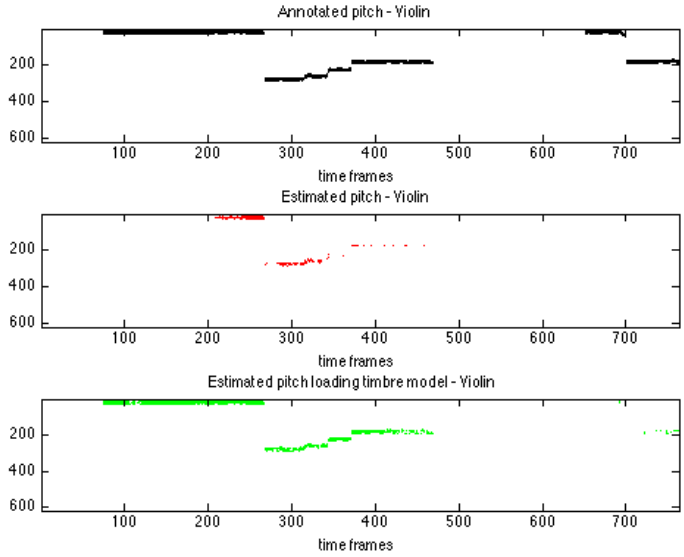


Figure 4.3: Pitch estimations of a 8 second violin excerpt (Schubert recording). Estimations with Yin (top), MEF IMM (middle) and MEF IMM loading supervised timbre models (bottom).

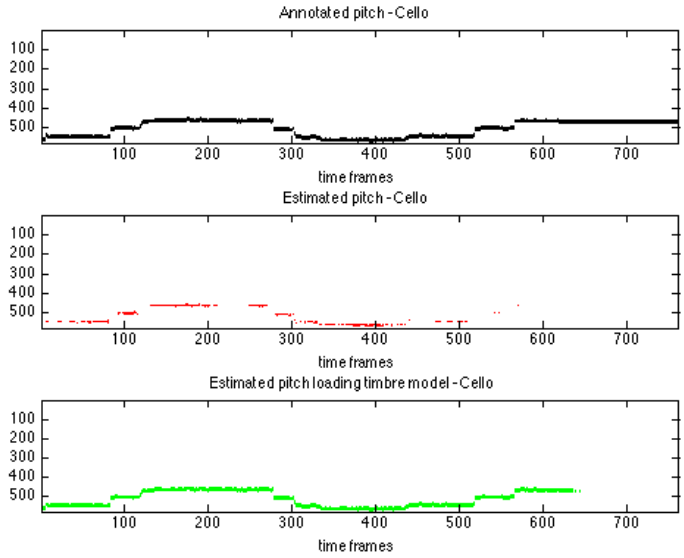


Figure 4.4: Pitch estimations over a 8 second cello excerpt (Schubert recording). Estimations with Yin (top), MEF IMM (middle) and MEF IMM loading supervised timbre models (bottom).

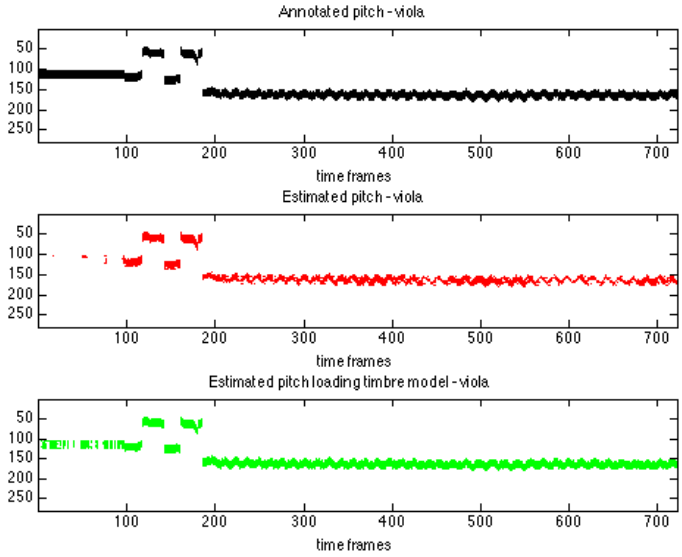


Figure 4.5: Pitch estimations over a 8 second viola excerpt (Mozart recording). Estimations with Yin (top), MEF IMM (middle) and MEF IMM loading supervised timbre models (bottom).

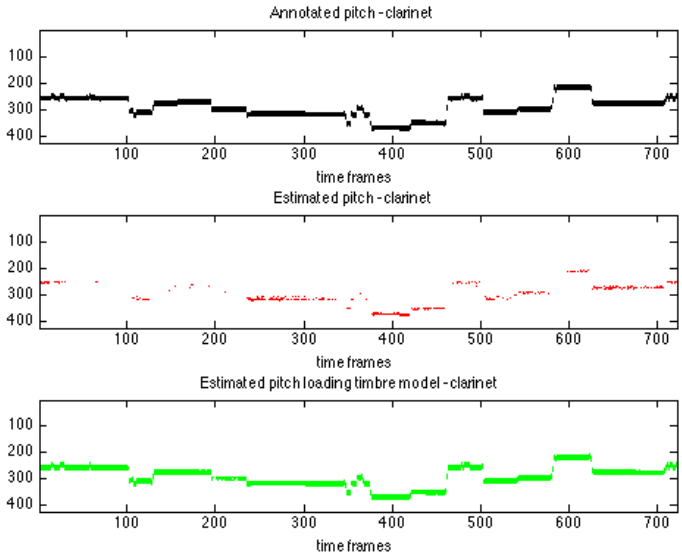


Figure 4.6: Pitch estimations over a 8 second clarinet excerpt (Mozart recording). Estimations with Yin (top), MEF IMM (middle) and MEF IMM loading supervised timbre models (bottom).

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 submodel, 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 ben 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

BSS Eval results for the Schubert excerpt of the parameter optimization test explained in 4.3.2.

	10 Iterations																	
	K1			K2			K3			K10			K20			K30		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
P10	5,4343	11,982	6,788	5,5533	11,724	7,0366	5,6007	11,435	7,2153	5,7193	11,825	7,2171	5,6216	11,852	7,0777	5,5292	11,193	7,2225
	5,7631	15,943	6,3102	5,7899	16,023	6,3295	5,8611	16,167	6,3897	6,0231	14,763	6,789	6,0444	15,053	6,7612	6,2354	15,07	6,9771
	5,9892	7,7445	11,446	5,8774	7,4322	11,814	5,7956	7,2458	12,014	6,467	8,2143	11,876	6,3862	8,1714	11,725	6,4708	8,2931	11,722
	5,7289	11,8898	8,1814	5,7402	11,7264	8,3934	5,7525	11,6159	8,5397	6,0698	11,6008	8,6274	6,0174	11,6921	8,5213	6,0785	11,5187	8,6405
P30	5,053	11,462	6,4798	5,7213	11,693	7,2717	5,7908	12,054	7,2246	5,6122	11,501	7,2034	5,6523	11,742	7,1602	5,5486	10,818	7,4259
	4,8023	14,76	5,4074	5,4228	14,654	6,121	5,3869	13,677	6,2661	5,6254	13,898	6,4988	5,731	13,607	6,6894	6,1382	14,454	6,9837
	5,7727	8,002	10,374	6,578	9,1481	10,577	6,951	9,6668	10,723	7,0428	9,8595	10,683	7,1751	10,076	10,705	7,398	10,796	10,399
	5,2093	11,4080	7,4204	5,9074	11,8317	7,9899	6,0429	11,7993	8,0712	6,0935	11,7528	8,1284	6,1861	11,8083	8,1849	6,3616	12,0227	8,2695
P70	5,5272	11,032	7,2927	5,8034	12,151	7,2057	5,7099	11,902	7,1754	5,7865	11,612	7,3927	5,7031	11,393	7,3728	5,5617	11,28	7,2286
	5,5452	16,513	6,0032	6,0398	14,629	6,8337	6,1184	14,491	6,9528	6,2557	14,826	7,0463	6,3057	14,666	7,1364	6,2608	14,172	7,1902
	6,4504	9,1124	10,342	7,3741	10,587	10,555	7,3264	10,563	10,487	7,515	11,027	10,406	7,56	11,062	10,456	7,8069	11,849	10,259
	5,8409	12,2191	7,8793	6,4058	12,4557	8,1981	6,3849	12,3187	8,2051	6,5191	12,4883	8,2817	6,5229	12,3737	8,3217	6,5431	12,4337	8,2259
P90	5,6269	12,386	6,8994	5,6419	12,05	7,0322	5,4128	10,752	7,266	5,6929	11,774	7,2016	5,5673	11,367	7,1988	5,7041	11,521	7,3193
	5,5453	14,422	6,3019	5,9256	14,472	6,7318	5,9195	15,217	6,5914	6,2981	14,518	7,1581	6,2863	14,345	7,1817	6,2948	14,263	7,2099
	6,9605	9,8832	10,485	7,2127	10,25	10,587	6,8147	9,7466	10,342	7,7533	11,463	10,461	7,6324	11,435	10,274	7,8274	11,698	10,404
	6,0442	12,2304	7,8954	6,2601	12,2573	8,1170	6,0490	11,9052	8,0665	6,5814	12,5850	8,2736	6,4953	12,3823	8,2182	6,6088	12,4940	8,3111

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

	30 Iterations																	
	K1			K2			K3			K10			K20			K30		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
P10	5,4712	12,647	6,625	6,03	12,592	7,3456	5,7172	12,054	7,1287	5,6865	12,721	6,8705	5,8796	12,293	7,2544	6,0538	12,529	7,3977
	6,4655	15,953	7,0929	5,7339	15,837	6,292	5,9082	16,918	6,3542	6,0843	14,719	6,8679	6,5791	15,497	7,2962	6,3711	15,361	7,0815
	6,3553	8,0148	11,974	6,2291	7,8373	11,984	6,2999	8,0826	11,656	6,9579	8,9949	11,74	7,1346	9,1436	11,948	6,9429	8,9251	11,826
	6,0973	12,2049	8,5640	5,9977	12,0888	8,5405	5,9751	12,3515	8,3796	6,2429	12,1450	8,4928	6,5311	12,3112	8,8329	6,4559	12,2717	8,7684
P30	5,516	13,182	6,5353	6,0804	13,569	7,1201	6,4193	13,202	7,6447	5,8077	12,83	6,9892	6,3414	13,07	7,5873	6,6839	13,383	7,9237
	6,2825	15,782	6,9126	6,659	16,089	7,2904	6,8854	16,45	7,4915	6,3223	15,567	6,9914	6,527	15,333	7,2654	6,2049	14,785	6,9953
	6,9794	9,372	11,185	7,0139	9,0983	11,706	7,5168	9,7577	11,899	6,9176	9,0433	11,55	7,3252	9,4898	11,85	7,4458	9,8771	11,55
	6,2593	12,7787	8,2110	6,5844	12,9188	8,7055	6,9405	13,1366	9,0117	6,3492	12,4801	8,5102	6,7312	12,6309	8,9009	6,7782	12,6817	8,8230
P70	5,8308	13,512	6,8322	6,4771	13,082	7,7565	6,35	13,258	7,5402	6,8866	13,518	8,1391	7,1159	13,562	8,4193	7,1391	13,757	8,3855
	6,1309	15,179	6,8383	6,9485	16,388	7,5717	6,9233	15,117	7,7688	6,245	15,104	6,982	6,7922	15,561	7,5299	6,4658	15,043	7,2484
	6,9437	9,3657	11,11	7,5257	9,8862	11,725	7,7005	10,047	11,904	7,3897	9,5277	11,952	7,8744	10,374	11,845	8,0474	10,731	11,764
	6,3018	12,6856	8,2602	6,9838	13,1187	9,0177	6,9913	12,8073	9,0710	6,8404	12,7166	9,0244	7,2608	13,1657	9,2647	7,2174	13,1770	9,1326
P90	5,8473	13	6,9891	6,3531	13,064	7,604	6,2651	13,167	7,4611	7,1837	13,691	8,4642	7,3023	13,847	8,5655	7,1899	13,611	8,4987
	5,7825	14,664	6,5302	6,6088	15,816	7,2765	6,9065	15,747	7,6284	7,3907	16,331	8,0836	7,3423	16,002	8,0857	6,9473	15,744	7,6759
	6,9177	9,2597	11,205	7,6831	10,461	11,313	7,5616	9,972	11,685	8,5001	11,394	11,935	8,6341	11,816	11,756	8,3546	11,303	11,738
	6,1825	12,3079	8,2414	6,8817	13,1137	8,7312	6,9111	12,9620	8,9248	7,6915	13,8053	9,4943	7,7596	13,8883	9,4691	7,4973	13,5527	9,3042
P100	5,5235	12,583	6,7085	5,9499	13,862	6,891	6,5349	13,314	7,756	6,4827	13,22	7,7193	7,1923	13,616	8,5	7,2567	13,75	8,5386
	5,7526	15,116	6,4188	6,353	15,2	7,0891	6,6174	15,606	7,3213	6,9552	15,918	7,6547	7,2596	16,513	7,9045	7,342	16,658	7,9754
	6,2385	8,0927	11,455	7,3515	9,9084	11,29	7,4758	9,7479	11,813	7,6363	10,207	11,531	8,3647	11,208	11,866	8,2874	11,275	11,633
	5,8382	11,9306	8,1941	6,5515	12,9901	8,4234	6,8760	12,8893	8,9634	7,0247	13,1150	8,9683	7,6055	13,7790	9,4235	7,6287	13,8943	9,3823
P120	6,0161	13,512	7,0567	7,1512	14,071	8,3047	6,6956	13,408	7,9306	6,7458	13,072	8,107	6,6956	13,478	7,9089	7,204	13,525	8,5464
	6,0949	15,001	6,828	6,9368	16,512	7,5402	6,762	16,225	7,386	6,8703	16,581	7,4556	6,8402	15,803	7,5426	7,501	15,976	8,2749
	7,3367	9,9772	11,167	8,0636	10,802	11,711	7,4623	9,7233	11,819	7,9421	10,639	11,649	7,8615	10,619	11,502	8,9586	12,385	11,834
	6,4826	12,8301	8,3506	7,3839	13,7950	9,1853	6,9733	13,1188	9,0452	7,1861	13,4307	9,0705	7,1324	13,3000	8,9845	7,8879	13,9620	9,5518

Table A.2: BSS Eval results for the Schubert excerpt over 30 iterations and different values for P smooth filters and K bases

	50 Iterations																	
	K1			K2			K3			K10			K20			K30		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
P10	5,027	13,652	5,8515	5,4336	13,129	6,4491	5,511	12,46	6,7301	5,3698	11,822	6,7607	5,0474	11,243	6,5546	6,3012	13,249	7,4818
	5,653	16,371	6,1367	6,2065	17,472	6,6208	5,3689	15,95	5,8754	5,4876	15,79	6,0258	5,4684	14,282	6,2394	5,7016	15,833	6,2567
	5,2303	6,642	11,649	5,1869	6,3362	12,43	5,2418	6,5942	11,827	4,6992	5,7058	12,576	6,0344	7,6143	11,886	6,1419	7,7544	11,899
	5,3034	12,2217	7,8791	5,6090	12,3124	8,5000	5,3739	11,6681	8,1442	5,1855	11,1059	8,4542	5,5167	11,0464	8,2267	6,0482	12,2788	8,5458
P30	5,3009	12,69	6,4033	5,4396	12,648	6,5856	5,3764	12,373	6,5883	4,7783	11,904	5,9855	4,7594	11,92	5,9573	6,1527	12,828	7,4248
	5,5829	16,215	6,0782	6,4622	15,971	7,0864	6,178	16,36	6,715	6,4899	16,061	7,1041	6,0379	15,249	6,7206	5,9853	15,792	6,5778
	5,7083	7,3481	11,467	6,3717	8,0629	11,916	5,8327	7,2984	12,004	5,6599	7,0987	11,931	5,294	6,6515	11,855	6,4579	8,2481	11,777
	5,5307	12,0844	7,9828	6,0912	12,2273	8,5293	5,7957	12,0105	8,4358	5,6427	11,6879	8,3402	5,3638	11,2735	8,1776	6,1986	12,2894	8,5932
P70	5,1418	12,72	6,2016	5,5506	12,663	6,7187	6,2975	12,898	7,587	6,8132	13,34	8,1029	6,4597	13,158	7,7096	7,088	13,803	8,3058
	6,081	15,74	6,6925	6,2946	17,163	6,7482	6,4785	16,485	7,0319	6,2747	16,74	6,7745	5,4639	15,088	6,0979	7,0206	16,612	7,6198
	6,1891	7,959	11,587	6,0879	7,7485	11,741	6,9461	8,8895	11,901	6,8651	9,0555	11,396	6,3113	8,2046	11,441	7,7294	10,057	11,958
	5,8040	12,1397	8,1604	5,9777	12,5248	8,4026	6,5740	12,7575	8,8400	6,6510	13,0452	8,7578	6,0783	12,1502	8,4162	7,2793	13,4907	9,2945
P90	6,3425	13,95	7,3414	6,1536	13,556	7,2127	6,1709	13,022	7,3867	5,8436	12,333	7,1941	6,4457	13,164	7,6902	6,9614	13,209	8,3405
	5,969	16,078	6,5204	6,6268	16,313	7,221	6,4633	17,037	6,9466	6,8828	17,423	7,3622	5,7395	15,916	6,2876	7,3075	16,983	7,8888
	6,6502	8,8992	11,111	6,7215	8,6524	11,727	6,5327	8,2882	11,915	6,047	7,5669	12,045	6,8046	9,0728	11,217	8,2405	11,031	11,812
	6,3206	12,9757	8,3243	6,5006	12,8405	8,7202	6,3890	12,7824	8,7494	6,2578	12,4410	8,8671	6,3299	12,7176	8,3983	7,5031	13,7410	9,3471
P100	6,104	13,339	7,2105	6,2914	14,221	7,2156	6,4458	13,57	7,5687	6,066	13,283	7,1793	6,931	14,276	7,9751	6,487	13,266	7,7105
	5,6507	16,374	6,1338	6,3278	15,991	6,9324	6,2915	17,143	6,747	6,4823	15,937	7,1142	6,7846	15,98	7,4498	7,1373	16,385	7,7857
	5,8857	7,7549	11,121	6,6253	8,4813	11,789	6,3707	8,1754	11,671	6,0459	7,4647	12,311	7,5703	10,029	11,624	7,2097	9,245	11,968
	5,8801	12,4893	8,1551	6,4148	12,8978	8,6457	6,3693	12,9628	8,6622	6,1981	12,2282	8,8682	7,0953	13,4283	9,0163	6,9447	12,9653	9,1547
P120	4,823	12,705	5,822	5,7908	13,549	6,7751	5,8582	13,129	6,966	5,6262	12,412	6,8901	7,0581	13,873	8,2466	6,9105	13,323	8,2342
	6,2594	15,864	6,8744	6,0311	16,908	6,4889	5,3179	15,803	5,8374	6,639	16,84	7,1637	7,2427	16,909	7,8266	7,5461	17,398	8,0988
	5,936	7,7938	11,187	5,8968	7,5438	11,609	6,0311	7,9261	11,195	6,4018	8,219	11,672	7,5306	9,7261	11,984	7,607	9,8902	11,915
	5,6728	12,1209	7,9611	5,9062	12,6669	8,2910	5,7357	12,2860	7,9995	6,2223	12,4903	8,5753	7,2771	13,5027	9,3524	7,3545	13,5371	9,4160

Table A.3: BSS Eval results for the Schubert excerpt over 50 iterations and different values for P smooth filters and K bases

	70 Iterations																	
	K1			K2			K3			K10			K20			K30		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
P10	5,1135	12,739	6,1625	4,7856	11,88	6,0018	5,0594	12,218	6,2404	5,1599	11,901	6,465	4,9979	11,826	6,2845	4,441	12,16	5,5019
	4,8886	17,415	5,2165	5,7218	18,505	6,0177	5,1803	16,456	5,6138	5,8189	17,277	6,2215	5,01	16,613	5,4149	4,7525	15,599	5,2435
	3,947	4,7123	13,128	4,6727	5,8315	11,985	4,279	5,1441	12,87	4,3546	5,2295	12,882	4,7966	6,0045	11,917	4,3695	5,5552	11,653
	4,6497	11,6221	8,1690	5,0600	12,0722	8,0015	4,8396	11,2727	8,2414	5,1111	11,4692	8,5228	4,9348	11,4812	7,8721	4,5210	11,1047	7,4661
P30	4,6782	12,828	5,6208	5,0264	12,68	6,0724	5,1752	12,692	6,2494	4,9387	12,225	6,0893	4,918	12,211	6,0679	5,3308	12,793	6,4119
	5,9054	16,679	6,3773	6,2618	16,987	6,7319	5,6456	15,827	6,1955	5,6802	16,597	6,1409	6,6818	14,994	7,5102	6,3003	15,45	6,9855
	4,9266	6,1602	11,936	5,3154	6,6536	11,93	4,8571	5,903	12,545	4,7071	5,7996	12,25	5,831	7,2211	12,209	5,394	6,6213	12,337
	5,1701	11,8891	7,9780	5,5345	12,1069	8,2448	5,2260	11,4740	8,3300	5,1087	11,5405	8,1601	5,8103	11,4754	8,5957	5,6750	11,6214	8,5781
P70	5,1802	12,511	6,3052	5,5307	12,67	6,6918	4,6824	12,224	5,7765	6,1874	13,593	7,2443	6,5058	13,759	7,5905	6,4163	13,151	7,6569
	5,364	16,061	5,8569	6,1763	17,908	6,5478	4,468	16,238	4,8691	5,6968	15,626	6,2796	6,806	17,345	7,2869	5,7743	16,416	6,2641
	5,0471	6,2804	12,034	5,3171	6,5213	12,35	4,3446	5,5735	11,49	5,905	7,5979	11,512	6,6566	8,5841	11,676	6,2019	7,9998	11,539
	5,1971	11,6175	8,0654	5,6747	12,3664	8,5299	4,4983	11,3452	7,3785	5,9297	12,2723	8,3453	6,6561	13,2294	8,8511	6,1308	12,5223	8,4867
P90	5,5358	13,924	6,3882	5,5715	12,533	6,7835	5,8965	13,098	7,0221	6,2644	13,163	7,4614	7,1337	13,898	8,3346	5,9302	13,348	6,9951
	6,4572	17,348	6,9054	5,6627	17,254	6,0556	5,3477	15,838	5,8659	5,8521	16,96	6,289	5,0415	15,536	5,5672	5,3643	15,682	5,9038
	5,1002	6,2453	12,374	4,9728	6,1243	12,25	5,4947	6,9061	11,869	6,0098	7,7633	11,468	6,3999	8,4255	11,269	5,8773	7,9161	10,792
	5,6977	12,5058	8,5559	5,4023	11,9704	8,3630	5,5796	11,9474	8,2523	6,0421	12,6288	8,4061	6,1917	12,6198	8,3903	5,7239	12,3154	7,8970
P100	5,2378	12,669	6,3321	5,6722	12,798	6,8303	4,8793	11,897	6,1135	4,9387	12,372	6,0482	5,975	13,043	7,1351	6,4397	13,422	7,6039
	5,8148	16,062	6,352	5,8158	17,259	6,2199	5,9238	18,037	6,267	5,8684	15,896	6,4332	5,7968	17,298	6,1957	6,7954	16,063	7,4487
	5,6096	7,0994	11,753	5,4342	6,8744	11,739	4,6263	5,6487	12,455	5,3992	6,8757	11,613	6,26	8,2816	11,154	7,2328	9,3494	11,847
	5,5541	11,9435	8,1457	5,6407	12,3105	8,2631	5,1431	11,8609	8,2785	5,4021	11,7146	8,0315	6,0106	12,8742	8,1616	6,8226	12,9448	8,9665
P120	5,0551	12,311	6,208	6,0025	12,971	7,1907	5,8349	13,122	6,9394	6,5779	14,106	7,5878	6,2298	12,891	7,5028	6,5284	13,427	7,7134
	5,1128	16,671	5,5198	5,662	17,942	5,996	5,5297	17,625	5,8808	6,2168	16,508	6,7393	6,682	16,98	7,1938	6,8548	16,774	7,4118
	4,7596	5,9837	11,833	4,7605	5,7407	12,733	4,8821	6,0429	12,145	6,3259	8,0301	11,848	6,2156	7,8028	12,023	7,1278	9,4834	11,373
	4,9758	11,6552	7,8536	5,4750	12,2179	8,6399	5,4156	12,2633	8,3217	6,3735	12,8814	8,7250	6,3758	12,5579	8,9065	6,8370	13,2281	8,8327

Table A.4: BSS Eval results for the Schubert excerpt over 70 iterations and different values for P smooth filters and K bases

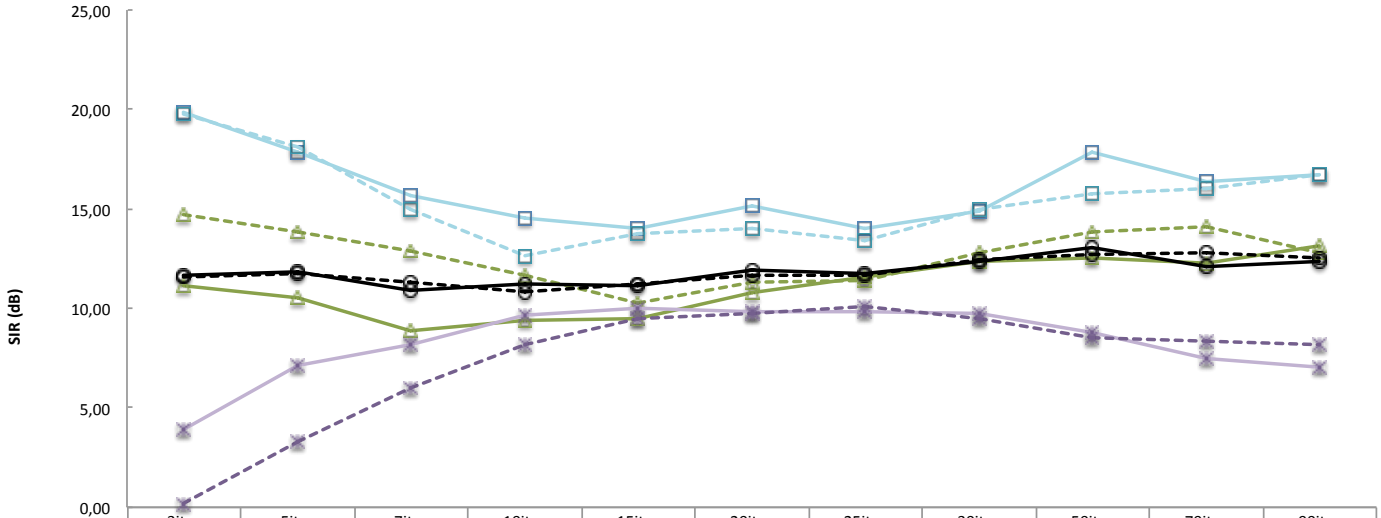
	90 Iterations																	
	K1			K2			K3			K10			K20			K30		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
P10	5,1413	13,47	6,0226	5,3349	12,42	6,5224	4,6615	12,845	5,5971	4,9723	11,918	6,2232	5,263	13,935	6,0688	4,5746	11,852	5,7494
	4,7717	17,306	5,1011	5,5866	16,535	6,0459	4,5997	15,442	5,0953	4,6454	15,075	5,1907	5,1817	14,633	5,8518	5,1418	15,175	5,7257
	3,929	4,7767	12,689	4,42	5,3255	12,791	3,8487	4,7312	12,462	3,9586	4,8009	12,738	4,2116	5,1046	12,688	4,558	5,7406	11,812
	4,6140	11,8509	7,9376	5,1138	11,4268	8,4531	4,3700	11,0061	7,7181	4,5254	10,5980	8,0506	4,8854	11,2242	8,2029	4,7581	10,9225	7,7624
P30	4,9797	12,659	6,0216	4,9431	12,515	6,0148	4,6708	12,346	5,7303	4,1699	11,494	5,3566	4,4079	11,976	5,5109	4,4301	11,872	5,5664
	4,9851	16,972	5,3554	4,8689	16,683	5,2569	4,9126	17,828	5,2114	5,0662	16,568	5,4796	6,0354	15,492	6,6787	5,0898	15,723	5,5972
	4,1376	5,0801	12,41	4,5585	5,7527	11,773	4,2481	5,3378	11,901	4,1647	5,337	11,54	5,1643	6,55	11,668	4,6061	6,0649	11,013
	4,7008	11,5704	7,9290	4,7902	11,6502	7,6816	4,6105	11,8373	7,6142	4,4669	11,1330	7,4587	5,2025	11,3393	7,9525	4,7087	11,2200	7,3922
P70	4,8659	12,642	5,889	5,5185	12,561	6,7089	4,314	11,656	5,4862	5,61	13,417	6,5901	7,319	14,836	8,306	5,4388	13,393	6,3913
	5,7908	17,103	6,2081	5,0501	16,852	5,4355	5,3254	16,959	5,7208	6,6777	18,061	7,0729	6,5226	18,152	6,8977	5,694	15,318	6,3212
	4,7966	6,0546	11,754	4,8434	6,0954	11,81	4,0816	5,0215	12,379	5,6764	7,2669	11,557	6,4793	8,254	11,828	5,7896	7,5093	11,354
	5,1511	11,9332	7,9504	5,1373	11,8361	7,9848	4,5737	11,2122	7,8620	5,9880	12,9150	8,4067	6,7736	13,7473	9,0106	5,6408	12,0734	8,0222
P90	5,3099	12,918	6,3536	4,4933	11,742	5,6813	4,9245	12,758	5,9302	6,3411	13,741	7,393	7,1651	14,231	8,2764	6,1046	12,918	7,3351
	5,6446	16,669	6,0945	5,8969	18,079	6,235	5,7965	17,488	6,1779	5,4196	16,302	5,89	5,7294	15,176	6,3829	6,5343	16,497	7,0922
	5,1527	6,469	11,861	4,6333	5,7566	12,079	4,6685	5,7837	12,136	5,4704	6,7561	12,215	6,0596	7,6822	11,805	6,6764	8,6344	11,636
	5,3691	12,0187	8,1030	5,0078	11,8592	7,9984	5,1298	12,0099	8,0814	5,7437	12,2664	8,4993	6,3180	12,3631	8,8214	6,4384	12,6831	8,6878
P100	5,3538	12,554	6,5066	5,6877	13,333	6,7043	5,5719	13,061	6,6342	6,0787	13,089	7,2508	6,9277	14,112	8,0149	6,9532	13,964	8,0881
	4,9882	16,688	5,3843	5,6902	17,367	6,0748	5,7713	16,939	6,2035	5,5057	16,733	5,9372	7,2755	16,717	7,8914	5,7462	17,015	6,1687
	4,922	6,3225	11,429	5,0583	6,2593	12,15	4,7817	5,8196	12,517	5,1448	6,3952	12,059	6,7709	8,3569	12,506	5,8897	7,7185	11,207
	5,0880	11,8548	7,7733	5,4787	12,3198	8,3097	5,3750	11,9399	8,4516	5,5764	12,0724	8,4157	6,9914	13,0620	9,4708	6,1964	12,8992	8,4879
P120	5,2078	12,797	6,2615	5,5086	12,899	6,6002	5,2635	13,386	6,185	6,9842	14,022	8,1102	6,9157	14,319	7,945	6,5336	13,78	7,6191
	5,2775	16,635	5,7007	5,5663	17,643	5,9184	5,9878	17,981	6,3399	6,6839	18,058	7,08	6,7053	18,659	7,0502	5,3032	15,036	5,9257
	5,0931	6,4656	11,648	5,121	6,5076	11,63	4,7704	5,8186	12,468	6,1688	7,545	12,534	6,171	7,9634	11,524	5,58	7,1182	11,605
	5,1928	11,9659	7,8701	5,3986	12,3499	8,0495	5,3406	12,3952	8,3310	6,6123	13,2083	9,2414	6,5973	13,6471	8,8397	5,8056	11,9781	8,3833

Table A.5: BSS Eval results for the Schubert excerpt over 90 iterations and different values for P smooth filters and K bases

Appendix B

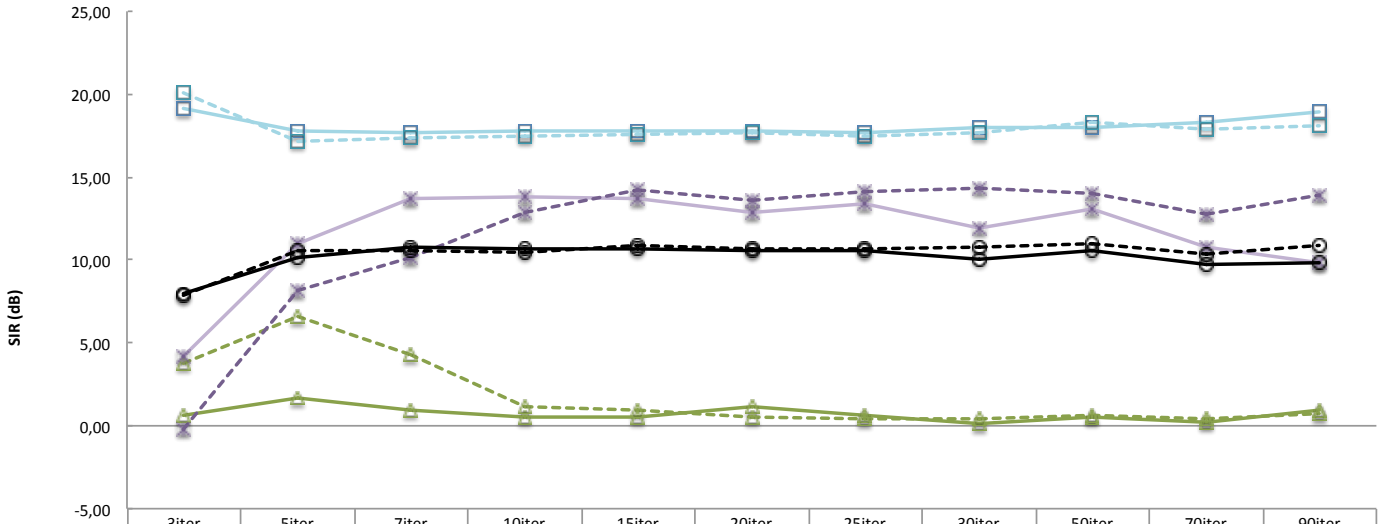
APPENDIX B

SIR and SAR error values of Schubert and Mozart recordings over different iteration values corresponding to the test presented in 4.3.3.



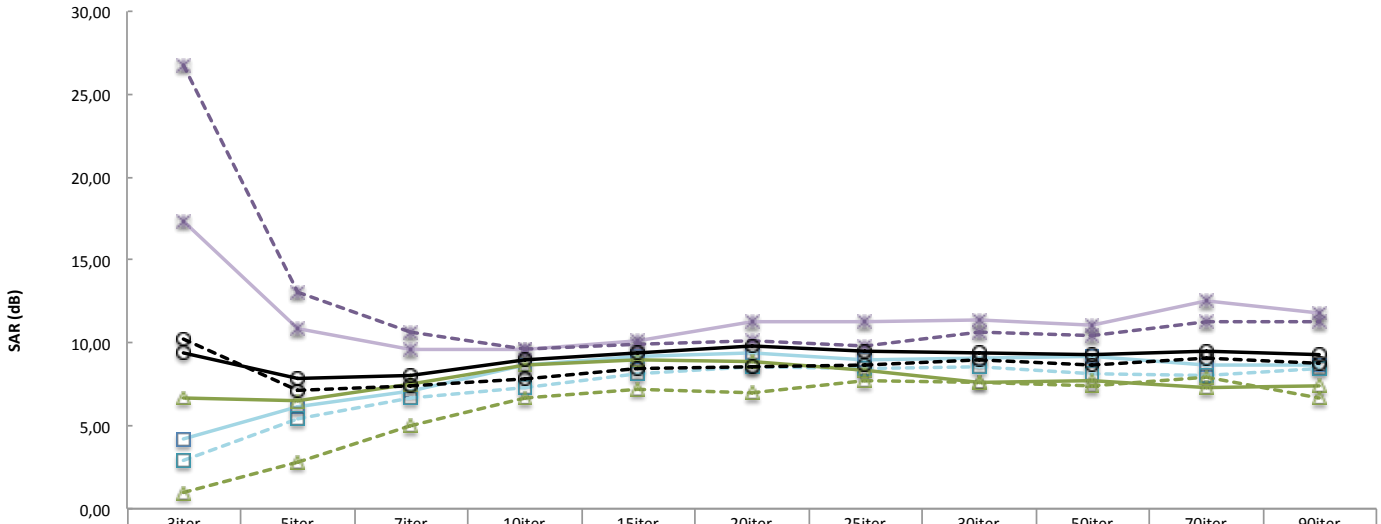
	3iter	5iter	7iter	10iter	15iter	20iter	25iter	30iter	50iter	70iter	90iter
violin	19,8926	17,8573	15,6695	14,5564	13,9760	15,1664	14,0037	14,8720	17,8458	16,3979	16,7476
violin filt	19,7308	18,1106	14,9717	12,6488	13,7885	14,0050	13,4238	14,9443	15,7703	16,0260	16,7369
cello	11,1352	10,5233	8,8571	9,3759	9,4785	10,7559	11,5485	12,3566	12,5777	12,2707	13,1644
cello filt	14,7014	13,8496	12,9280	11,6985	10,2647	11,2751	11,3983	12,8391	13,8399	14,0943	12,8329
accomp.	3,9033	7,1252	8,1884	9,7018	10,0439	9,7973	9,8108	9,7730	8,7884	7,5221	7,0757
accomp. filt	0,1958	3,3072	6,0053	8,1506	9,5255	9,7132	10,0710	9,5178	8,5566	8,3675	8,1532
average	11,6437	11,8353	10,9050	11,2113	11,1661	11,9065	11,7877	12,3339	13,0706	12,0636	12,3292
average filt	11,5426	11,7558	11,3017	10,8326	11,1929	11,6644	11,6310	12,4337	12,7223	12,8292	12,5743

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



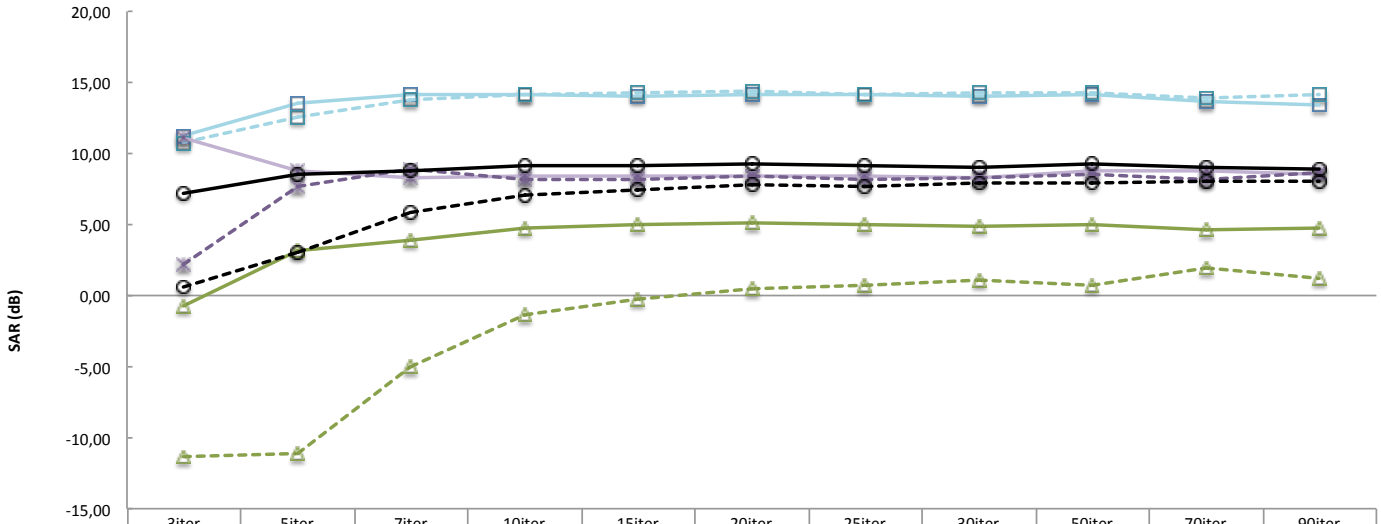
	3iter	5iter	7iter	10iter	15iter	20iter	25iter	30iter	50iter	70iter	90iter
clarinet	19,1733	17,7808	17,7113	17,7810	17,8058	17,8200	17,6388	17,9463	17,9918	18,2758	18,8958
clarinet filt	20,0425	17,1303	17,3280	17,4515	17,5745	17,7260	17,5200	17,6993	18,2950	17,9300	18,0640
viola	0,5795	1,7123	0,9037	0,5618	0,5442	1,1040	0,5997	0,1295	0,5682	0,2348	0,9245
viola filt	3,8174	6,5438	4,3413	1,1360	0,9429	0,5683	0,4299	0,3893	0,6316	0,4730	0,7251
accomp.	4,1717	10,9620	13,7098	13,8138	13,6938	12,8430	13,4053	11,9483	13,0860	10,7913	9,8143
accomp. filt	-0,2579	8,1416	10,1032	12,8974	14,2330	13,6268	14,1110	14,3150	13,9710	12,7970	13,9423
average	7,9748	10,1517	10,7749	10,7188	10,6812	10,5890	10,5479	10,0080	10,5486	9,7673	9,8782
average filt	7,8674	10,6052	10,5908	10,4950	10,9168	10,6403	10,6870	10,8012	10,9659	10,4000	10,9104

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



	3iter	5iter	7iter	10iter	15iter	20iter	25iter	30iter	50iter	70iter	90iter
violin	4,2059	6,1759	7,0721	8,6130	9,1381	9,3796	8,9373	9,0997	9,1673	8,6888	8,6250
violin filt	2,8855	5,4684	6,6413	7,2694	8,1381	8,5800	8,4125	8,5241	8,1058	8,0863	8,5007
cello	6,6255	6,5123	7,5291	8,6758	8,9855	8,8325	8,3874	7,5916	7,6742	7,2595	7,3960
cello filt	0,9754	2,8389	4,9748	6,6243	7,2464	6,9654	7,7047	7,6328	7,3990	7,9781	6,7126
accomp.	17,3832	10,8757	9,6223	9,5758	10,1191	11,2940	11,3169	11,3450	11,0945	12,5687	11,7667
accomp. filt	26,7175	13,1043	10,6884	9,6242	9,9079	10,1214	9,8309	10,6287	10,4454	11,3264	11,2392
average	9,4049	7,8546	8,0745	8,9549	9,4142	9,8354	9,5472	9,3454	9,3120	9,5057	9,2626
average filt	10,1928	7,1372	7,4348	7,8393	8,4308	8,5556	8,6494	8,9285	8,6500	9,1302	8,8175

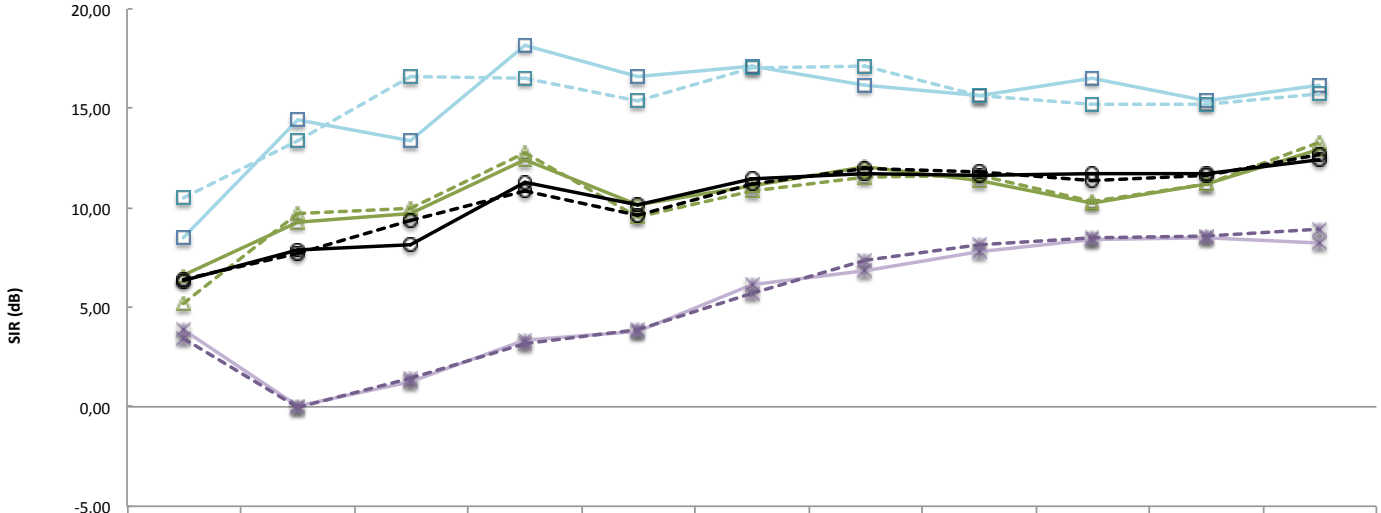
Table B.3: SAR error of the Schubert recording for the Multi-Excitation and Multi-Filter Model with estimated filters, trained filters for different iteration values



	3iter	5iter	7iter	10iter	15iter	20iter	25iter	30iter	50iter	70iter	90iter
clarinet	11,2444	13,5613	14,0863	14,1318	14,0628	14,1373	14,0998	13,9955	14,1543	13,6550	13,4618
clarinet filt	10,7713	12,5543	13,7250	14,1758	14,3253	14,3818	14,1995	14,2720	14,2860	13,8923	14,1968
viola	-0,7345	3,1182	3,9351	4,7457	5,0027	5,0645	5,0194	4,9079	4,9732	4,6150	4,6946
viola filt	-11,3382	-11,0941	-5,0338	-1,3281	-0,2847	0,5026	0,7427	1,0461	0,7530	1,9040	1,1534
accomp.	11,0368	8,7273	8,2431	8,4307	8,3800	8,4118	8,4176	8,2711	8,7650	8,7463	8,5493
accomp. filt	2,1821	7,6346	8,9347	8,2232	8,2136	8,4278	8,1727	8,2763	8,5530	8,1535	8,6597
average	7,1823	8,4689	8,7548	9,1027	9,1485	9,2045	9,1789	9,0582	9,2975	9,0054	8,9019
average filt	0,5384	3,0316	5,8753	7,0236	7,4181	7,7707	7,7050	7,8648	7,8640	7,9832	8,0033

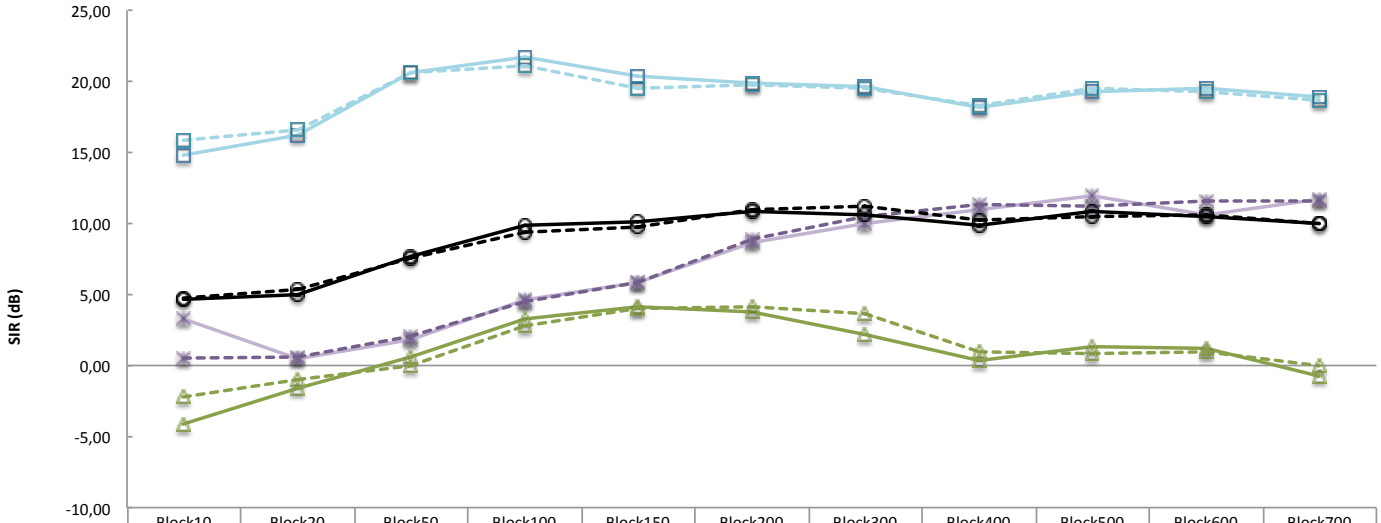
Table B.4: SAR error of the Mozart recording for the Multi-Excitation and Multi-Filter Model with estimated filters and trained filters for different iteration values

SIR and SAR error values of Schubert and Mozart recordings over different *block size* values corresponding to the test presented in 4.3.3.



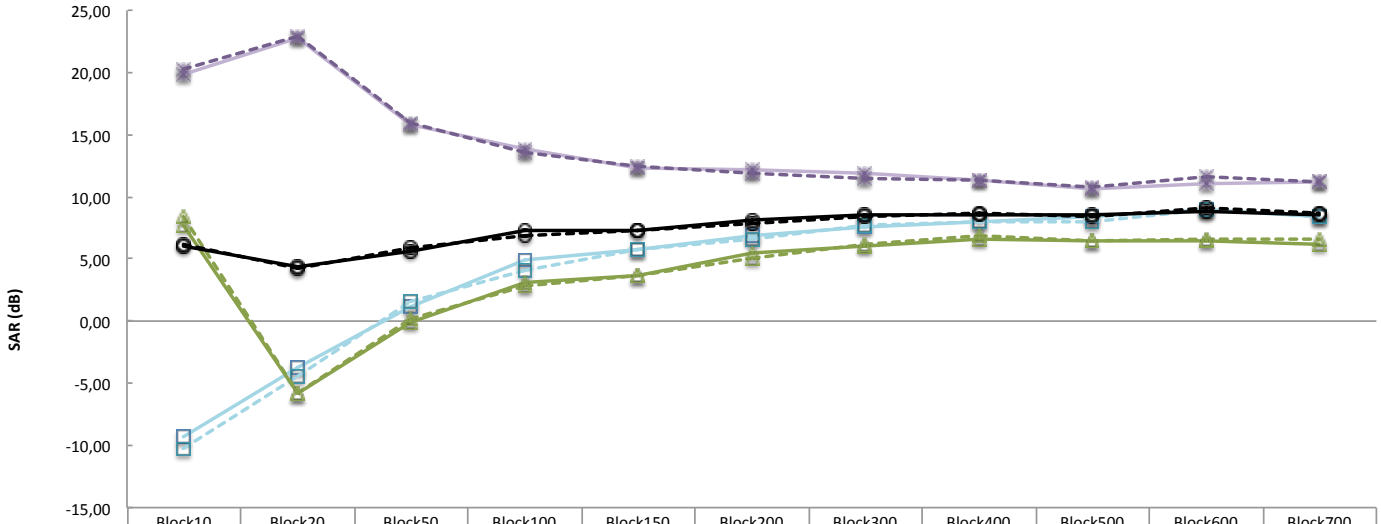
	Block10	Block20	Block50	Block100	Block150	Block200	Block300	Block400	Block500	Block600	Block700
violin	8,50418	14,41708	13,39212	18,18635	16,60117	17,13912	16,16410	15,66495	16,51865	15,36448	16,18713
violin filt	10,48158	13,35117	16,58977	16,49122	15,38913	17,02967	17,16715	15,67672	15,17485	15,18492	15,75893
cello	6,61620	9,29285	9,74283	12,45767	10,14300	11,10080	12,08295	11,40525	10,24067	11,18913	12,89950
cello filt	5,19405	9,70233	9,98312	12,78693	9,55943	10,84342	11,51422	11,62672	10,34410	11,19415	13,31348
accomp.	3,91852	0,00470	1,27750	3,32451	3,78593	6,16138	6,83475	7,81785	8,40350	8,50247	8,19582
accomp. filt	3,48006	-0,02867	1,45225	3,18548	3,89134	5,68663	7,38138	8,17107	8,52703	8,61720	8,94108
average	6,34630	7,90488	8,13748	11,32284	10,17670	11,46710	11,69393	11,62935	11,72094	11,68536	12,42748
average filt	6,38523	7,67494	9,34171	10,82121	9,61330	11,18657	12,02092	11,82483	11,34866	11,66542	12,67117

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



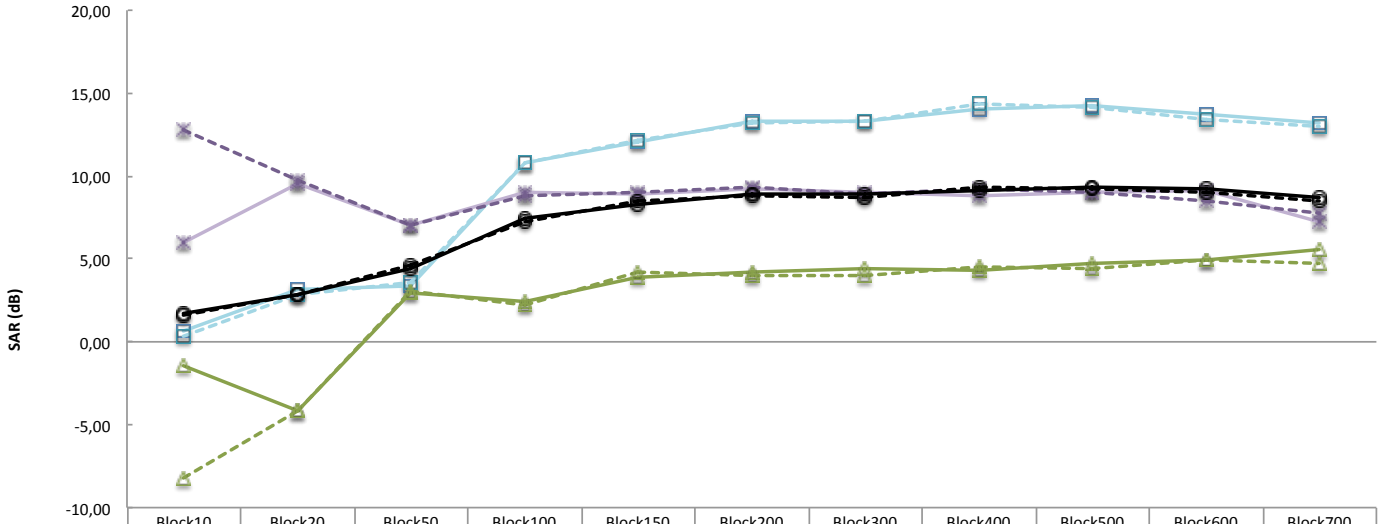
	Block10	Block20	Block50	Block100	Block150	Block200	Block300	Block400	Block500	Block600	Block700
clarinet	14,80223	16,23135	20,54725	21,70400	20,39675	19,87100	19,66500	18,11050	19,30500	19,55775	18,92825
clarinet filt	15,85118	16,54790	20,62175	21,06525	19,53025	19,80550	19,49200	18,31175	19,45625	19,31250	18,63575
viola	-4,11718	-1,57781	0,57600	3,27125	4,11800	3,82120	2,12915	0,39913	1,34183	1,17725	-0,74380
viola filt	-2,23300	-0,98678	0,01005	2,79508	4,01158	4,15400	3,63033	1,00443	0,82605	0,98853	-0,06683
accomp.	3,27935	0,47578	1,83598	4,66670	5,88268	8,68218	10,04010	10,96223	11,93275	10,63268	11,68183
accomp. filt	0,52145	0,64693	2,04993	4,45788	5,81323	8,83670	10,47985	11,30038	11,20950	11,58008	11,54975
average	4,65480	5,04310	7,65308	9,88065	10,13248	10,79146	10,61142	9,82395	10,85986	10,45589	9,95543
average filt	4,71321	5,40268	7,56058	9,43940	9,78502	10,93207	11,20073	10,20552	10,49727	10,62703	10,03956

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



	Block10	Block20	Block50	Block100	Block150	Block200	Block300	Block400	Block500	Block600	Block700
violin	-9,3399	-3,7892	1,1178	4,9459	5,7573	6,8336	7,5788	7,9725	8,3502	8,9449	8,4120
violin filt	-10,2416	-4,4093	1,5493	4,1263	5,7003	6,5230	7,6943	8,0298	7,9447	8,9919	8,3845
cello	7,7488	-5,8113	-0,1253	3,0476	3,6453	5,5112	6,0539	6,5641	6,5061	6,4490	6,1708
cello filt	8,4428	-5,7764	0,1935	2,7886	3,6283	5,0238	6,1908	6,8650	6,4511	6,6174	6,6163
accomp.	19,8548	22,7160	15,7728	13,8962	12,2554	12,2128	11,8379	11,2691	10,6352	11,1161	11,2334
accomp. filt	20,1904	22,9007	15,8808	13,6308	12,4721	11,8770	11,5308	11,3300	10,7317	11,5820	11,2126
average	6,0879	4,3718	5,5884	7,2966	7,2193	8,1859	8,4902	8,6019	8,4972	8,8367	8,6054
average filt	6,1305	4,2383	5,8745	6,8486	7,2669	7,8079	8,4719	8,7416	8,3759	9,0637	8,7378

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



	Block10	Block20	Block50	Block100	Block150	Block200	Block300	Block400	Block500	Block600	Block700
clarinet	0,6415	3,1820	3,3758	10,7590	12,0965	13,2650	13,3420	14,0755	14,2885	13,6735	13,1860
clarinet filt	0,2921	2,8967	3,6111	10,7813	12,1735	13,2475	13,3148	14,3613	14,1518	13,4585	13,0240
viola	-1,4645	-4,2040	2,9746	2,4514	3,8609	4,1698	4,4184	4,3246	4,6900	4,9005	5,5531
viola filt	-8,2175	-4,1657	3,0837	2,1874	4,2463	3,9588	4,0288	4,4898	4,4339	4,9119	4,7489
accomp.	5,9550	9,5905	6,9833	9,0359	8,9241	9,2237	8,9782	8,8526	9,0419	9,0903	7,2424
accomp. filt	12,7508	9,7809	7,0288	8,7825	9,0391	9,3141	8,8694	9,1801	8,9977	8,5330	7,8100
average	1,7107	2,8562	4,4446	7,4154	8,2938	8,8861	8,9128	9,0842	9,3401	9,2214	8,6605
average filt	1,6085	2,8373	4,5745	7,2504	8,4863	8,8401	8,7377	9,3437	9,1945	8,9678	8,5276

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

[Bregman, 1990] Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press.

[Burred, 2009] Burred, J. J. (2009). *From sparse models to timbre learning: new methods for musical source separation*. PhD thesis, Technical University of Berlin.

[Carabias-Orti et al., 2011] Carabias-Orti, J. J., Virtanen, T., Vera-Candeas, P., Ruiz-Reyes, N., and Canadas-Quesada, F. J. (2011). Musical Instrument Sound Multi-Excitation Model for Non-Negative Spectrogram Factorization. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1144–1158.

[Casey, 2000] Casey, M. (2000). Separation of mixed audio sources by independent subspace analysis. *Proceedings of the International Computer Music Conference*, pages 154–161.

[C.Févotte et al., 2005] C.Févotte, Gribonval, R., and E.Vincent (2005). Bss-eval toolbox user guide : Revision 2.0. Technical Report 1706, IRISA.

[Comon, 1994] Comon, P. (1994). Independent component analysis, A new concept? *Signal Processing*, 36(3):287–314.

[Comon and Jutten, 2010] Comon, P. and Jutten, C. (2010). *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 1st edition.

[De Cheveigné and Kawahara, 2002] De Cheveigné, A. and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4):1917.

[Durrieu et al., 2009a] Durrieu, J., Ozerov, A., and Févotte, C. (2009a). Main instrument separation from stereophonic audio signals using a source/filter model. *European Signal Processing Conference (EUSIPCO)*.

[Durrieu et al., 2009b] Durrieu, J. L., Richard, G., and David, B. (2009b). An iterative approach to monaural musical mixture de-soloing.

[Emiya et al., 2011] Emiya, V., Vincent, E., Harlander, N., and Hohmann, V. (2011). Subjective and Objective Quality Assessment of Audio Source Separation.

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

[Fletcher, 1999] Fletcher, N. H. (1999). The nonlinear physics of musical instruments. *Rep. Prog. Phys.*, 62:723–764.

[Fritsch, 2012] Fritsch, J. (2012). High quality musical audio source separation.

[Hahn et al., 2010] Hahn, H., Roebel, A., Burred, J. J., and Weinzierl, S. (2010). Source-filter model for quasi-harmonic instruments. In *Digital Audio Effects (DAFx)*, Graz, Austria.

[Heittola et al., 2009] Heittola, T., Klapuri, A., and Virtanen, T. (2009). Musical instrument recognition in polyphonic audio using source-filter model for sound separation. *Information Retrieval*, 2009(Ismir):327–332.

[Helmholtz, 1860] Helmholtz, H. (1860). On The Action Of The Strings Of A Violin. *Proceedings of the Glasglow Philosophical Society*.

[Hérault et al., 1985] Hérault, J., Jutten, C., and Ans, B. (1985). Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervis. In *Actes du Xème colloque GRETSI*, volume 3, pages 1017–1022.

[Hopgood and Rayner, 2003] Hopgood, J. R. and Rayner, P. J. W. (2003). Single channel nonstationary stochastic signal separation using linear time-varying filters.

[Huber and Kollmeier, 2006] Huber, R. and Kollmeier, B. (2006). PEMO-Qamp;amp;#8212;A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception.

[Jolliffe, 2002] Jolliffe, I. T. (2002). Principal Component Analysis, Second Edition. *Encyclopedia of Statistics in Behavioral Science*, 30(3):487.

[Klatt and Klatt, 1990] Klatt, D. H. and Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2):820–857.

[Lee and Seung, 1999] Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–91.

[Lee and Seung, 2001] Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *In NIPS*, pages 556–562. MIT Press.

[Lee and Seung, 2006] Lee, D. D. and Seung, H. S. (2006). Algorithms for Non-negative Matrix Factorization. *Lecture Notes in Computer Science*, 4029(1):548–562.

[Lefevre et al., 2011] Lefevre, A., Bach, F., and Févotte, C. (2011). Itakura-Saito nonnegative matrix factorization with group sparsity. *IEEE International Conference on Acoustics Speech and Signal Processing*, 31(1):1–4.

[Liutkus et al., 2012] Liutkus, A., Pinel, J., Badeau, R., Girin, L., and Richard, G. (2012). Informed source separation through spectrogram coding and data embedding. *Signal Processing*, 92(8):1937 – 1949.

[Paatero and Tapper, 1994] Paatero, P. and Tapper, U. (1994). Positive matrix factorization - a nonnegative factor model with optimal utilization of error-estimates of data values. *Environmetrics*, 5(2):111–126.

[Rodriguez-Serrano et al., 2012] Rodriguez-Serrano, F. J., Carabias-Orti, J. J., Vera-candeas, P., Virtanen, T., and Ruiz-reyes, N. (2012). Multiple Instrument Source Separation Evaluation Using Instrument-Dependent NMF Models. *10th International Conference, LVA/ICA 2012*, 7191:380–387.

[Schelleng, 1974] Schelleng, J. C. (1974). The physics of the bowed string. *j-SCI-AMER*, 230(1):87–95.

[Schmidt, 2008] Schmidt, M. N. (2008). *Single-channel source separation using non-negative matrix factorization*. PhD thesis, Technical University of Denmark.

[Vincent et al., 2003] Vincent, E., Févotte, C., Gribonval, R., Benaroya, L., Rodet, X., Röbel, A., Le Carpentier, E., and Bimbot, F. (2003). A tentative typology of audio source separation tasks. In *4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA)*, pages 715–720, Nara, Japon.

[Vincent et al., 2007] Vincent, E., Sawada, H., Bofill, P., Makino, S., and Rosca, J. P. (2007). First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results. In *7th Int Conf on Independent Component Analysis and Blind Source Separation ICA*, pages 552–559. Springer.

[Virtanen et al., 2008] Virtanen, T., Cemgil, A. T., and Godsill, S. (2008). Bayesian extensions to non-negative matrix factorisation for audio signal modelling.

[Virtanen and Klapuri, 2006a] Virtanen, T. and Klapuri, A. (2006a). Analysis of polyphonic audio using source-filter model and non-negative matrix factorization.

[Virtanen and Klapuri, 2006b] Virtanen, T. and Klapuri, A. (2006b). Analysis of polyphonic audio using source-filter model and non-negative matrix factorization.

[Wang and Brown, 2006] Wang, D. and Brown, G. J. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press.

[Woodhouse, 2004] Woodhouse, J. (2004). On the Synthesis of Guitar Plucks. *Engineering*, 90(5):928 – 944.