TECHNISCHE UNIVERSITÄT ILMENAU

Fakultät für Elektrotechnik und Informationstechnik
der Technischen Universität Ilmenau

Pitch-Informed Solo and Accompaniment Separation

Estefanía Cano Cerón

Dissertation zur Erlangung des
akademischen Grades Doktor-Ingenieur (Dr.-Ing)

urn:nbn:de:gbv:ilm1-2014000351

Anfertigung im:	Fachgebiet Elektronische Medientechnik Institut für Medientechnik Fakultät für Elektrotechnik und Informationstechnik
Gutachter:	Prof. Dr.-Ing Dr. rer. nat. h.c. mult. Karlheinz Brandenburg Dr. Derry Fitzgerald Prof. Dr.-Ing Gerald Schuller
Vorgelegt am:	15.12.2013
Verteidigt am:	27.10.2014

Acknowledgements

I would like to thank my colleagues and friends at the Fraunhofer Institute for Digital Media Technology IDMT for all the hard work and good times during these years of my PhD. I would also like to thank Prof. Gerald Schuller for his supervision, and Prof. Mark Plumbley for a very much enjoyed research stay at the Center for Digital Music C4DM, at Queen Mary University of London.

Special thanks to the big boss Christian Dittmar and to my partner in crime Sascha Grollmisch for many hours of hard work, for all the good times with Songs2See and freaky Feierabend. I would also like to thank my friend and favorite proof-reader Jakob Abeßer for all his help through all these years of papers, articles, and project proposals.

Most importantly, I would like to thank my family as none of this would have been possible without them.

Abstract

This thesis addresses the development of a system for pitch-informed solo and accompaniment separation capable of separating main instruments from music accompaniment regardless of the musical genre of the track, or type of music accompaniment. For the solo instrument, only pitched monophonic instruments were considered in a single-channel scenario where no panning or spatial location information is available.

In the proposed method, pitch information is used as an initial stage of a sinusoidal modeling approach that attempts to estimate the spectral information of the solo instrument from a given audio mixture. Instead of estimating the solo instrument on a frame by frame basis, the proposed method gathers information of tone objects to perform separation. Tone-based processing allowed the inclusion of novel processing stages for attack refinement, transient interference reduction, common amplitude modulation (CAM) of tone objects, and for better estimation of non-harmonic elements that can occur in musical instrument tones. The proposed solo and accompaniment algorithm is an efficient method suitable for real-world applications.

A study was conducted to better model magnitude, frequency, and phase of isolated musical instrument tones. As a result of this study, temporal envelope smoothness, inharmonicty of musical instruments, and phase expectation were exploited in the proposed separation method. Additionally, an algorithm for harmonic/percussive separation based on phase expectation was proposed. The algorithm shows improved perceptual quality with respect to state-of-the-art methods for harmonic/percussive separation.

The proposed solo and accompaniment method obtained perceptual quality scores comparable to other state-of-the-art algorithms under the SiSEC 2011 and SiSEC

2013 campaigns, and outperformed the comparison algorithm on the instrumental dataset described in this thesis.

As a use-case of solo and accompaniment separation, a listening test procedure was conducted to assess separation quality requirements in the context of music education. Results from the listening test showed that solo and accompaniment tracks should be optimized differently to suit quality requirements of music education. The Songs2See application was presented as a commercial music learning software which includes the proposed solo and accompaniment separation method.

Zusammenfassung

Das Thema dieser Dissertation ist die Entwicklung eines Systems zur Tonhöhen-informierten Quellentrennung von Musiksignalen in Soloinstrument und Begleitung. Dieses ist geeignet, die dominanten Instrumente aus einem Musikstück zu isolieren, unabhängig von der Art des Instruments, der Begleitung und Stilrichtung. Dabei werden nur einstimmige Melodieinstrumente in Betracht gezogen. Die Musikaufnahmen liegen monaural vor, es kann also keine zusätzliche Information aus der Verteilung der Instrumente im Stereo-Panorama gewonnen werden.

Die entwickelte Methode nutzt Tonhöhen-Information als Basis für eine sinusoidale Modellierung der spektralen Eigenschaften des Soloinstruments aus dem Musikmischsignal. Anstatt die spektralen Informationen pro Frame zu bestimmen, werden in der vorgeschlagenen Methode Tonobjekte für die Separation genutzt. Tonobjekt-basierte Verarbeitung ermöglicht es, zusätzlich die Notenanfänge zu Verfeinern, transiente Artefakte zu reduzieren, gemeinsame Amplitudenmodulation (Common Amplitude Modulation CAM) einzubeziehen und besser nichtharmonische Elemente der Töne abzuschätzen. Der vorgestellte Algorithmus zur Quellentrennung von Soloinstrument und Begleitung ermöglicht eine Echtzeitverarbeitung und ist somit relevant für den praktischen Einsatz.

Ein Experiment zur besseren Modellierung der Zusammenhänge zwischen Magnitude, Phase und Feinfrequenz von isolierten Instrumententönen wurde durchgeführt. Als Ergebnis konnte die Kontinuität der zeitlichen Einhüllenden, die Inharmonizität bestimmter Musikinstrumente und die Auswertung des Phasenfortschritts für die vorgestellte Methode ausgenutzt werden. Zusätzlich wurde ein Algorithmus für die

Quellentrennung in perkussive und harmonische Signalanteile auf Basis des Phasenfortschritts entwickelt. Dieser erreicht ein verbesserte perzeptuelle Qualität der harmonischen und perkussiven Signale gegenüber vergleichbaren Methoden nach dem Stand der Technik.

Die vorgestellte Methode zur Klangquellentrennung in Soloinstrument und Begleitung wurde zu den Evaluationskampagnen SiSEC 2011 und SiSEC 2013 eingereicht. Dort konnten vergleichbare Ergebnisse im Hinblick auf perzeptuelle Bewertungsmaße erzielt werden. Die Qualität eines Referenzalgorithmus im Hinblick auf den in dieser Dissertation beschriebenen Instrumentaldatensatz übertroffen werden.

Als ein Anwendungsszenario für die Klangquellentrennung in Solo und Begleitung wurde ein Hörtest durchgeführt, der die Qualitätsanforderungen an Quellentrennung im Kontext von Musiklernsoftware bewerten sollte. Die Ergebnisse dieses Hörtests zeigen, dass die Solo- und Begleitspur gemäß unterschiedlicher Qualitätskriterien getrennt werden sollten. Die Musiklernsoftware Songs2See integriert die vorgestellte Klangquellentrennung bereits in einer kommerziell erhältlichen Anwendung.

Contents

1. Introduction

1.1. Motivation and Scope

What is sound source separation?

Sound source separation is the signal processing task that attempts to recover unknown signals or *sources* from an audio *mixture* by computational means. In the case of musical signals, a possible sound source separation task would be to obtain independent signals for the saxophone, piano, bass, and percussion, given a recording or *audio mixture* of a jazz quartet. In a more general scenario, a sound separation problem could be to recover the original signals of a car passing by, a male speaker, and a dog barking, given a sound field recording of a certain street intersection.

Why is it relevant to solve the sound source separation problem?

Many applications benefit from robust sound source separation approaches. In most cases however, more than being the final goal, source separation appears as an intermediate step to allow and improve complex types of content analysis. This is the case, for example, of automatic music transcription, up-mixing to multi-channel formats, parametric audio coding, musicological analysis, automatic music classification, search, and recommendation.

Automatic music transcription (AMT), which deals with the extraction of melodic, rhythmic, and harmonic parameters of music recordings to allow its representation as a musical score, could greatly benefit from having independent sound sources. After many years of work in the Music Information Retrieval (MIR) community, automatic music transcription is still considered today, an unsolved problem. The main difficulty of this task lies in the complexity of extracting meaningful rhythmic, melodic and harmonic information, when the sound sources greatly overlap in the spectral domain [9]. Ideally, a system for sound source separation would extract independent signals for each one of the sources in the mixture. The transcription problem would then be limited to monophonic signals where parameter extraction is much more robust.

In the music production field, sound separation allows up-mixing of monaural and stereo recordings to multi-channel formats. Numerous recordings are only available as monaural (single-channel) tracks. Starting in the 1960s, stereo recording became popular and has been the standard for audio distribution since then. However, due to developments in audio systems, increased capacity of handling information, and the popularity of immersive sound, multi-channel formats are becoming more and more common. Sound separation gives the possibility to create new mixes in new audio formats from any type of existing recording [10].

More advanced tools for musicological analysis can be achieved if separation of sound sources is available. Style, melody, genre and artistic elements in music could be more thoroughly studied and characterized. Similarly, tools for automatic music classification, search and recommendation would become more powerful as more detailed information from the different sources could be extracted [O1].

Audio coding also benefits from sound separation as new coding schemes, that reduce the amount of information needed to characterize the different channels or instruments, can be developed [11]. The capability of separating audio sources in the decoder's side by only transmitting a series of signal parameters is very powerful.

In the particular case of this work, the motivation to work on sound separation is two-fold: on the one hand, this work aims at improving quality of sound separation through better understanding and characterization of musical instrument signals and their time-frequency characteristics. On the other hand, this works aims at exploring the potential of sound separation technologies in music education applications. Even when music education has been an extremely active field for many decades, its methods and tools still remain very traditional. Only in the past few years, music technologies have started to reach the music education context. This work attempts to further push these boundaries and study the usability of sound separation technologies for content and practice material creation.

Which specific separation task does this work attempt to solve?

The work presented in this thesis focuses on a particular case of sound source separation called *solo and accompaniment separation*, also referred to as *lead or main instrument separation* or *de-soloing*. For this specific task, the goal is to separate the audio mix into two sources only: the main instrument or solo, and the accompaniment. The accompaniment refers to one or more instruments playing along with the solo. In this thesis, the terms accompaniment tracks and *backing tracks* are used interchangeably, both referring to extracted musical accompaniment. The solo is assumed to be the instrument playing the *main melody* of the piece. The Music Information Retrieval (MIR) community commonly refers to main melody as the single (monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of polyphonic music, and that a listener would recognize as being the essence of that music [12].

Results of many years of sound separation research suggest that separation performance can be improved when prior information about the sources is available. The inclusion of known information about the sources in the separation scheme is referred to as *Informed Sound Source Separation (ISS)* and comprises, among others, the use of MIDI-like musical scores, the use of pitch tracks of one or several sources, oracle sound separation where the original sources are available, and the extraction of model parameters from training data of a particular sound source [13]. The work described in this thesis focuses on solving the solo and accompaniment separation problem in polyphonic music using pitch as prior information. This approach is referred to as *pitch-informed solo/accompaniment separation*.

> ***Problem Definition and Scope***
>
> The goal of this work is the development of a system for pitch-informed solo and accompaniment separation capable of separating main instruments from music accompaniment regardless of the musical genre of the track, or type of music accompaniment. For the solo instrument, only *pitched instruments* are

considered and no attempt is made to separate percussive instruments. This work focuses on the *monophonic* case, where the solo instrument is assumed to play only one note at a time. Only the *single-channel* separation problem is considered and no panning or spatial location information is used for separation. The algorithm should be lightweight and processing times should be minimized to allow its use in real-world applications.

1.2. Thesis structure

Figure 1.1 shows a block diagram of the structure of this thesis. The contents of this work are distributed in five chapters:

Chapter 1 presents a general introduction to the work presented in this thesis, clearly presenting the problem definition and scope.

In Chapter 2, the relevant theoretical background and a general survey of state-of-the-art approaches in sound source separation is presented. As they are the core foundations of this thesis, special attention was given to informed sound source separation approaches and to the characterization of the spectral parameters of musical instruments (magnitude, frequency, and phase).

In Chapter 3, the studies conducted as part of this research and the proposed methods for pitch-informed sound separation are described. Three main sections compose this chapter. Section 3.1 presents the proposed method for frame-based pitch-informed solo and accompaniment separation. The method is described and its performance is evaluated and put into context of state-of-the-art approaches. In Section 3.2, different studies conducted to better describe the characteristics of spectral parameters (magnitude, frequency and phase) of musical instrument and their contributions to the quality of audio signals, are presented. Section 3.3 collects the findings from the two previous sections to propose a pitch-informed tone-based solo and accompaniment separation approach. Being a fundamental component of the proposed methods for solo and accompaniment separation, an evaluation of the

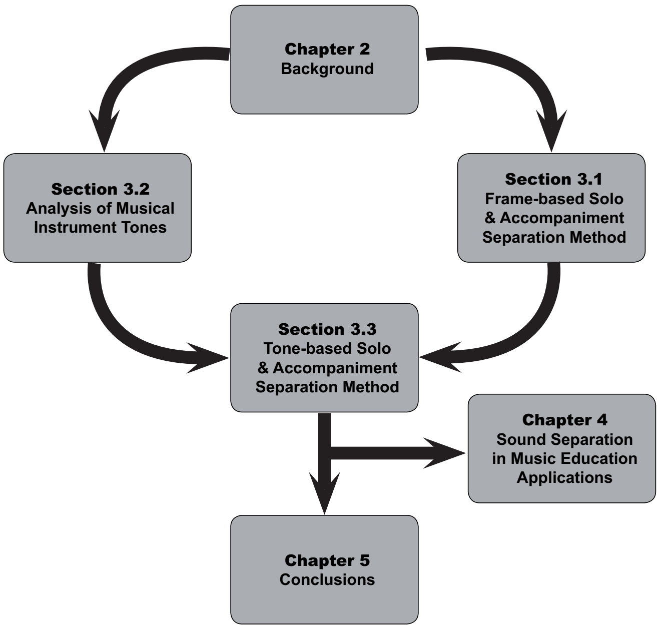


Figure 1.1.: Structure of the thesis

performance of pitch detection algorithms within the separation context is also presented in Section 3.3.8.1. The performance of the tone-based separation algorithm is evaluated and placed into context of state-of-the-art approaches.

In Chapter 4, the use of solo and accompaniment separation technologies in music education applications is described as a case study. Quality requirements posed by music education applications on separation methods are studied through a listening test procedure. Songs2See is described as a commercial application that incorporates the separation method described in this thesis for content creation in a music practice scenario.

Finally, Chapter 5 draws some conclusions and depicts future directions for this research.

1.2.1. Format and conventions

Throughout this thesis, citations will be presented using bracket notation [x]. In the cases where own publications are cited in the text, the convention [Ox] will be used, where O stands for own and x is the publication number. All the cited publications are listed in the Reference section of this thesis.

In the different chapters, terms that are not defined directly in the text but that are presented in the Glossary for reference are shown in blue. In all the figures where several observations are simultaneously displayed, the *Categorical Color Scheme in 12 steps* is used as a color scheme. This color scheme was designed for easier the visualization of scientific data [14].

1.3. Associated Publications

This thesis covers work conducted by the author as a research assistant and PhD student at the Fraunhofer Institute for Digital Media Technology from January 2010 until September 2013. The majority of the work presented in this thesis has been presented in international peer-reviewed conferences and journals. Additionally, it should be noted that parts of this work have been linked to industry-related projects:

1. Songs2See Project: The goal of this project was the development of a music learning application to be used with real musical instruments that could offer real-time performance feedback to the user. The separation algorithm presented in Section 3.1 is included in this application as a tool for content creation. This project was conducted from 2010-2012 at the Fraunhofer Institute for Digital Media Technology IDMT in collaboration with the following academic and industry partners: Tampere University, Grieg Music, Kids Interactive, and Sweets for Brains. The project website can be accessed in [15] and a brief description of the application is provided in Section 4.3 of this thesis.

   Several publications resulted from the involvement in the Songs2See project and several sections of this thesis refer to the results presented in them, specially Chapter 4 which refers to the use of sound separation research in Music Education applications.

The following list of publications is related to the work presented in this thesis:

Book Chapters

[i] Dittmar, C., Cano, E., Grollmisch, S., Abeßer, J. & Maennchen A. Music Technology and Music Education. To appear in Springer Handbook on Systematic Musicology, (Springer Berlin Heidelberg, 2014).

[ii] Dittmar, C., Cano, E., Abeßer, J. & Grollmisch, S. Music Information Retrieval Meets Music Education. In Müller, M., Goto, M. & Schedl, M. (eds.) Multimodal Music Processing, chap. Music Info, 1-24 (Dagstuhl Publishing, 2012). ISBN 978-3-939897-37-8

[iii] Dittmar, C., Grossman, H., Cano, E., Grollmisch, S., Lukashevich, H., Abeßer, J. Songs2see and Globalmusic2one: Two Applied Research Projects in Music Information Retrieval at Fraunhofer IDMT. In Ystad, S., Aramaki, M., Kronland-Martinet, R. & Jensen, K. (eds.) Exploring Music Contents, vol. 6684 of Lecture Notes in Computer Science, 259-272 (Springer Berlin Heidelberg, 2011).

Journal Papers

[vi] Cano, E., Schuller, G. & Dittmar, C. Pitch-Informed Solo and Accompaniment Separation: Towards its use in Music Education Applications. In EURASIP Journal on Advances on Signal Processing, special issue on Informed Source Separation. 2013 (Submitted)

Peer-reviewed Conference Papers

[v] Weiss, Christof, Cano, E. & Lukashevich Hanna. A Mid-level Approach to Local Tonality Analysis: Extracting Key Signatures from Audio. To appear in AES 53rd International Conference on Semantic Audio, (London, UK, 2014).

[vi] Cano, E., Dittmar, C. & Schuller, G. Re-thinking Sound Separation: Prior Information and Additivity Constraint in Separation Algorithms. In 16th International Conference on Digital Audio effects (DAFx-13), 1-7 (Maynooth, Ireland, 2013).

[vii] Krasser, J., Abeßer, J., Grossmann, H., Dittmar, C. & Cano, E. Improved Music Similarity Computation based on Tone Objects Categories. In Audio Mostly Conference, 47-54 (Corfu, Greece, 2012).

[viii] Cano, E., Dittmar, C. & Schuller, G. Efficient Implementation of a System for Solo and Accompaniment Separation in Polyphonic Music. In 20th European Signal Processing Conference (EUSIPCO 2012), Eusipco, 285-289 (Bucharest, Romania, 2012).

[ix] Cano, E., Grollmisch, S. & Dittmar, C. Songs2See: Towards a New Generation of Music Performance Games. In 9th International Symposium on Computer Music Modeling and Retrieval CMMR, June, 19-22 (London, UK, 2012).

[x] Cano, E., Dittmar, C. & Schuller, G. Influence of Phase, Magnitude, and Location of Harmonic Components in the Perceived Quality of Extracted Solo Signals. In AES 42nd International Conference on Semantic Audio, 1-6 (Ilmenau, Germany, 2011).

[xi] Grollmisch, S., Dittmar, C., Cano, E. & Dressler, K. Server Based Pitch Detection for Web Applications. In AES 41st International Conference on Audio for Games, 1-5 (London, UK, 2011).

[xii] Grollmisch, S., Cano, E. & Dittmar, C. Songs2See: Learn to Play by Playing. In AES 41st International Conference on Audio for Games, 2-7 (London, UK, 2011).

[xiii] Cano, E., Dittmar, C. & Grollmisch, S. Acoustics and Signal Processing in the Development of Music Education Software. In Proceeding of the 2nd Vienna Talk, 19-22 (Vienna, Austria, 2010).

[xiv] Cano, E., Schuller, G. & Dittmar, C. Exploring Phase Information in Sound Source Separation Applications. In 13th International Conference on Digital Audio Effects (DAFx-10), 1-8 (Graz, Austria, 2010).

[xv] Cano, E. & Cheng, C. Melody Line Detection and Source Separation in Classical Saxophone Recordings. In 12th International Conference on Digital Audio Effects (DAFx-09), 1-6 (Como, Italy, 2009).

Other Publications

[xvi] Cano, E., Dittmar, C.& Grollmisch, S. Songs2See: Learn to Play by Playing. In 12th International Society for Music Information Retrieval Conference: Late-breaking Demo (ISMIR 2011) (Miami, USA, 2011).

In [vii], the frame-based separation algorithm presented in Section 3.1 of this thesis was used to improve music similarity estimation by extracting relevant features directly on tone objects. The proposed method was tested in a 5 class genre classification system using Mel Frequency Cepstral Coefficients (MFCC) and Octave-based Spectral Contrast (OSC) features and showed improved performance compared to a baseline system.

2. Background

As described in the introductory sections, sound source separation is the signal processing task that attempts to extract unknown sound sources from a given audio mixture. Attempting to classify separation approaches into distinct categories can be a difficult task: as systems become more complex, many of the distinctions between categories have become blurred and a clear classification border between them can no longer be drawn. An important categorization of separation systems that remains valid refers to the proportion between sounds sources and channels in the problem to be addressed. When the number of channels available is larger than the number of sources, the separation task is named *overdetermined.* Similarly, if the number of channels is the same as the number of sound sources in the mixture, the problem is referred to as *determined.* The opposite case appears when less channels than sources are available and the separation task becomes *underdetermined.* This last case poses greater difficulties than the previous two, as with a reduced number of channels, less information from the sources is available and the mathematical means to solve the problem greatly rely on strong assumptions about the sources.

Another distinction between separation algorithms that is still very frequently used refers to the available knowledge of the sources to be separated. A source separation approach is said to be blind if little or no knowledge about the sources is available. This type of separation is referred to as *Blind Source Separation (BSS).* It is important to mention that no separation method is completely blind as at least some probabilistic assumptions have to be made to address the task. In contrast, those separation approaches that make use of available high-level information about the sources are categorized as *Informed Source Separation (ISS).*

Being the core of this work, this section focuses on underdetermined separation methods where prior information of the sources is accessible. Here, the separation process is described in different stages depicted in Figure 2.1:

(1) *Source Parameter Estimation:* before any separation can be performed, all methods need to estimate the parameters corresponding to the desired source. Depending on the method used, different parameters might be required: magnitude envelopes, frequency locations of harmonic components, activation coefficients, etc. The estimation stage often makes use of *prior information* about the sources to guide and make the estimation more robust.

(2) *Separation Procedure:* After having estimated the source parameters, this stage refers to the actual separation of the spectral content from the different sources.

The stages of the separation process are further explained in the following sections. The separation procedure is explained in Section 2.1, different methods used to estimate source parameters are described in Section 2.2, and finally, types of prior information frequently used in separation tasks are described in Section 2.3.

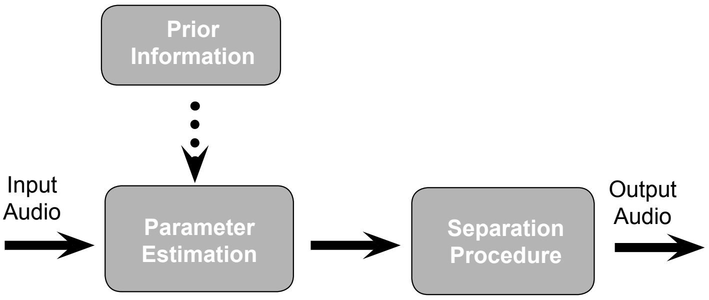


Figure 2.1.: Block diagram of a separation process where prior information about the sources is available.

2.1. Separation Stage

As shown in Figure 2.1, the separation procedure in a typical sound source separation method is only performed after a series of parameters have been estimated and have defined the elements in the time-frequency representation that should be separated. Here, the separation procedure is explained first, as it is usually the simplest stage of the processing chain.

Wiener filtering has been the most commonly used separation procedure in the last few years. To exploit the short-term stationarity of audio signals, Wiener filtering is most commonly applied on time-frequency representations such as the Short Time Fourier Transform (STFT). In the following section, the multi-channel generalized Wiener filter is first introduced. Being probably the most used procedure so far, the particular case of time-frequency masking is also described. Finally, the concept of binary masking is introduced.

2.1.1. Wiener Filtering

Given a multi-channel system with $j = 1, \ldots, J$ sources and $i = 1, \ldots, I$ channels, the STFT coefficients $S_j(f, n)$ of each source are modeled as statistically independent Gaussian random variables with variance $v_j(f, n)$ [16]. Furthermore, let the relationship between each source j and its image in channel i be given by the spatial covariance matrix $\boldsymbol{R}_j(f)$ of size $I \times I$ [13]. Namely, for each source j, the matrix $\boldsymbol{R}_j(f)$ describes its contribution in each of the $i = 1, \ldots, I$ channels. The spatial image $\hat{\boldsymbol{y}}_{j_0}(f, n)$ of source j_0 is given in the Minimum Mean-Square Error (MMSE) sense by [17]:

$$\hat{\boldsymbol{y}}_{j_0}(f, n) = \frac{v_{j_0}(f, n)\boldsymbol{R}_{j_0}(f)}{\sum_{j=1}^{J} v_j(f, n)\boldsymbol{R}_j(f)} \, \boldsymbol{x}(f, n) \tag{2.1}$$

with $\boldsymbol{x}(f, n)$ an $I \times 1$ vector holding the STFT coefficients of the observed mixture.

In a single-channel separation scenario, the generalized Wiener filter becomes the commonly used time-frequency masking approach. The classical Wiener Filter estimate for $\hat{S}_j(f,n)$ is then given by [18]:

$$\hat{S}_j(f,n) = \frac{v_j(f,n)}{\sum_{j=1}^{J} v_j(f,n)}\, x(f,n) \tag{2.2}$$

2.1.2. Binary Masking

In many applications, efficiency of computation plays an important role. In these cases, Wiener filtering approaches might not be suitable as its calculation requires the estimation of the STFT coefficients $\hat{S}_j$ of all sources and possibly large matrix divisions before separation. Binary masks represent a simple alternative with a good trade-off between efficiency and separation performance.

Under the assumption that the observed mixture $x[n]$ is the sum of a target source $s_1[n]$ and an interference source $s_2[n]$:

$$x[n] = s_1[n] + s_2[n] \tag{2.3}$$

with STFT coefficients given by $X(f,n) = S_1(f,n) + S_2(f,n)$, the Ideal Binary Masks (IBM) is defined as [19]:

$$M(f,n) = \begin{cases} 1 & \text{if } |S_1(f,n)|^2 > |S_2(f,n)|^2 \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

2.1.3. Divergence-based Masks

As mentioned in the Introduction, Wiener masking has been the most commonly used masking approach in separation research. While being a relatively simple approach, Wiener masking can result in good separation quality. However, while Wiener masks are optimal in a least-square sense, this does not necessarily mean that the masks are optimal from a perceptual point of view. With this in mind, [20] proposes the use of divergence-based masks for separation. The choice of divergence-based masks was made due to the fact that in Non-negative Matrix Factorization (NMF)-based separation algorithms, least-squares approximations have been outperformed by divergence-based cost functions. In their study, the authors compare quality of separation when Wiener masking and two types of divergence masks — Itakura-Saito and Kullback-Leibler divergence— are used. Results from this study show that with three different separation algorithms, the use of divergence masks outperforms Generalized Wiener filtering in terms of overall perceptual quality. The divergence-based masks are defined as follows [20]:

$$M(f,n) = 1 - \frac{D(S_k(f,n), Q(f,n))^t}{(J-1)\sum_{j=1}^{J} D(S_j(f,n), Q(f,n))^t} \tag{2.5}$$

where $Q(f,n)$ is the spectrogram of the original mixture, $S_k(f,n)$ is the estimated spectrogram of the target source k, t is a parameter used to vary the characteristics of the mask, J the total number of sources, and D denotes a suitable diverge metric. The Kullback-Leibler and Itakuro-Saito divergences are given by Equations (2.17) and (2.18), respectively, and are further explained in Section 2.2.3.

2.2. Source Parameters Estimation

In most sound separation methods, a series of signal parameters are first estimated given certain assumptions made about the sources. The role that the parameter estimation stage plays in a source separation method is depicted in Figure 2.1. In the following sections, several models and methods are introduced which have been applied in a sound separation scenario. Even if their formal formulation does not explicitly conceive them as such, their main function in a separation context is either to facilitate parameter extraction or to directly extract source parameters that will guide the final separation.

2.2.1. Signal Models

In general, music information retrieval tasks attempt to identify perceived signal characteristics as the pitch, duration, timbre, etc. It is then desirable to have modeling approaches that allow easy control of these high-level categories of the physical domain. One modeling alternative that provides such a link with the physical world is the use of physical models of the sound sources, e.g., the physical model of a trumpet. In practice, physical models are not frequently used in Music Information Retrieval (MIR) as they tend to be complex, still require extensive research, and learning them automatically from data is difficult. Furthermore, physical models are often very specific, making it difficult to develop general solutions with them [21]. A second alternative and a more frequently used one is the use of signals models. The idea here is not to model the sound source that produces the signal, but the resulting signal itself. Almost all signal processing tasks rely on the assumption that the signal $x[n]$ can be modeled as a weighted sum of a set of expansion functions [22]. This assumption is referred to as *linear signal model* and is described as follows:

$$\hat{x}[n] = \sum_{m=1}^{M} g_m b_m[n] \tag{2.6}$$

where M is the number of expansion functions, g_m are the expansion coefficients, and $b_m[n]$ are the expansion functions. Two special cases of the linear signal model model in (2.6) have become very widespread in MIR research as they are especially successful in establishing the link between the processing and physical worlds: *Sinusoidal Model* and *Source/Filter Model*.

2.2.1.1. Sinusoidal Model

The sinusoidal signal model decomposes the signal as a sum of sinusoids with varying frequencies and amplitudes. The main idea behind this model is to represent the individual vibration modes of the excitation source [23]. The linear signal model in (2.6) becomes:

$$\hat{x}[n] = \sum_{m=1}^{M[n]} g_m[n] cos\phi_m[n] \tag{2.7}$$

where $g_m[n]$ and $\phi_m[n]$ represent the amplitude variation and total phase of the *m-th* sinusoidal partial. $M[n]$ is the total number of sinusoids which may also vary in time.

2.2.1.2. Source/Filter Model

The linear signal model described in (2.6) can also be used to model the magnitude spectrogram $x_n[k]$ of a signal $x[n]$, with k the frequency index, and n the time frame. The magnitude spectrogram is then modeled as a weighted sum of expansion functions:

$$\hat{x}_n[k] = \sum_{m=1}^{M} g_{m,n} b_m[k] \tag{2.8}$$

where $g_{m,n}$ is the gain of the expansion function m in time frame n, and $b_m[k]$ the expansion functions with $m = 1, \ldots, M$. In the context of polyphonic signals, the expansion functions in (2.8) represent the magnitude spectra of the different musical instrument tones. In this model, $\hat{x}_n[k]$ is represented as a sum of fixed spectra. This implies that a distinct excitation function is required for each pitch value of each musical instrument. The source/filter model builds upon this definition and further models the expansion functions $b_m[k]$ as a product of the magnitude spectra of an excitation source $e_{i,n}[k]$, and a resonator structure or filter $h_j[k]$ [24] .

$$\hat{x}_n[k] = \sum_{i=1}^{I} \sum_{j=1}^{J} g_{i,j,n} e_{i,n}[k] h_j[k] \tag{2.9}$$

In this model, the source represents a vibrating object such as a violin string, and the filter represents the resonance structure of the instrument. In the polyphonic signal context, the excitations sources $e_{i,n}[k]$ correspond to pitch values of individual notes $i = 1, \ldots, I$ at time frame n. The filters $h_j[k]$ correspond to the spectral shapes of the different musical instruments $j = 1, \ldots, J$. One of the main advantages of the source/filter model described in (2.9) over the initial linear model in (2.8) is that the number of parameters to estimate is highly reduced as it only assigns one excitation per pitch and one filter per instrument.

In a separation context, source/filter models are often used to model solo instruments and the voice. The system described in [25] proposes a mid-level representation of the audio signal assuming an instantaneous mixture model (IMM) of the target source and the residual. The solo instrument is represented with a source/filter model where the source carries pitch information, and the filter timbral information. Non-negative-matrix factorization (NMF) and soft masking are used to separate the solo instrument from its accompaniment.

2.2.2. Signal Sparsity

The linear signal model in (2.6) can be represented in matrix form as follows:

$$\hat{x} = \hat{g}\boldsymbol{B} \tag{2.10}$$

where $\boldsymbol{B}$ is a matrix of expansion functions and $\boldsymbol{B}_{m,n} = b_m[n]$. A signal $x[n]$ is said to be sparse if most of its expansion coefficients g_m are zero or close to zero, or equivalently, if only a small number of coefficients in $\hat{g}$ are non-zero. Sparse representations have proven to be very powerful in the analysis of audio signals [26]. In the context of sound source separation, signal sparsity is a desired characteristic as it implies that a given sound source can be described with a small number of coefficients. Furthermore, assuming signal sparsity in a separation problem implies that the degree of overlapping between sources is small, and if the separation scheme succeeds in detecting the coefficients that correspond to each one of the sources in the mixture, the amount of unwanted interference in the resulting separated sources will be minimal. However, solving (2.10) to recover a sparse representation needs to be handled with care. In systems where the number of basis functions M is larger than the number of signal samples N, that is $M > N$, the matrix $\boldsymbol{B}$ is rectangular and (2.10) cannot be solved by simple matrix inversion. A common approach in these cases is the use of the *Moore-Penrose pseudoinverse* $\boldsymbol{B}^{\dagger}$ such that $\hat{g} = \hat{x}\boldsymbol{B}^{\dagger}$. This solution however, is not guaranteed to be sparse. A sparse representation can be obtained for example, by using the *Basis Pursuit* relaxation [26]:

$$\underset{\hat{g}}{argmin}\{\|\hat{g}\|_1 \,|\hat{x} = \hat{g}\boldsymbol{B}\} \tag{2.11}$$

where 1-norm $||\hat{g}||_1 = \sum_{m=1}^{M} |\hat{g}|$ is the sum of absolute values of the expansion coefficients.

Another possibility to retrieve sparse representations is the use of *greedy algorithms* such as *Matching Pursuits (MP)* and *Orthogonal Matching Pursuits (OMP)* to find an approximation to:

$$\underset{\hat{g}}{argmin}\{\|\hat{g}\|_0 \,|\hat{x} = \hat{g}\boldsymbol{B}\} \tag{2.12}$$

where 0-norm $||\hat{g}||_0$ is the number of non-zero elements in $\hat{g}$. A comparative study of different greedy algorithms is presented in [27].

Signal *sparsity* for example, has been exploited in several approaches. In [28] a system for singing voice separation is proposed where the singing voice is modeled as a high-rank but *sparse* signal in the time-frequency domain. The accompaniment is modeled as a low-rank signal due to its assumed repetitive structure. Robust Principal Component Analysis (RPCA) is used as a factorization scheme to extract the desired sources. Another approach that takes advantage of the repetitive structure of the accompaniment is presented in [29]. The system first identifies the repeating period p of the signal using an autocorrelation approach to calculate a beat spectrum. The algorithm then models the repeating segment S as the element-wise median of the r segments of length p in the spectrogram.The repeating patterns are finally extracted using a soft masking approach.

2.2.3. Non-negative Matrix Factorization

NMF is a dimensionality reduction technique employed to represent non-negative data [30]. Given a data matrix $\boldsymbol{V}$ of size $n \times m$ with non-negative entries, the goal is to find a factorization given by:

$$\boldsymbol{V} \approx \mathbf{WH} \tag{2.13}$$

such that $\boldsymbol{W}$, $\boldsymbol{H}$ are also non-negative matrices of dimensions $n \times r$ and $r \times m$ respectively, with r less than n and m.

When used for sounds source separation, NMF is usually applied to the magnitude or power spectrogram where the matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ represent frequency and amplitude basis respectively. The columns of $\boldsymbol{W}$ can be interpreted as the spectral basis contained in the spectrogram. The rows of $\boldsymbol{H}$ can be interpreted as the weights of the spectral basis in each time frame. The assumption behind this approach is that the sum of all the spectrograms generated as the combination of basis functions, is the equal to the mixture spectrogram.

The approximate solution to the factorization problem is usually obtained through a minimization problem:

$$\underset{\boldsymbol{W},\boldsymbol{H} \geq 0}{min} D(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}) \tag{2.14}$$

where $D(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H})$ is the cost defined as:

$$D(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}) = \sum_{n=1}^{N}\sum_{m=1}^{M} d([\boldsymbol{V}]_{nm} \mid [\boldsymbol{W}\boldsymbol{H}]_{nm}) \tag{2.15}$$

where $d(x \,|\, y)$ is a scalar cost function usually taken from the family of β-Divergence cost functions. Commonly used cost functions such as the Euclidean Distance, Kullback-Leibler (KL) divergence, and the Itakura-Saito (IS) divergence are all special cases of the β-Divergence with $\beta = 2$, $\beta = 1$, and $\beta = 0$, respectively [31]:

$$d_{EUC}(x \mid y) = \frac{1}{2}(x-y)^2 \tag{2.16}$$

$$d_{KL}(x \mid y) = x \log \frac{x}{y} - x + y \tag{2.17}$$

$$d_{IS}(x \mid y) = \frac{x}{y} - \log \frac{x}{y} - 1 \tag{2.18}$$

The choice of cost function is highly dependent on the application at hand. In the case of audio signals, [31] reports that the IS divergence outperforms both the Euclidean and the KL divergence costs. However, [32] explored the use of β-Divergence masks in the context of audio source separation. In the experiments reported, the KL divergence used in the magnitude spectrogram resulted in better quality of separation.

Besides the basic NMF formulation, numerous variants of the factorization scheme have been developed. A comprehensive overview of NMF algorithms is presented in [33]. In this review, the authors categorize the NMF model in four subclasses: Constrained NMF, Basic NMF, Structured NMF, and Generalized NMF. Particularly relevant in the sound separation context is the *Convolutive* NMF under the Structured NMF subclass. It was originally proposed by Smaragdis in [34], and the goal is to incorporate time domain information in the decomposition; namely, the dependency between neighboring columns of the input data matrix $\boldsymbol{V}$. It is formally described as follows [34]:

$$\boldsymbol{V} \approx \sum_{t=0}^{T-1} \boldsymbol{W}_t \overset{t\rightarrow}{\boldsymbol{H}} \qquad (2.19)$$

where $\boldsymbol{W}_t$ varies across time and the $\overset{i\rightarrow}{(\cdot)}$ operator shifts the columns of its argument by i spots to the right. This formulation expresses the input data matrix $\boldsymbol{V}$ as the convolution between the basis matrix $\boldsymbol{W}$ and the weights matrix $\boldsymbol{H}$. The Convolutive NMF formulation can express more effectively the temporal continuity of the input signal (whose frequency varies in time) in the time-frequency domain.

Other variants of the NMF formulation that have been used for sound source separation include: *Sparse* NMF from the Constrained category of NMF algorithms. The sparse NMF formulation penalizes non-zero gains through a sparseness criterion in the cost function to be minimized [35]. *Weighted* NMF has also been applied to source separation [36, 37]. In [36] for example, the time-frequency components are given a perceptually motivated gain so that their contribution corresponds to their perceptual significance. Each weight is selected so that the weighted sum of spectral bins is equal to the estimated loudness of each time frame.

Finally, Probabilistic Latent Component Analysis (PLCA) is an important methodology for single-channel separation of sounds proposed by Smaragdis in [38]. PLCA is a statistical model that uses the product of marginal distributions to model an N-dimensional distribution. In the context of sound separation, the magnitude spectrogram is modeled with a 2-d PLCA formulation as the product of a frequency and a temporal marginals:

$$P(f, n) = \sum_{z} P(z)P(f/z)P(t/z) \qquad (2.20)$$

with $P(f, n)$ the magnitude spectrogram, $P(f/z)$ and $P(t/z)$ the frequency and time marginals respectively, and z the latent variable. When written in matrix form, the PLCA formulation becomes $\boldsymbol{V} = \boldsymbol{W}\boldsymbol{S}\boldsymbol{H}$, with $\boldsymbol{V}$ the magnitude spectrogram, $\boldsymbol{W}$ containing the columns of $P(f/z)$, $\boldsymbol{S}$ a diagonal matrix containing the elements of

$P(z)$, and $\boldsymbol{H}$ containing the rows of $P(t/z)$. If the matrix $\boldsymbol{S}$ is absorbed by $\boldsymbol{W}$ and $\boldsymbol{H}$, the model becomes the NMF decomposition: $\boldsymbol{V} = \bar{\boldsymbol{W}}\bar{\boldsymbol{H}}$. PLCA is a statistical interpretation of NMF that allows the incorporation of prior distributions. It has been applied to sound separation for example in [39, 40].

2.2.4. Non-negative Tensor Factorization

Non-negative Tensor Factorization (NTF) belongs to the Generalized NMF category of NMF-based algorithms [33]. Here, the factorization techniques from NMF are extended to deal with stereo and multi-channel signals. The idea behind the basic NTF model is that in a multi-channel scenario, the same basis function can be used to describe the contribution of a given instrument in each of the channels. However, the gains of the basis functions in each channel should be incorporated [41]. The factorization then becomes:

$$\hat{\mathcal{X}} \approx \sum_{k=1}^{K} \boldsymbol{G}_{:\boldsymbol{k}} \circ \boldsymbol{A}_{:\boldsymbol{k}} \circ \boldsymbol{S}_{:\boldsymbol{k}}. \tag{2.21}$$

where $\hat{\mathcal{X}}$ is a $c \times n \times m$ tensor containing the magnitude spectrograms of each of the c channels. $\boldsymbol{G}$ is a $c \times k$ matrix containing the gains of the k basis in each channel. $\boldsymbol{A}$ is a $n \times k$ matrix of frequency basis functions and $\boldsymbol{S}$ is a $m \times k$ matrix with the time activations of these functions. Finally, $\circ$ denotes the outer product multiplication and $:k$ the k-th column of a given matrix. Tensor factorization has been applied to drum separation from polyphonic music [42], score-informed separation [43], and Coding-based Informed Source Separation (CISS) [44] (see Section 2.3.4) among others.

2.3. Prior Information in Sound Source Separation

Many years of sound separation research have shown that separation performance can be improved when prior information about the sources is available. The inclusion of known information about the sources in the separation scheme is referred to as Informed Source Separation (ISS) and comprises, among others, the use of MIDI-like musical scores, the use of pitch tracks of one or several sources, and the extraction of model parameters from training data of a particular sound source [13]. This section describes the most common types of prior information used in separation approaches. As shown in Figure 2.1, the role of prior information in the separation process is to make the parameter estimation stage more robust, leading to better performance of the separation algorithms.

2.3.1. Pitch-informed Separation

Some approaches make use of pitch as prior information to perform separation. In these cases, an initial pitch detection stage extracts sequences of $f0$ (fundamental frequency) values of the target source. Due to the complexity of multi-pitch extraction, these algorithms mostly focus on the extraction of pitch values and consequent separation of the lead instrument only. Consequently, most of the pitch-informed separation approaches in the literature focus on the solo and accompaniment separation problem. For the purposes of this thesis, only those methods that directly extract pitch information before separation are included into the pitch-informed separation category. Those methods that make use of available MIDI-like representations such as musical scores are categorized as score-informed methods and are further explained in Section 2.3.2.

Initially, two pitch detection approaches that have shown superior performance in the Music Information Retrieval Evaluation eXchange (MIREX) in recent years, and that are of special relevance for the work presented in this thesis are described.

In the first approach [45], the author addresses the task of melody extraction from polyphonic music with an approach divided in four processing stages: (1) Spectral representation, (2) Pitch candidate detection and tone formation, (3) Voice formation, and (4) Main melody selection. A spectral representation is obtained starting with a multi-resolution spectrogram that provides a good trade-off between time resolution for higher frequencies and frequency resolution in the lower range. Magnitude and instantaneous frequency (IF) values are obtained for each peak within the frequency range of 55 Hz to 5 kHz. The magnitude of each spectral peak is weighted with its instantaneous frequency value to obtain a 6 dB magnitude boost per octave. Each spectral peak is either assigned to a previously existing tone (if it can be explained by the spectral envelope of such a tone), or is used to detect new salient pitches. To detect new salient pitches, a pair-wise evaluation of spectral peaks, which tries to detect partials with successive harmonic numbers, is used in conjunction with a set of perceptually motivated ratings. These ratings include a harmonicity threshold defined as a maximum deviation of 120 cents from the exact harmonic interval between the peaks, a measure to guarantee a degree of spectral smoothness, and a harmonic impact measure that reduces the impact of higher harmonics. In the voice formation stage, each voice is characterized by its magnitude and frequency range. A tone is assigned to a voice if it passes the magnitude threshold and lies within the frequency range of the voice. After different voices have been created, the most salient stream is selected as the main melody. In cases where no clear difference exists between the magnitude of two voices, a frequency weighting is applied that gives lower weight to voices in the lower frequency range. This approach will be referred to as **Alg1** in Section 3.3.8.1 of this thesis.

In the second approach [12], the authors propose a method for melody extraction from polyphonic music by pitch contour extraction and characterization. In this approach, pitch contours are defined as time continuous sequences of $f0$ candidates grouped based on auditory streaming cues such as harmonicity, pitch continuity, and exclusive allocation. This approach is divided in four processing stages: (1) Sinusoid extraction, (2) Salience function, (3) Pitch contour creation, and (4) Melody selection. For the sinusoid extraction an equal loudness filter is first applied to enhance the frequencies to which the human auditory system is more sensitive. The STFT is applied and instantaneous frequency and instantaneous amplitude values are obtained using phase differences. In order to obtain a salience function, an approach which computes the salience of a given frequency as the sum of the weighted magnitudes at integer multiples of that frequency is used. A compression parameter and a magnitude threshold are defined to prune the peak candidates and a frequency range of 55 Hz to 1.76 kHz is considered. To create the pitch contours, initial peak candidates are filtered using a salience threshold and a deviation threshold. The salience threshold is computed in relation to the highest peak in the frame and the deviation threshold is calculated using the salience mean and standard deviation of all remaining peaks. The final peaks are grouped into contours using heuristics based on auditory streaming cues. For each contour a set of features is calculated: pitch mean, pitch standard deviation, contour mean salience, contour total salience, contour salience deviation, length, and vibrato presence. For the melody selection stage, an initial voicing detection stage determines when the main melody is present and when it is not by setting a voicing threshold slightly below the average contour mean salience. Octave errors are also addressed by comparing pitch trajectories, which in case of octave relationships, will be almost identical with an octave separation. The correct contour is always assumed to be the most salient of the two and has to be somehow continuous with the other melody contours. If more than one contour is still present in a certain frame, the melody is selected as the peak belonging to the contour with the highest total salience. This approach will be referred to as **Alg2** in Section 3.3.8.1 of this thesis.

Several approaches for pitch-informed separation have also been proposed in the literature. The system described in [25] proposes a mid-level representation of the audio signal which is on the one hand invertible, which makes it suitable for sound separation applications, and gives, on the other hand, access to some semantically rich salience functions for pitch and timbre content analysis. The system uses an instantaneous signal model which represents the audio signal as the sum of a signal of interest, i.e., the lead instrument, and a residual, i.e., accompaniment. A source-filter model is used to represent the signal of interest. Information from the source is related to the pitch of the lead instrument and information from the filter is related to the timbre of the instrument. The residual is modeled using NMF. The mid-level representation is used to separate lead instrument from accompaniment in conjunction with a Wiener masking approach.

An interesting approach is presented in [46] where Computational Auditory Scene

Analysis (CASA) elements are introduced in the separation scheme. The system attempts to separate sound sources in monaural recordings using multi-pitch information of the sources obtained either from a MIDI-like score or from the multi-pitch detection algorithm presented by Klapuri in [47]. The multi-pitch information is used to differentiate overlapped harmonics from non-overlapped ones. This is performed by assuming harmonicity of the sources and by the use of a frequency threshold that assigns a set of frequency bins to a given harmonic. Harmonic masks are created for each of the sources by first refining the pitch estimates as the weighted average of the instantaneous frequency of the harmonics divided by their harmonic number. A new set of frequency bins is then assigned to each harmonic based on the refined pitch estimate. In the case of overlapped harmonics, the Common Amplitude Modulation (CAM) principle is applied in a least square estimation. The underlying assumption here is that the amplitude envelopes of the harmonic components of a source are correlated. In this system, the envelope of the strongest non-overlapped harmonic is used to estimate the envelopes of the overlapped ones. The system is evaluated with a dataset created from 20 MIDI files of Bach quartets where either 2 or 3 of the voices are created by inserting instrument notes taken from the RWC music instruments dataset.

In [48], an approach for singing voice extraction in stereo recordings is presented. The system is designed to specifically deal with real-time constraints achieving a latency of 250 ms (which is enough for on-line processing). To address the problem, two different spectral masks are obtained. An initial binary mask, i.e., spectral bin classification mask, based on panning information, phase difference between the stereo channels, and absolute frequency is calculated. A second mask, i.e., harmonic mask, is calculated based on a probabilistic pitch tracking approach composed of three steps: pitch likelihood estimation, timbre classification, and instrument pitch tracking. Two assumptions are made for the harmonic mask calculation: (1) the vocal component is fully localized in the spectral bins around the partials and (2) the singing voice is the only source present in those bins. Even though these assumptions are often violated, they greatly simplify the problem and lower the computational load. A final mask is obtained by combining the harmonic and the spectral bin classification masks.

2.3.2. Score-informed Separation

Musical scores are often used as prior information in separation methods. A score is a form of music notation where the parts of the different instruments (which contain their corresponding pitch and timing information) are specified independently. Having pitch and rhythm already as prior information naturally removes some of the difficulties inherent to pitch extraction; nevertheless, score-informed separation brings its own challenges and difficulties. Namely, due to artistic liberties taken in musical performances, audio recordings will never be completely synchronized with music scores. It is then necessary, before any separation is attempted, to align score and audio as precisely as possible. A common approach to address this issue is the use of Dynamic Time Warping (DTW) to find the optimal match between the two sequences. However, the spectrogram is not usually the best representation to perform the alignment as it can be very sensitive to possible differences between the sequences. The audio sequences are usually transformed into chroma representations as they have proven to be a very powerful and robust tool for synchronization of music [49]. Chroma features encode the energy distribution over 12 pitch classes of the equal-tempered scale , i.e., C, C#, D, D#, E,..., B. Generally speaking, chroma features give information about the harmonic content of a musical piece. A thorough overview of chroma representations can be found in [50]

In [49] for example, a system for high-resolution synchronization of audio streams via chroma-based onset features is presented. This system takes advantage of the robustness of chroma features and enhances them by calculating a set of chroma onset features. The onset features are obtained by selecting salient peaks in the time derivative of the energy curve of each pitch subband . The authors report increased accuracy in audio synchronization when the onset features are used in music with clear note attacks. In music where attacks are not so clearly marked, the authors report that no performance degradation occurs.

Similarly, [51] proposes the use of chroma-based DTW to address global misalignments between the score and the audio signal. A MIDI confidence measure is proposed to deal with small-scale misalignments. The confidence measure gives a lower weight to attack and offset regions of each note and a higher weight to the sustained part of the notes. The authors report that within a score-informed scheme for lead instrument separation, the proposed alignment produces results comparable to those obtained with manually synchronized scores. Another approach that proposes the use of score information to separate leading voice is presented in [52]. The authors propose a NMF routine with time and pitch constraints taken from the score, and a source/filter model to represent the leading voice.

In [53], information from the score is used both to provide signal models of the different sources, and to incorporate temporal and harmonic constraints in a NMF approach. The system divides the separation process in two phases: In an initial learning phase, synthesized signals from the score are decomposed with a NMF routine to generate models for the spectral bases and amplitude of each note. To impose the constraints, the activation coefficients of each note are initialized with a binary mask (1 if the note is playing, 0 if not) which results in a piano-roll representation of the score used to initialize the gain matrix. For the final separation phase, a new NMF routine is performed using the learned spectral bases and activation coefficients as initialization. A similar approach based on PLCA was proposed in [39].

An interesting online approach to score-informed separation is presented in [54]. As opposed to DTW-based approaches which need to have the entire score and audio signal to perform the alignment, this approach processes each incoming frame at a time. A hidden Markov approach is used to model each time frame as a 2D-state vector with score position and tempo as entries. The aligned score position and tempo are used to guide the separation stage where a pitch refinement stage is first conducted and a harmonic mask, that tries to take into account overlapping of harmonic components in the spectral domain, is extracted.

In [43], a Generalized Coupled Tensor Factorization (GCTF) approach is used to jointly include harmonic information from an approximate score and spectral information from instrument recordings. The system uses music scores as prior information but relaxes the alignment constraint between score and audio. The authors showed that a strict alignment between audio and score is not necessary when note co-occurrences, which are the same in the score and audio signal, are exploited.

A general overview of score-informed separation approaches is presented in [55].

2.3.3. Source Specific Prior Information

In those separation tasks where the type of source to be separated is known a priori, source specific information can be exploited as prior information in the separation scheme. That is the case for example of several approaches for lead voice separation that take advantage of characteristics of voice signals directly in their signal models or in their parameter estimation stages [52, 56, 57].

In [48] for example, a system for lead voice separation is presented where source specific information is included in a trained Support Vector Machine (SVM) for timbre classification of spectral envelopes of pitch candidates. The classifier attempts to predict the probability of a pitch candidate being a voiced envelope. The system is based on Mel Frequency Cepstral Coefficients (MFCC).

For the case of musical instruments, [58] presents a system for solo and accompaniment separation based on a probabilistic approach to NMF. In this system, source specific information is included in the form of temporal smoothness priors modeled as Inverse Gamma (IG) distributions. The smoothness priors are learned from a database of isolated instruments notes.

Similarly, timbre models for different musical instruments are used as time-frequency templates to guide a separation scheme in [22]. The separation method is based on sinusoidal modeling, followed by an onset detection stage that allows grouping of tracks based on their start and end points. A timbre matching stage attempts to assign an instrument to each track group before re-synthesis.

2.3.4. Oracle Source Separation

Oracle source separation refers to separation methods where the original sources are available as prior information. These methods are mainly relevant within an audio coding context where the main idea is not to transmit independent audio objects, but to transmit the parameters required to recover them using the mixture and a separation scheme. For this matter, the original sources are analyzed to extract relevant side information that is transmitted along with the mixture. In the decoder's side, the available side information is used to retrieve the independent sources from the transmitted mixture. In the audio coding community, this is known as Spatial Audio Object Coding (SAOC).

Given that in past years, the audio coding community and the separation community independently worked on this topic, [59] presents a general overview of the relationships between the approaches taken by the two communities and draws theoretical connections between ISS and source coding.

In [11], the concept of Coding-based Informed Source Separation (CISS) is extended to multi-channel mixtures. As opposed to source coding (which encodes the signals using their distributions only), CISS encodes the signals of interest using a probabilistic model relying on their a posteriori distribution given the mixture. The residual is encoded using posterior covariances as signal statistics. This represents one of the main characteristics of CISS, as the posterior dependencies between the sources are exploited. In this system a non-negative tensor factorization source model is used.

A similar approach is presented in [60] where information about the sources is extracted in the encoder's side and embedded in the mixture by means of a quantization-based watermarking technique on the Modified Discrete Cosine Transform (MDCT) coefficients.

2.3.5. User-assisted Source Separation

In the past few years, a number of methods have been proposed where the separation scheme is guided by information directly provided by the user. The idea here is to overcome some of the major difficulties faced by separation schemes such as the lack of reliable signal models or the difficulty of extracting accurate pitch sequences, by introducing information provided by the user.

Some systems, often referred to as *exemplar-based* methods, use example signals to perform separation. In [61], the user provides a version of the target source by humming the melody line of the source to be extracted. The user rendition is then used as prior in a PLCA factorization scheme. An extension to multi-channel signals that uses example signals as prior information is proposed in [62]. The systems makes use of a NTF scheme to perform separation.

In [63], the authors take advantage of the great number of multi-track recordings of cover songs commercially available and use them as prior information in a NMF approach. The use of cover songs in a separation scheme is referred to as *cover-informed separation*. A slightly related method is presented in [64] where an approach for common signal extraction is proposed with the goal of extracting common music in soundtracks in different languages.

Other methods have been proposed with the goal of overcoming one of the main difficulties of pitch-informed sound separation: errors in the pitch detection stage inevitably propagate to the separation stage. Some approaches that propose supervised pitch extraction with a consequent separation scheme have been presented in [65, O2]. In [66] for example, a probabilistic model of the Constant-Q Transform (CQT) is proposed for the estimation of polyphonic pitch content. Notes are modeled with time-frequency activations and normalized harmonic spectra. A sparseness prior is introduced for the note activations to guarantee that the data will be represented with the least amount of active notes possible. Through a especially designed interface, the user can choose the pitches to be extracted by clicking on them. Separation is performed with time-frequency masking in the CQT domain.

2.4. Performance Evaluation in Sound Separation

An important part of sound separation research is the development of appropriate methods to consistently evaluate the quality of extracted sources. Standardized evaluation metrics or evaluation setups allow on the one hand, a systematic evaluation and comparison of algorithms' performance under a unified dataset. They allow, on the other hand, the optimization of certain metrics that might be especially relevant in some applications. In this section, an overview of methods for quality assessment used in sound separation research is presented.

2.4.1. Subjective Quality Assessment

Subjective evaluation of audio quality is usually achieved by means of listening tests. In the source separation community however, listening tests have not been very common so far [67]. It is mostly in the audio coding and in the audio systems evaluation communities where active research in this field has been conducted in the past years. Specifically for audio quality evaluation, several standards have been published. Some relevant standards are: General methods for the subjective assessment of sound quality (ITU-R BS.1284-1) [68], Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems (ITU-R BS.1116-1) [69], and Method for the subjective assessment of intermediate quality level of coding systems (ITU-R BS.1534-1) [70]. This last standard is particularly relevant for the source separation community, as this is where the Multiple Stimulus with Hidden Reference and Anchors (MUSHRA) test is defined. The main goal of MUSHRA tests is to evaluate signals of intermediate quality by assessing the degradation of a test signal relative to a known reference. In the specific context of sound separation, the test signal represents the estimated source $\hat{s_j}(t)$ and the reference would be the original recording of the source $s_j(t)$. An adaptation of the MUSHRA test for sound separation evaluation is presented in [67] and in [71] similar listening tests have been conducted.

2.4.2. Objective Quality Assessment

This section describes two sets of objective evaluation metrics that have been proposed in the literature and that have been widely used in the research community: (1) the BSS evaluation metrics designed to evaluate the quality of single-channel source estimates, and (2) an extension of BSS to multi-channel environments that evaluates the quality of the spatial images of a source in the different channels.

The BSS is a set of four performance metrics that evaluates the quality of the extracted source $\hat{s_j}$ by means of energy ratios between the different signal components [72]. These metrics first attempt to decompose the signal into different signal distortions: interference from unwanted sources, sensor noise, and burbling artifacts (musical noise). The extracted source is then decomposed as follows:

$$\hat{s_j} = s_{target} + e_{interf} + e_{noise} + e_{artif} \tag{2.22}$$

where $s_{target} = f(s_j)$ is a version of the original source s_j modified by an allowed distortion f. The terms e_{interf}, e_{noise}, and e_{artif} are the interference, noise, and artifacts error terms, respectively. The following design requirements were considered for the development of these metrics:

- The performance metrics can be applied to underdetermined separation problems, i.e., the number of sources is larger than the number of available channels.

- Mixing and de-mixing system do not need to be known.

- True sources are available.

- The user can select the set of available signal distortions F according to the application: time-invariant gain distortions, time-invariant filters, time-varying gains, and time-varying filters.

The numerical performance criteria are then computed as energy ratios expressed in dB. Namely, Source to Distortion Ratio (SDR), Source to Interference Ratio (SIR), Source to Noise Ratio (SNR), and Source to Artifacts Ratio (SAR) [72]:

$$SDR = 10\log_{10}\frac{\|s_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2} \tag{2.23}$$

$$SIR = 10\log_{10}\frac{\|s_{target}\|^2}{\|e_{interf}\|^2} \tag{2.24}$$

$$SNR = 10\log_{10}\frac{\|s_{target} + e_{interf}\|^2}{\|e_{noise}\|^2} \tag{2.25}$$

$$SAR = 10\log_{10}\frac{\|s_{target} + e_{interf} + e_{noise}\|^2}{\|e_{artif}\|^2} \tag{2.26}$$

As an extension of BSS to multi-channel environments, a set of four objective performance measures was proposed to evaluate the contribution of source j to channel i. This is referred to as, the *spatial image* of source j in channel i, with $j = 1, \dots, J$ images, and $i = 1, \dots, I$ channels. The estimated image of source j in channel i is modeled as follows:

$$\hat{s}_{ij}^{imag}(t) = s_{ij}^{imag}(t) + e_{ij}^{spat}(t) + e_{ij}^{interf}(t) + e_{ij}^{artif}(t) \tag{2.27}$$

where $s_{ij}^{imag}(t)$ is the true source image, and $e_{ij}^{spat}(t)$, $e_{ij}^{interf}(t)$, and $e_{ij}^{artif}(t)$ are error components of spatial, interferences, and artifacts distortions respectively. The set of objective measures is once again defined as a an energy ratio between signal components and is expressed in dB. In this case, the following measures are defined: Source Image to Spatial Ratio (ISR), Source to Interference Ratio (SIR), Source to Artifacts Ratio (SAR), and Source to Distortion Ratio (SDR).

$$ISR_j = 10\log_{10}\frac{\sum_{i=1}^{I}\sum_t s_{ij}^{imag}(t)^2}{\sum_{i=1}^{I}\sum_t e_{ij}^{spat}(t)^2} \tag{2.28}$$

$$SIR_j = 10\log_{10}\frac{\sum_{i=1}^{I}\sum_t (s_{ij}^{imag}(t) + e_{ij}^{spat}(t))^2}{\sum_{i=1}^{I}\sum_t e_{ij}^{interf}(t)^2} \tag{2.29}$$

$$SAR_j = 10\log_{10}\frac{\sum_{i=1}^{I}\sum_t (s_{ij}^{imag}(t) + e_{ij}^{spat}(t) + e_{ij}^{interf}(t))^2}{\sum_{i=1}^{I}\sum_t e_{ij}^{artif}(t)^2} \tag{2.30}$$

$$SDR_j = 10\log_{10}\frac{\sum_{i=1}^{I}\sum_t s_{ij}^{imag}(t)^2}{\sum_{i=1}^{I}\sum_t (s_{ij}^{artif}(t) + e_{ij}^{spat}(t) + e_{ij}^{interf}(t))^2} \tag{2.31}$$

An important characteristic of these two sets of objective measures is that they assign equal weights to the different error terms. This assumes that in terms of quality, all types of distortions contribute equally to the overall quality of the extracted source. However, experience has shown that different applications pose different quality requirements and while artifacts, for example, might play a very important role in hearing aid research, their importance might not as critical when it comes to karaoke applications [73]. Another important characteristics of these sets of measures is that they do not take into consideration perceptual aspects of hearing for their calculations. This comes as a major drawback as the measures do not necessarily correlate to perceptual attributes and consequently, their ability to fit subjective ratings can be questioned [O3, O4]. In cases for example, when the perceived loudness of interference or artifacts is much smaller than the power of the corresponding signals, resulting numeric values can be misleading.

2.4.3. Objective Perceptual Quality Assessment

The development of objective perceptual measures for source quality assessment came as an attempt to exploit the main strengths of both subjective and objective measures into a single evaluation scheme: while subjective measures take into account the perceptual phenomena of hearing, objective measures allow consistent numerical evaluation of the extracted sources.

The PEASS Toolkit –Perceptual Evaluation Methods for Audio Source Separation– was developed as a set of four objective perceptual measures that attempt to predict subjective scores by decomposing the signal into different types of distortions; namely, interference, artifacts, and target distortions [71]. The spatial image of source j in channel i is modeled as follows:

$$\hat{s}_{ij}(t) = s_{ij}(t) + e_{ij}^{target}(t) + e_{ij}^{interf}(t) + e_{ij}^{artif}(t) \tag{2.32}$$

Subjective scores were obtained by means of a listening test protocol designed to address the perceptual characteristics of the distortions components: target, interference, and artifacts. Objective scores were obtained by calculating the perceptual salience of each specific distortion and of the overall distortion using the PEMO-Q auditory model [74]. Subjective and objective results were joined using non-linear mappings which aim at combining the salience features obtained with PEMO-Q into a single scalar value, and at adapting the feature scale to the subjective scale from the listening test.

A family of four objective perceptual measures was proposed: Overall Perceptual Score (OPS), the Target-related Perceptual Score (TPS), the Interference-related Perceptual Score (IPS) and the Artifacts-related Perceptual Score (APS). The PEASS Toolkit, freely accessible in [75], is considered the state-of-the-art evaluation scheme for separation research and it is used in current public evaluation campaigns described in the next section.

2.5. Public Evaluation

Several public evaluation campaigns for sound source separation have taken place in the last few years. These campaigns play an important role in the research community as they allow regular evaluation of the effects of algorithm design, they specify a common evaluation method, and promote the results and advances in the field.

A structured evaluation of a source separation system requires four important elements: (1) a dataset, (2) a task to be addressed, (2) one or more evaluation criteria, and (4) performance bounds [76, 77].

Several datasets have been compiled and released in the last few years and are freely available for the research community. Table 2.1 presents a list of some of datasets, generally described as *Professionally Produced Music Recordings*, that are relevant to the work presented in this thesis.

Table 2.1.: Public dataset that can be used for sound source separation

Name	**Year**	**Published by**	**Ref**
BASS_dB	2006	IRISA	[78]
SiSEC2008	2008	SiSEC	[79]
SiSEC2010	2010	SiSEC	[80]
TRIOS	2012	Queen Mary University	[81]
SiSEC2013	2013	SiSEC	[82]
SA_DS2	2013	Fraunhofer IDMT	[83]

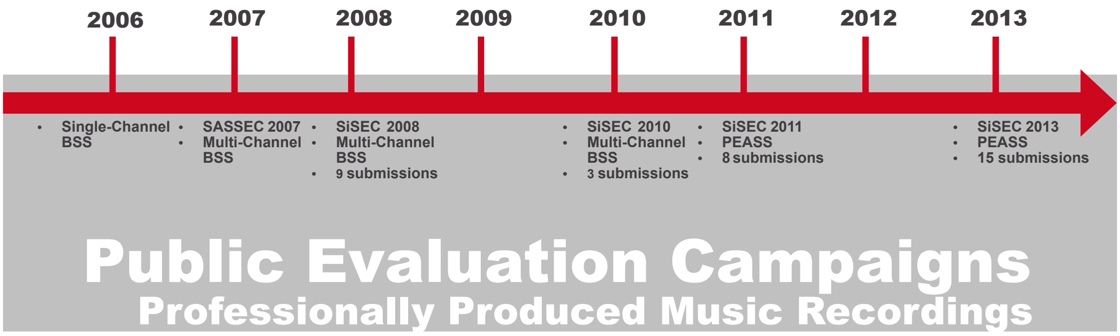


Figure 2.2.: Timeline of public evaluation campaigns. Name of the campaigns, evaluation criteria, and number of submissions for the Professionally Produced Music Recordings Dataset are displayed.

In terms of the tasks to be addressed in a structured evaluation, some examples that have been proposed in separation campaigns are: source counting which attempts to determine the number of sources in a mix, source signal estimation, source spatial image estimation, and source direction of arrival DOA estimation.

The third element of a structured evaluation refers to the evaluation criteria. In terms of the evaluation criteria used in separation campaigns, the metrics described in sections 2.4.2 and 2.4.3 have been used in previous separation campaigns. See Figure 2.2 for a chronological overview of the usage of these metrics.

Finally, performance bounds can be defined in terms of a reference algorithm which is evaluated under the same criteria. Oracle estimators, for example, can be used to provide a theoretical upper bound on performance.

With these requirements in mind, several evaluations campaigns have been held in the last few years, including the Stereo Audio Source Separation Campaign (SASSEC 2007) [73], and the 2008, 2010, 2011, and 2013 Signal Separation Evaluation Campaign (SiSEC) [84]. While SASSEC2007 was restricted to audio applications only, the following SiSEC campaigns were open to other applications, such as biomedical data. The SiSEC campaigns are run in conjunction with the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA) as community-based scientific evaluations. Figure 2.2 shows a chronological description of the different campaigns, the evaluation criteria used in them, and the number of submissions for the *Professionally Produced Music Recordings* estimation.

2.6. Spectral Parameters of Musical Instrument Signals

In this section, the general characteristics of the spectral parameters of musical instrument signals are described. Namely, the spectral magnitude, frequency, and phase of musical instrument signals are described in an attempt to draw general directions that can be used within a solo and accompaniment separation context. As will be explained in Section 3.1, our proposed method for solo and accompaniment separation is built upon the idea that the spectrum of the solo instrument can be estimated with a certain accuracy when pitch information is available, and simple acoustical and spectral characteristics of musical instrument tones are considered. In this section, the theoretical background that allows the understanding of spectral parameters of musical tones is described.

2.6.1. Magnitude

Throughout the years of sound separation research, spectral magnitude has received a lot of attention from the research community. Most sound separation approaches work directly on the magnitude spectrogram and perform source estimation entirely in this domain. When it comes to representing the magnitude information of a given source, two different representations are commonly used : one exhibits the *frequency* characteristics of the source in a given time instant, and the other shows the *time* evolution of a given harmonic of a source. The first representation is called the *spectral envelope* and it is defined as a smooth function of frequency that tracks the individual partial peaks of a source in a given time frame. The second representation called the *amplitude envelope* or *temporal envelope*, exhibits the frame-wise evolution of the amplitude of a given partial of a sound source. In Figure 2.3 the amplitude envelope of the first harmonic *h1* of an A4 trombone tone without vibrato, and the spectral envelope of the same tone in a given time frame, are displayed.

Some approaches have attempted to describe musical instruments by their spectral envelopes. In [85] for example, isolated instrument tones from the clarinet, saxophone, and trombone are used to create a library of spectra to be used within a harmonic source separation system. The method first attempts to perform polyphonic pitch detection by modeling the spectrum of the target source as a mixture of Gaussian distributions located at integer multiples of the fundamental frequency. Overlapped harmonics are detected based on the estimated $f0s$. To replace corrupted harmonics, the library of pre-stored spectra is searched for the best match to the uncorrupted harmonics. The best match match is found using a simple Euclidean distance measure.

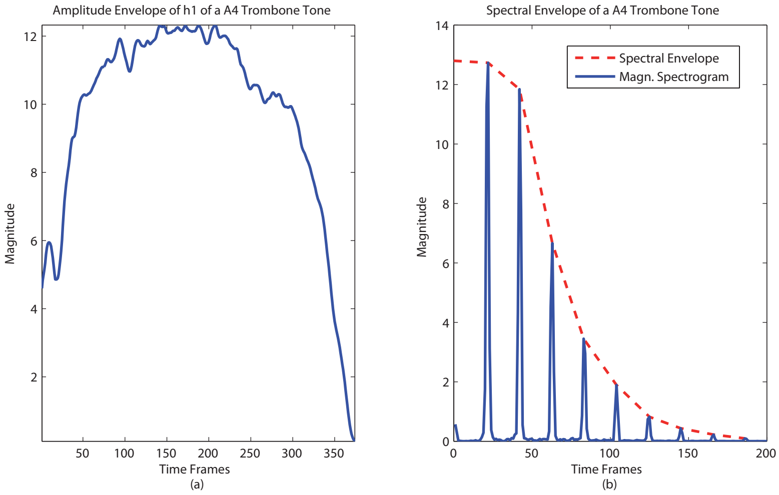


Figure 2.3.: Temporal and spectral representations of the magnitude spectrogram: (a) Amplitude envelope of the first harmonic *h1* of an A4 trombone tone without vibrato. (b) Spectral envelope in time frame $n = 200$ of the A4 trombone tone without vibrato shown in (a). The original magnitude spectrogram is displayed with a continuous blue line, the spectral envelope is displayed with a red dashed line.

Results have shown however, that spectral envelopes of musical instruments are difficult to model. The prediction of unknown amplitude values of harmonic components based on known information from neighboring harmonics of the same source has not shown consistent results [46, 86]. With this in mind, several approaches have attempted to only enforce smoothness in the spectral envelopes of the estimated sources without trying to model a particular behavior for the different musical instruments. Even when a smoothness constraint is not as powerful as the hypothetical idea of finding a representative spectral envelope for each musical instrument, it is a more robust assumption that holds for many harmonic musical instrument tones. Some approaches have used the smoothness of spectral envelope to resolve overlapped harmonics of different sources. In [47], the smoothness of the spectral envelope is exploited within a multiple fundamental frequency estimation method.

The spectral envelope of overlapped sections of the spectrum is smoothed using a moving average filter. The amplitude of overlapped harmonics is chosen as the minimum between the original amplitude and its filtered version. Similarly, [87] exploits the spectral envelope smoothness for separation of synchronous pitched notes. To deal with overlapped harmonics, filters are designed to split the spectral content shared by M overlapping harmonics into M parts using M overlapping filters. The amplitude of the harmonics is predicted using simple linear interpolation between the amplitudes of the nearest non-overlapped harmonics.

The spectral envelope is often used to describe characteristics of the singing voice. The vibration of the vocal folds produces a varying air flow of periodic nature. The vocal tract acts as a variable filter which can change its response depending on the position of the tongue and shape of the mouth opening. This variable nature occurs due to the fact that any changes in the tongue position and mouth opening, change the physical dimensions of the vocal tract. A resonant frequency R_1 in the vocal tract gives rise to a formant at frequency F_1. The term *formant* refers to a broad peak in the spectral envelope of the singing voice. The *singer's formant* is a broad band of enhanced power, noticed in the spectral envelope of classically trained singers. The singer's formant clusters the 3rd, 4th, and 5th resonances of the vocal tract and is usually evident in the frequency range of $[2\,\text{kHz}, 4\,\text{kHz}]$. Except for high voices as the soprano voice, the fundamental frequencies of the singing voice usually fall below any of the resonances, leading in many cases, to fundamental frequencies weaker than the other harmonics [88].

In contrast, some approaches have focused on the temporal information of the harmonics and use the amplitude envelope to perform separation. Here, the concept of Common Amplitude Modulation (CAM) becomes relevant. CAM refers to the observed characteristic that amplitude envelopes of the harmonics of the same source tend to be similar and correlated. Correlation studies of amplitude envelopes in musical instruments are reported in [46, 89]. Results from these studies conclude that correlation values are highest between adjacent partials and decrease exponentially with increasing distance. Results also suggest that strong harmonics are highly correlated with one another, but approximating low-energy harmonics using strong non-overlapped harmonics may be less accurate.

In [90], CAM is used within a multi-channel separation approach. The system first estimates regions in the time-frequency plane that cover the main energy of one or more overlapped harmonics. Each region is mapped to a certain source using the harmonicity principle, spatial cues, and CAM. The amplitude envelopes of the non-overlapped sections are used as models for the overlapped sections. A similar approach that uses CAM for pitch-informed sound separation is presented in [46]. The details of this system were described in Section 2.3.1 of this thesis.

2.6.2. Frequency

It is a well known fact that the spectrum of a musical instrument tone has peaks at approximately harmonic ratios. Given a fundamental frequency $f0$, harmonic peaks will most likely be evident in the spectrum, in the regions around $2f0$, $3f0$, $4f0$, etc. The degree to which the frequencies of the overtones deviate from integer multiples of the fundamental frequency is called *inharmonicity*. In real musical instruments where assumptions such as infinitely thin and flexible strings, or perfectly cylindrical or conical pipes do not completely hold, it is expected that the frequencies of the overtones deviate to different extents (depending on the instrument and $f0$) from harmonic ratios.

In Figure 2.4 the magnitude spectrogram of an A#3 trumpet tone without vibrato is displayed. The red vertical lines show the expected location of the harmonic components when integer multiples of the fundamental frequency ($f0 = 236\,\mathrm{Hz}$) are calculated. It is evident in the figure not only that harmonics deviate from their harmonic locations but also that the deviation changes with harmonic number: the higher the harmonic number, the larger the frequency deviation. Even when these exact observation cannot be generalized to all musical instruments, the tone displayed in Figure 2.4 is a clear example of an inharmonicity pattern in musical instrument sounds.

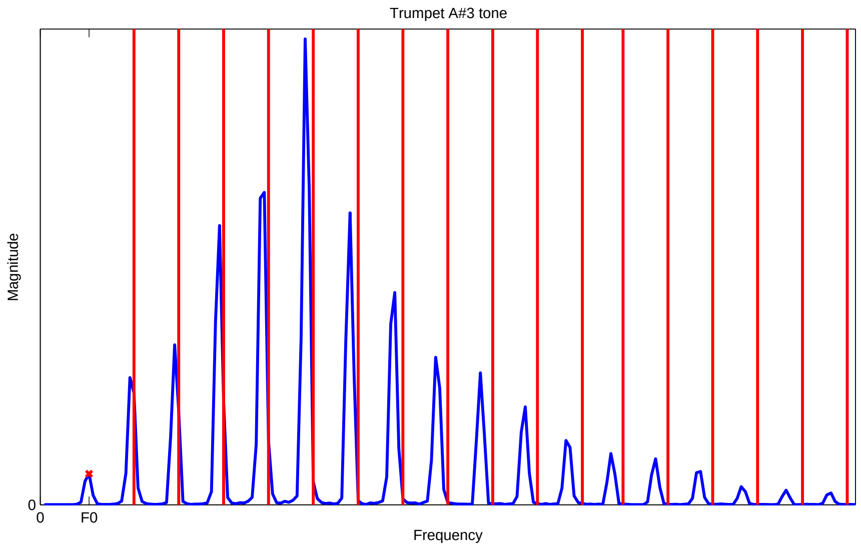


Figure 2.4.: Magnitude spectrogram of an A#3 trumpet tone (blue): the red vertical lines mark the expected location of the harmonic components when calculated as multiple integers of the fundamental frequency $f0 = 236\,\mathrm{Hz}$. The frequency deviation of the observed harmonic peaks from their harmonic locations can be clearly seen.

Different aspects contribute to the harmonicity (or deviation from it) of a musical tone. As described in [91], musical instruments capable of producing sustained tones consist of one or more resonant systems (air columns, cavities, strings), excited by a non-linear source (lips, reed, air jet, strings) with which they are coupled. In the resonant system, natural modes are never in exact harmonic relation because of second order effects like end corrections and string stiffness. In other words, the harmonic components (overtones), given a fundamental frequency $f0$ are never in exact harmonic relation. End corrections refer to the phenomenon occurring at the open end of a pipe: When a pulse of high pressure air gets to the end of the pipe, it spreads out and this allows reflection; however, the air outside the pipe has its own mass and inertia and thus, the reflection does not happen immediately at the open end but slightly beyond it. Consequently, the pipe appears to be longer (*effective* *length*) than it physically is (*geometrical length*). This effect causes flattening of upper resonances in conical bore instruments. In the case of string instruments, inharmonicity effects are caused by the fact that real strings are not infinitely thin or flexible and as such, they do not bend perfectly. This bending stiffness affects especially the higher modes of vibration and they become stretched, often appearing at higher frequencies than the calculated harmonic ones.

The excitation source of musical instruments also plays an important role in the harmonicity of musical instruments. The effects of the bow on string instruments, and the effects of reeds and lips on wind instruments, provide a locking mechanism that results in spectra of nearly harmonic tones. The effect of the bow on string instruments (*stick-slip phenomenon*) drives all of the resonances of the string at nearly exact harmonic ratios, even if it means driving the resonant modes slightly off their natural frequency. A similar effect occurs in wind instruments where reeds (as in the clarinet and saxophone) and lips (as in the trumpet, horn, or flute) undergo their own periodic vibrations which act as a locking mechanism to find a compromise between all the slightly inharmonic modes of vibration. This phenomenon is referred to as *mode locking.* For plucked string instruments such as the piano, guitar, or strings played with pizzicato, such locking mechanisms are not present and inharmonicity is much more noticeable.

Several studies have been conducted to characterize the frequency relationship between the harmonic components and the fundamental frequency of a tone. In [92], a study is presented were clarinet, voice, alto flute, piano, violin, viola, and cello tones are analyzed to calculate frequency ratios of their spectral components. As a frequency tracker, the Single Frame Approximation method, proposed by the authors, was used. This method is equivalent to the phase vocoder but with a time advance of one sample only. The benefit of this method over the phase vocoder was its computational efficiency (which was still relevant at the time for such calculations) as only one Fast Fourier Transform (FFT) frame needed to be calculated. This study reported harmonic ratios (to a 0.2% achievable accuracy) for all instruments but the piano and string instruments played with pizzicato. For the piano, the study found that deviation from harmonicity increases with partial number, and was found to be proportional to the harmonic number squared. For the case of string instruments played with pizzicato, the ratios were also found to deviate from integer relationships but no quantitative values are presented. The measurement of pizzicato notes was reported to be problematic due to the rapidly decaying amplitudes of the partials. An interesting finding was the fact that for bowed string instruments played with vibrato, frequency variations of the fundamental frequency exceeded those of the tracked partials. In the current separation context where the harmonic components are always estimated based on an initial estimate of the fundamental frequency, this effect will prove to be problematic and could possibly lead to wrong estimations.

In [93] a statistical study of the spectral parameters of seven C5 instrument tones is presented. The statistical study was directed to finding spectral characteristics that could lead to better synthesis models of musical instruments. Violin, flute, and oboe were analyzed with and without vibrato, while the trumpet was only analyzed without vibrato. Results show mean frequency deviations from harmonicity that are nearly constant for all harmonics (in the ±0.3% range); however, standard deviations of percentage frequency deviations tend to be slightly larger for higher partials than for lower ones. An interesting finding is that for vibrato tones, standard deviations of the percentage frequency deviations are much larger than for the non vibrato tones (5 times larger for the violin, and twice as large for the oboe and flute)

Because inharmonicity effects are much more notorious in plucked string instruments, special attention has been given to instruments such as guitar, piano, and harpsichord, and several studies that attempt to characterize different aspects of inharmonicity have been presented. The inharmoncity of plucked strings is described by the following formulation [94]:

$$f_k = kf_0\sqrt{1+\beta k^2} \tag{2.33}$$

In this equation, f_k represents the frequency of the kth partial given a fundamental frequency f_0, and an inharmonicity constant β determined by the physical characteristics of the string.

In [95] for example, the inharmonicity of the electric guitar is studied, showing that human listeners slightly prefer synthesized sounds with inharmonic characteristics than the perfectly harmonic ones. It was also shown that lower strings produce larger inharmonicities than higher strings. In [96] the audibility of inharmonicity is studied for guitar and piano sounds as a function of β. Results from the listening test show that detection of inharmonicity is dependent on the fundamental frequency, being more easily detected in the lower frequency range. Other studies that have focused on inharmonicity of string instruments are [97, 98, 99]

In the case of the singing voice, the source/filter model explained in Section 2.2.1.2 is often used to describe the process of voice production. The *source* in this case comes from the vibration of the vocal folds which produces a varying air flow often treated as a periodic source. Thus, harmonic components are expected to appear in nearly integer multiples of the fundamental frequency. The vocal tract acts as a variable *filter* which can change its response depending on the position of the tongue and shape of the mouth opening. As explained in Section 2.6.1, the filter modifies the spectral envelope depending on the resonant frequencies of the vocal tract. An important aspect in the characterization of the singing voice is the presence of both voiced and unvoiced sounds which need to be considered in any spectral analysis. *Voiced* sounds are produced by the vibration of the vocal chords producing a periodic wave. Normal vowel sounds (as opposed to whispered vowel sounds) are examples of voiced sounds. In *unvoiced* sounds, the vocal chords do not vibrate but are held very close together. This produces a flow of air of turbulent quality which produces a *broadband* sound characterized by a flat spectrum [100]. Whispered vowels and consonant sounds such as *ss, sh, f, p, t, k* are unvoiced. Consonant sounds are classified as *fricatives*, where the vocal tract is constricted and a turbulent flow contributes a broadband sound to the spectrum, and *plosives*, which involve the opening and closing of the tract and produce a brief broadband sound. Both fricatives and plosives can be voiced or unvoiced. Examples of voiced fricatives are *z, j, v*; unvoiced fricatives are *ss, sh, f*. Similarly, voiced plosives are *b, d, g* and unvoiced ones are *p,t,k* [88].

As explained in this section, musical instruments and the singing voice can exhibit very diverse frequency characteristics. This diversity naturally represents a big challenge for the development of algorithms capable of dealing with different types of sound sources.

2.6.3. Phase

In the most general sense, phase refers to the fraction of the cycle of a harmonic wave that has elapsed with respect to the origin at a given time. The unit circle is commonly used to describe the characteristics of phase, as for any given harmonic wave, a full period T_0 occurs every 2π radians or every full turn around the unit circle. The unit circle is shown in Figure 2.5. Here, phase is represented by the angle ϕ with respect to the origin.

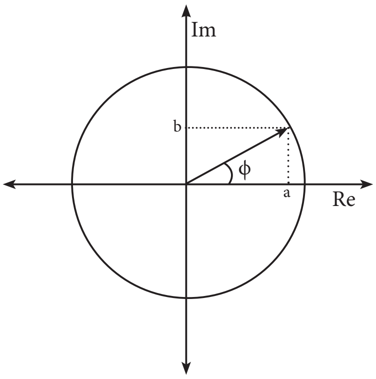


Figure 2.5.: Representation of phase in the unit circle. The real components are plotted in the horizontal axis. The imaginary components are plotted in the vertical axis. The angle ϕ with respect to the origin represents the phase.

The Fourier transform of signal $x(t)$ at frequency bin k and time frame n is given by:

$$X(k,n) = |X(k,n)| \cdot \exp^{i\phi(k,n)} \tag{2.34}$$

where $X(k, n)$ represents the complex valued spectrogram, $\phi(k, n)$ the spectral phase, and $|X(k, n)|$ the spectral magnitude. If the Fourier transform is analyzed from the unit circle point of view, it can be easily seen that $X(k, n) = a + ib$. Here, $a = X_R(k, n)$ and $b = X_I(k, n)$ are the real and imaginary parts of $X(k, n)$, respectively. See Figure 2.5 for reference.

The spectral phase is denoted by $arg[X(k, n)]$ and can be calculated as follows:

$$\phi_{\mathcal{W}}(k, n) = \arctan(\frac{b}{a}) \tag{2.35}$$

Two important aspects should be considered when dealing with phase:

- From the definition of phase, it is evident that phase takes values within the $-\pi < \phi_{\mathcal{W}}(k, n) \leq \pi$ range. This is often referred to as *wrapped* phase, as it is wrapped around $\pm\pi$. Following the notation presented in [101], the subscript $\mathcal{W}$ refers to the wrapped phase.
- There are inherent discontinuities in phase at $\pm\pi$. This makes phase in its original form, difficult to predict.

To avoid confusions, it is important to note that the concept of phase can in general be interpreted in two different ways. On the one hand, phase can refer to the change in the arrival time of a signal's frequency component due to variations in the path length between the signal source and the ear [102]. On the other hand, the term phase can be used to refer to the short-time phase spectrum. In this thesis, the terms *phase* and *phase spectrum* are used interchangeably always to refer to the short-time phase spectrum in a Fourier analysis.

Of all the spectral parameters described in this section, phase has probably received the least amount of attention from the separation community. Until now, very few separation methods have explored the analysis of phase within their processing schemes. Several factors have contributed to the lack of studies related to phase in the separation context:

- The phase spectrum is an abstract representation where information is difficult to extract and model. In contrast, the spectral magnitude is easy to analyze and parameterize into features.

- The assumption that humans are insensitive to phase. This assumption comes from the frequent confusion between the two definitions of phase (explained in the previous paragraph). The human ear is **relatively** insensitive to changes in the arrival time of a signal's frequency component due to variations in the path length between the signal source and the ear; however, spectral phase has proven to highly contribute to speech quality and naturalness and as such, it is used for speech coding algorithms and high-quality speech synthesis [102]. In the case of musical signals, no formal studies have been conducted (to the author's knowledge) to evaluate the contribution of phase in signal quality. In Section 3.2.2 of this thesis a study of the influence of phase in separation quality is presented.

- Phase is very sensitive to modifications and even minor changes can lead, in some cases, to heavily distorted signals.

- The development of factorization schemes that allow the inclusion of phase is still limited.

Most of the studies on the importance of phase have been conducted within the speech community. A very early study on phase is presented in [103]. The authors conducted a study where noisy speech signals are generated by adding white Gaussian noise to clean speech signals. The Signal to Noise Ratio (SNR) of the Fourier magnitude and phase were systematically changed. To assess the effects of phase on speech enhancement for example, the SNR of the magnitude was kept unchanged and the SNR of the phase was varied. A listening test was conducted where subjects were asked to rate the quality of the audio signals with different SNRs. Results showed that only for very low magnitude SNRs, the equivalent SNR of the reconstructed signal improves significantly with a more accurate estimate of the phase.

Another study presented in [104], evaluated the importance of phase for speech intelligibility. The authors conducted a listening test where subjects were asked to recognize utterances synthesized with different spectral parameters. Magnitude only signals, synthesized with random phase and phase only signals, synthesized with unit magnitude were used during the test. Results showed that phase contributes to speech intelligibility as much as the spectral magnitude. An important open question was raised by the authors in the conclusion of the paper: even when both magnitude and phase proved to be equally important for speech intelligibility, it was not clear whether their effects are complementary or independent.

In [105] the authors evaluate the effects of uncertainty (error) in the phase of speech signals on the word recognition error rate of human listeners. Results from the listening test showed that the importance of phase is SNR-dependent. At lower SNRs, the effects of phase uncertainty are more pronounced than at higher SNRs. For all the experiments, speech signals were corrupted using Gaussian noise.

When it comes to sound source separation, only a few systems have attempted to address phase. In [106] for example, reconstruction of separated sources is obtained by applying the spectrogram inversion technique originally proposed by Griffin and Lim in [107]. The spectrogram inversion technique attempts to estimate a signal from its modified STFT by minimizing the mean squared error between the STFT of the estimated signal and the Modified Short Time Fourier Transform (MSTFT). The complex MSTFT is obtained with the estimated magnitude spectrogram of the target source and the original phase of the audio mixture. Most recently, the problem of spectrogram consistency has been addressed by [16]. Spectrogram consistency refers to the guarantee that a given STFT actually corresponds to a time domain audio signal. In the separation context, most algorithms work on the magnitude spectrogram to obtain estimates of the target sources. The original phase from the audio mixture is often used to obtain a complex valued spectrogram. This procedure often results in arrays of complex numbers that do not correspond to any audio signal. In [16], the authors address the problem of spectrogram consistency within a generalized Wiener Filtering or time-frequency masking approach.

In a slightly different context, phase information was used in [108] to resolve overlapped harmonics within a separation scheme. The systems works under two assumptions: (1) The harmonic components of the same source have correlated magnitude envelopes. This is known as Common Amplitude Modulation (CAM) and was further explained in Section 2.6.1. (2) Phase change of harmonics can be accurately predicted from an instrument's pitch. Overlapped harmonics are detected using known pitch information about the sources. The system uses a sinusoidal modeling approach to formulate a set of equations representing the time-frequency bins where the overlapping occurs. A solution in a least squares sense is found.

Another aspect of phase relevant to this thesis is the capability of predicting the phase spectrum of an audio signal. In [O4] we presented a preliminary study of the use of phase prediction in the context of sound separation. This study is described in Section 3.2.3.4 of this thesis. In [101] the authors address the prediction of phase spectra of audio signals from two different approaches: a least squares estimation and a neural network approach. The goal of this study was to estimate the phase at a specific point in the phase spectrum by using the observed values of the phases at neighboring points. Their study concluded that there exists structure in phase that allows the prediction of phase spectrum to a certain extent.

An important aspect in the prediction of phase spectrum is the concept of *unwrapped phase*. To better handle phase information it is a common procedure to unwrap the phase spectrum to obtain a continuous representation where the discontinuities are removed. The most common process of phase unwrapping corrects the phase values by adding multiples of $\pm 2\pi$ when absolute jumps between adjacent values are greater than or equal to a pre-defined tolerance. The tolerance is normally chosen to be π. This is the approach taken in Matlab's unwrap function.

In the remainder of this thesis, the unwrapped phase is referred to as $\phi_{k,n}$ (the superscript $\mathcal{W}$ has been dropped). In Figure 2.6 the wrapped and unwrapped phase of the fundamental frequency of an A#3 alto saxophone tone without vibrato is displayed. It can be seen how phase unwrapping results in a continuous phase curve much easier to model.

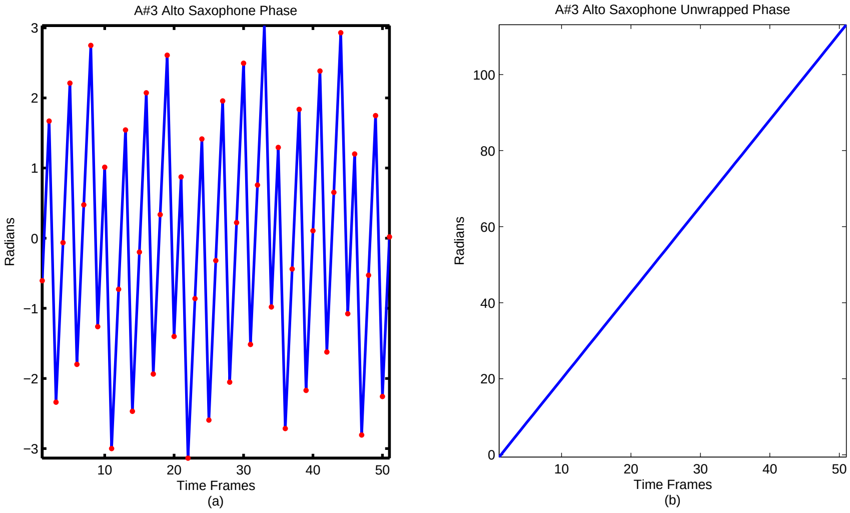


Figure 2.6.: Phase spectrum of the fundamental frequency of an A#3 alto saxophone tone without vibrato: (a) Wrapped phase $\phi_{\mathcal{W}}(k,n)$, (b) Unwrapped phase $\phi(k,n)$. It can be observed how phase unwrapping removes the discontinuities of standard Fourier phase.

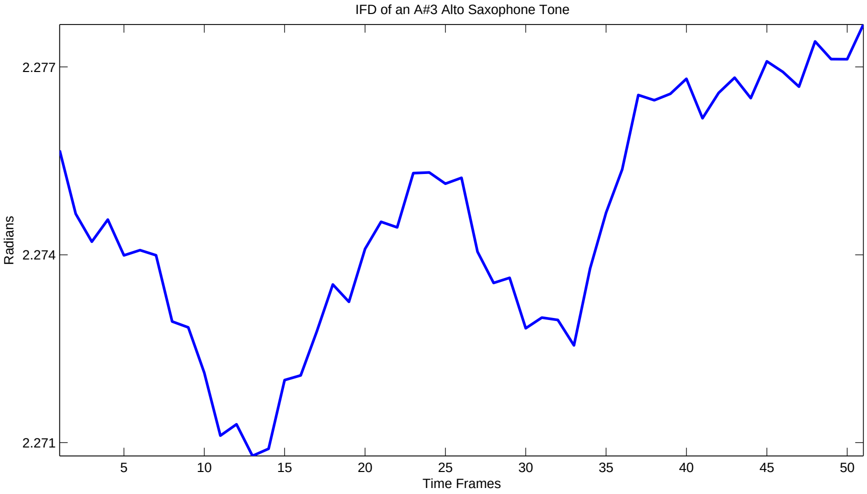


Figure 2.7.: Instantaneous Frequency Distribution (IFD) $\Phi(k,n)$ of the A#3 alto saxophone tone without vibrato presented in Figure 2.6. The IFD representation clearly shows the variations of the instantaneous frequency over time.

Another important representation of phase derived from the short-time phase spectrum is its first-order time derivative. This representation is called Instantaneous Frequency Distribution (IFD) and is defined as follows:

$$\Phi(k,n) = \frac{1}{2\pi}\frac{d\phi(k,n)}{dn} \tag{2.36}$$

where $\phi(k,n)$ is the unwrapped phase spectrum. In practice, the differentiation in (2.36) is approximated by taking the difference between two consecutive values of the phase spectrum. The division by 2π is used to normalize the instantaneous frequency (IF). The normalized IF can be used to obtain the IF in Hertz simply by multiplying it by the sampling frequency fs [102]. In Figure 2.7, the IFD of the saxophone tone presented in Figure 2.6 is displayed.

3. Pitch-informed Solo and Accompaniment Separation: Studies and Proposed Methods

This chapter presents a series of studies and methods developed in the attempt to reach the goals of this thesis. Here, a summary of our goals is presented again for reference:

> *The goal of this work is the development of a system for pitch-informed solo and accompaniment separation capable of separating main instruments from music accompaniment, regardless of the type of solo instrument used, musical genre of the track, or type of music accompaniment. For the solo instrument, only pitched instruments are considered and no attempt is made to separate percussive instruments. We focus on the monophonic case, where the solo instrument is assumed to play only one note at a time. Only the single-channel separation problem is considered and no panning or spatial location information is used for separation. The algorithm should be lightweight and processing times should be minimized to allow its use in real-world applications.*

For the development of these studies, different datasets have been compiled always attempting to have a sample as general as possible, but somehow limited by the availability of multi-track recordings that allow proper evaluation of separation results. As the human voice and musical instruments can have very different acoustic and spectral characteristics, both vocal and instrumental solos are included in the datasets. In each section, the relevant datasets used for evaluation are explained in detail.

For evaluation of results, the PEASS Toolkit introduced in Section 2.4.3 is used. This choice was made based on the fact that current public evaluation campaigns such as the Signal Separation Evaluation Campaign (SiSEC) have chosen this set of measures to present their results, and as such, the evaluations presented in this thesis could be used as reference for future works.

As the goal of this study is to develop a method that can produce solo and backing tracks of similar quality, in all the tests conducted through this chapter, independent evaluations of solo and backing tracks are always conducted. Additionally, and taking into consideration the acoustical and spectral difference between musical instruments and the human voice, the evaluation is also conducted independently for vocal and instrumental tracks. This distinctions in the evaluation allows a thorough analysis of results, a better understanding of the differences between the signals, and more specific conclusions specific to each signal case.

Notation

For the remainder of this thesis, the following notation applies: let $f(t)$ be a monaural signal representing the audio mixture. The mixture $f(t)$ is assumed to be the sum of a monaural signal $s(t)$ representing the solo instrument, and a monaural signal $a(t)$ representing the accompaniment such that:

$$f(t) = s(t) + a(t) \tag{3.1}$$

The complex-valued spectrogram of the mixture obtained by means of the Short Time Fourier Transform (STFT) is given by $F(k, n)$. The complex-valued spectrogram is also assumed to be the sum of the complex-valued spectrograms of the solo and accompaniment signals:

$$F(k, n) = S(k, n) + A(k, n) \tag{3.2}$$

where $S(k, n)$ is the complex valued spectrogram of the solo signal, and $A(k, n)$ the complex valued spectrogram of the accompaniment. The index k will always denote frequency bins while the index n will denote time frames in the time-frequency representation. The window length will always be denoted by N, and the sampling frequency by fs. For simplicity of notation, the magnitude spectrogram of the mixture will be denoted $M(k, n)$. The magnitude spectrogram of the solo and accompaniment signals will be denoted by $|S(k, n)|$, and $|A(k, n)|$ respectively.

3.1. Frame-based Solo and Accompaniment Separation

The algorithm described in this section attempts to separate the solo instrument from its music accompaniment by estimating on a frame by frame basis, a spectral representation of the solo instrument given a fundamental frequency value $f0$. In Figure 3.1, a block diagram of the algorithm is shown for reference.

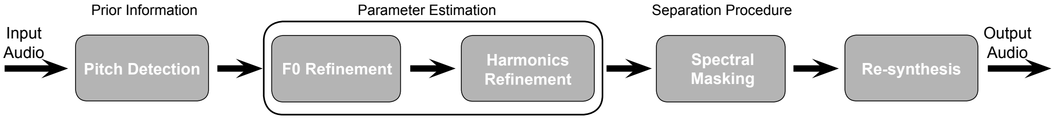


Figure 3.1.: Block diagram of the proposed frame-based solo and accompaniment separation algorithm. After an initial pitch detection stage, parameter estimation is composed of an F0 Refinement stage and a Harmonics Refinement stage. To complete separation a spectral masking approach is proposed.

As explained in Section 2 of this thesis, the separation process is generally composed of three main stages: parameter estimation, inclusion of prior information (when used), and a final separation stage. As shown in Figure 3.1, prior information is used in this algorithm in the form of $f0$ sequences of the solo instrument extracted by a pitch detection front-end. The pitch sequences obtained as prior information are used in the parameter estimation stage to obtain a frame-wise spectral representation of the solo instrument. Parameter estimation is composed of an initial $f0$ refinement stage followed by a harmonics refinement stage. Finally, separation is conducted using a masking approach. The details of the different processing stages of the proposed algorithm are further explained in the next sections.

3.1.1. Pitch Detection

As pitch detection front-end, the algorithm proposed in [45] is used. This method was already described in Section 2.3.1 of this thesis and only the main processing steps are mentioned here for reference. The algorithm extracts pitch information in four different stages: (1) Spectral representation, (2) Pitch candidate detection and tone formation, (3) Voice formation, and (4) Main melody selection. The idea behind this method is to use a multi-resolution Fast Fourier Transform (FFT) to extract pitch candidates based on a pair-wise evaluation of spectral peaks. After an initial peak detection stage, each peak is either assigned to an existing voice or starts a new one. Each voice is characterized by its mean magnitude and frequency range. A tone is assigned to a voice if it falls in its frequency range and passes its magnitude threshold. The most salient voice is selected as the main melody.

During pitch extraction, an analysis frame of 46 ms is used in conjunction with a hopsize of 5.8 ms. The pitch detection algorithm returns fundamental frequency sequences of the main melody on a frame by frame basis. Let $f0(n)$ be the frame-wise sequence of fundamental frequency values returned by the pitch detection algorithm. For those frames where no melodies are detected (silent frames) or where the solo instrument is assumed to be silent (only the accompaniment is playing), an $f0(n) = 0\,\text{Hz}$ is returned.

3.1.2. F0 Refinement

To further refine the fundamental frequency values delivered by the pitch detection algorithm, a refinement stage is proposed where the magnitude spectrogram is interpolated in a narrow band around each $f0(n)$ and its constituent harmonics. For a given $f0(n)$, low and high quarter tone deviations in Hz are given by:

$$f^{\downarrow}(n) = f0(n)/2^{(50/1200)} \tag{3.3}$$

$$f^{\uparrow}(n) = f0(n) \cdot 2^{(50/1200)} \tag{3.4}$$

Expressions (3.3) and (3.4) are derived from the definition of cent, the logarithmic unit of measure for music intervals. Here, a semitone spans 100 cents and an octave spans 1200 cents. The definition of cent states that given an initial frequency value f_1, and the number of cents covering the desired ascending interval c the frequency of the second note in the interval is given by $f_2 = f_1 \cdot 2^{(c/1200)}$. For descending intervals, the expression simply changes from a multiplication to a division: $f_2 = f_1/2^{(c/1200)}$. Given that a semitone spans 100 cents (and thus, a tone spans 200 cents), quarter tone intervals are represented by a difference of 50 cents, which explains Equations (3.3) and (3.4). The choice of cents was made because as a logarithmic unit of measure, interval calculations can be made in a simple and musically meaningful manner.

For each partial with partial index $p = 1, \ldots, p_{max}$, the same strategy is used to calculate the boundaries of its interpolation band. The initial (ideal) frequency location in Hz of the partial with partial index p is given by $f_p(n) = f0(n) \cdot p$, and the boundaries of the interpolation band are calculated with Equations (3.3) and (3.4).

The interpolation band around each partial is given by $[f_p^{\downarrow}(n), f_p^{\uparrow}(n)]$. For a given window size N and a sampling frequency fs, the frequency bin $k(n)$ where a frequency value $f(n)$ falls is given by $k(n) = \lfloor f(n) \cdot \frac{N}{fs} \rceil$. Here, the notation $\lfloor x \rceil$ indicates the round operator. The interpolation band expressed in frequency bins is given by $[k_p^{\downarrow}(n), k_p^{\uparrow}(n)]$. Let $k_p^{\downarrow}(n)$ and $k_p^{\uparrow}(n)$ be the frequency bins of $f_p^{\downarrow}(n)$ and $f_p^{\uparrow}(n)$, respectively.

For each time frame n, linear interpolation is used to refine the location of each partial within its interpolation band. Here, $i = 1, \ldots, i_{max}$ is the interpolation step. Each interpolation step i results in a new magnitude value $M(k_p^i(n), n)$ for each partial, with $k_p^i(n)$ the frequency bin (not necessarily integer) of partial p in the interpolation step i. The value of $k_p^i(n)$ is not necessarily an integer because the interpolation is conducted between two integer frequency bins $[k_p^{\downarrow}(n), k_p^{\uparrow}(n)]$.

A cumulative magnitude sum $E_i(n)$ is obtained for each interpolation step i. The optimal frequency bin $k_{opt}(n)$ is taken as the one that maximizes $E_i(n)$:

$$k_{opt}(n) = \underset{i}{argmax}\ (E_i = \sum_{p=1}^{p_{max}} M(k_p^i(n), n)) \tag{3.5}$$

The new refined fundamental frequency $\hat{f}0(n)$ in time frame n is simply obtained by:

$$\hat{f}0(n) = \lfloor k_{opt}(n) \cdot \frac{fs}{N} \rceil \tag{3.6}$$

Again here, the notation $\lfloor x \rceil$ indicates the round operator. This approach assumes that all the partials deviate from their harmonic locations by exactly the same factor.

3.1.3. Harmonic Series Refinement

To estimate the complete harmonic series of the solo instrument, the location of each one of the harmonic components is also refined. Two principles are followed at this stage: (1) Each harmonic component is allowed to have an *independent* deviation from its ideal harmonic location, i.e., multiple integer of the fundamental frequency. (2) The acoustic differences between the voice, strings, and wind instruments are considered when harmonic components are located. Namely, inharmonicity characteristics differ between instrument families. As explained in Section 2.6.2, string instruments tend to show harmonic components at frequencies slightly higher than the calculated harmonic ones. In contrast, wind instruments tend to show deviations to frequencies which are lower than the harmonic ones. The voice is in general assumed to be harmonic. Consequently, the harmonic estimation stage is kept consistent with either lower or higher deviations from harmonic locations, but never a mix of both.

Let $k_p(n)$ be the frequency bin of the ideal partial location of partial p (calculated as integer multiple of the fundamental frequency). To keep control of harmonic deviations, each partial is allowed a maximum deviation ρ_{max} from its harmonic location $k_p(n)$ of one quarter tone. This will guarantee that tones will remain perceptually harmonic. For each time frame n a frequency band given by $[k_p(n) - \delta_{max}, k_p(n) + \delta_{max}]$ is defined to conduct an iterative search in the vicinity of the ideal partial location $k_p(n)$ for all partials with partial index $p = 2, \ldots, p_{max}$.

For each partial index p, the search returns the frequency bin $\hat{k}_p(n)$ where the observed harmonic with the largest amplitude is detected. A detection mask $D(k,n)$ where the observed harmonics are marked with 1 for each time frame, is defined for k in the $[1, N/2]$ range:

$$D(k,n) = \begin{cases} 1 & \text{for } k \in \{\hat{k}_p(n), p = 1, \ldots, p_{max}\} \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

A simple smoothness constraint δ_k is set in the refinement stage in the attempt to avoid sudden frequency bin jumps in the harmonic estimation. The assumption here is that the frequency variation of a tone (and thus, of its harmonic components) in the time interval spanned by two processing frames should be relatively small and as such, large frequency bin jumps between two frames are unlikely. For each time frame n, the frequency bin $\hat{k}_p(n)$ where each partial p appears is stored in memory. In the next time frame $n+1$, this information is used to guarantee that the time-frequency bins assigned to each harmonic are a maximum of $\delta_k = 2$ bins apart. When the refinement stage finds a frequency bin $\hat{k}_p(n+1)$ with a larger bin jump than δ_k with respect to $\hat{k}_p(n)$, the smoothness constraint is enforced and the harmonic component is located in $\hat{k}_p(n) \pm \delta_k$. (the sign in the expression is determined by whether the deviation is to higher or lower frequency bins). This approach only requires the storage of a vector of length p_{max} which saves the frequency bin $\hat{k}_p(n)$ of each harmonic component in time frame n.

A clear distinction needs to be made between the smoothness constraint δ_k and the allowed frequency deviation ρ_{max}. While ρ_{max} allows harmonic components to deviate from their calculated harmonic location, δ_k guarantees that this deviation is relatively smooth over time. Here, only the previous time frame is used to impose smoothness and as such, it could be understood as a low-pass filter of length $L = 2$.

3.1.4. Spectral Masking

After the complete harmonic series that represents the solo instrument has been estimated, a binary mask for the solo $Z_S(k,n)$, and a binary mask for the accompaniment $Z_A(k,n)$ are calculated. To compensate for spectral leakage in the time frequency transform, a tolerance band Δ centered at the estimated location $\hat{k}_p(n)$ is included in the masking procedure. Thus, for a frequency range $[\hat{k}_p(n) - \Delta, \hat{k}_p(n) + \Delta]$ and time frame n we have:

$$(Z_S(k,n), Z_A(k,n)) = \begin{cases} (1,0) & \forall k, n \,\text{with}\, D(k,n) = 1 \\ (0,1) & \text{otherwise} \end{cases} \tag{3.8}$$

As in any binary masking procedure, the underlying assumption is that the energy in each time-frequency bin mainly belongs to one sound source and as such, each bin will be assigned to one source only in the masking procedure. With this in mind, (3.8) is equivalent to: $Z_A(k,n) = 1 - Z_S(k,n)$.

3.1.5. Re-synthesis

To obtain time domain signals for the solo and accompaniment, the complex valued spectrogram of the mixture is masked to obtain estimates of the complex spectrograms of the solo and accompaniment :

$$\hat{S}(k,n) = F(k,n) \odot Z_S(k,n) \tag{3.9}$$

$$\hat{A}(k,n) = F(k,n) \odot Z_A(k,n) \tag{3.10}$$

where $\odot$ denotes the Hadamard product. The estimated solo and accompaniment tracks are given by the Inverse Short Time Fourier Transform (ISTFT) of the masked spectrograms:

$$\hat{s}\,(t) = ISTFT\left(\hat{S}(k,n)\right) \tag{3.11}$$

$$\hat{a}\,(t) = ISTFT\left(\hat{A}(k,n)\right) \tag{3.12}$$

3.1.6. Experiments and Results

In this section, the performance of the frame-based separation algorithm is evaluated. The details of the dataset used, the evaluation criteria and the conclusions drawn from the experiments are presented in the following sections.

3.1.6.1. Implementation Details

For the experiments described in this section, the following algorithm parameters were used: in the $f0$ refinement stage, $i_{max} = 50$ iterations were used in the interpolation stage. As the assumption that partials deviate by the same factor from their harmonic locations does not necessarily hold true for higher partials, only the $p < 5$ partials were used for interpolation. In general, this assumption within this range can be loosely held as true. For the harmonic series refinement the total number of partials was set to $p_{max} = 25$. This choice was made given that informal observations show that wind instruments in particular can exhibit up to 25 clear harmonic components. A tolerance band $\Delta = 1$ was used for spectral binary masking. An analysis frame of 46 ms in conjunction with a hop size of 5.8 ms were used, given a sampling frequency $fs = 44.1$ kHz.

3.1.6.2. Dataset

For the evaluation of the algorithm, a dataset of 27 tracks was compiled. For the remainder of this thesis, this dataset will be referred to as **SA_DS1** (Solo and Accompaniment DataSet 1). A full description of the tracks used in the evaluation, as well as the copyright and availability information is provided in Appendix B. Here, only a general description is provided. The dataset is composed of two parts: (1) SA_DS1a with copyright free tracks. A total of 17 tracks, 10 with voice as main instrument, and 7 instrumental tracks are included. (2) SA_DS1b composed of 10 commercial instrumental tracks with saxophone as the solo instrument.

All the tracks in SA_DS1a were mixed from the available multi-track recordings into a solo signal, an accompaniment track, and a final mixture. Solo and accompaniment tracks were already available in the commercial distribution of SA_DS1b.

3.1.6.3. Analysis of Results

The Perceptual Evaluation Methods for Audio Source Separation (PEASS) Toolkit was used for evaluation of separation results. Mean values with 95% confidence intervals for the entire dataset are presented in Figure 3.2. For reference purposes and to allow future comparison of results, the full table of results with independent scores for each track in the dataset are presented in Appendix C of this thesis.

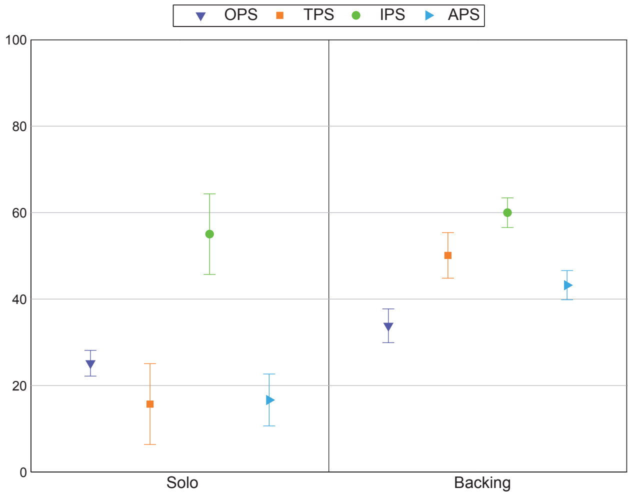


Figure 3.2.: Results obtained with the frame-base separation algorithm for the entire dataset. Objective perceptual quality measures: Overall Perceptual Score (OPS), Target-Related Perceptual Score (TPS), Interference-Related Perceptual Score (IPS), Artifact-Related Perceptual Score (APS). Mean values with 95% confidence intervals are presented. It can be observed that the algorithm results in particularly high IPS scores for both the solo and the backing tracks. The backing tracks show in general more homogeneous scores than the solo tracks where clear differences between the IPS and TPS scores can be observed.

For the solo tracks, an Overall Perceptual Score (OPS) of 25.07 was obtained. For the backing tracks, slightly better results were obtained with an OPS of 34.55. The high Interference-related Perceptual Scores (IPS) scores obtained for both the solo and backing tracks are particularly noticeable, being in both cases, the highest score from the four measures. Given that the IPS measures the distortion in the signal produced by interferences from other sources, these results suggest that the iterative estimation of the algorithm is very effective in detecting spectral information that corresponds specifically to the solo instrument.

For the solo tracks however, a much higher variation in the IPS is observed, resulting in a much larger confidence interval than for the backing IPS. These results evidence that the algorithm can better handle some musical instruments than others.

Particularly noticeable are also the lower Target-related Perceptual Scores (TPS) and Artifacts-related Perceptual Scores (APS) obtained for the solo instrument in comparison to the backing tracks. The TPS for the solo also showing large confidence intervals that evidence the great variability in the results. These low TPS and APS scores for the solo instrument suggest that while the estimation stage is very effective in detecting spectral information that belongs to the solo, its estimation might be too restrictive and does not guarantee the continuity and smoothness expected in spectral representations of harmonic sources. This naturally results in more artifacts in the solo signal and in the loss of important spectral information mistakenly assigned to the backing track. Simply stated, results suggest that the information detected by the algorithm as belonging to the solo, does belong to the solo in most cases (and thus, high IPS); however, there is valuable information from the solo being assigned to the backing track (and thus, low APS and TPS for the solo). Given that some information from the solo is being mistakenly assigned to the backing track, it could be expected that low IPS scores would be obtained for the backing tracks. However, this is not the case. These results can be understood taking into consideration the fact that backing tracks are composed of several musical instruments playing together, including percussive instruments. This naturally poses different conditions for the backing tracks as minor artifacts and interferences from the solo might be somehow masked by the information from other instruments. Consequently, quality measures for the backing tracks are not so strongly affected by mis-estimations as they are for the solo.

A final remark can be made regarding the low APS obtained for the solo signals. It should be noted that the frame-based algorithm performs separation by applying binary masks to the complex valued spectrogram of the original mixture (See Section 2.1.1 for a definition of binary masks). While binary masking allows very efficient processing, it also results in separated signals with more artifact distortions in comparison to soft time-frequency masking from a generalized Wiener filtering approach. Artifacts often appear due to sudden discontinuities in the magnitude or phase spectrograms produced by the binary approach. Soft Wiener masks act as low-pass filters, smoothing the resulting signals and reducing artifacts at the expense of more interference between the sources. This approach can often result in reduced sharpness of the solo signals and requires longer processing times. However, depending on the application, it can be an option to reduce artifact distortions.

As stated in the introduction of this chapter, the different spectral characteristics of musical instruments and the human voice call for independent evaluations which are now presented in Figure 3.3. For these figures, mean values with 95% confidence intervals where calculated by dividing the dataset (SA_DS1) into vocal and instrumental tracks. This distinction is only made to better assess the performance of the algorithm when dealing with different types of signals.

It can be seen that even when the OPS obtained for the instrumental and vocal solo tracks are very similar, 26.16 and 23.21, respectively, very different behavior is observed for the APS and TPS of the two types of signals. Vocal solo tracks exhibit considerably lower TPS and APS when compared to the instrumental solo tracks. These results most likely come from the fact that the algorithm fails to capture unvoiced sounds present in vocal signals (see Section 2.6.2 for a definition of unvoiced sounds). The estimation algorithm assumes completely harmonic sources at all times, and while this assumption holds in most cases for the sustained parts of musical instrument tones, fricatives and plosives in vocal sounds evidence broadband spectral characteristics that the algorithm fails to handle. It is also evident from the results in Figure 3.3a that the solo tracks for the instrumental database still result in considerable large confidence intervals. Considering that the database used contains solo tracks from different musical instruments, this variability can be expected. Even when general assumptions about musical instrument tones can be used for the estimation, particular differences between brass and string instruments,

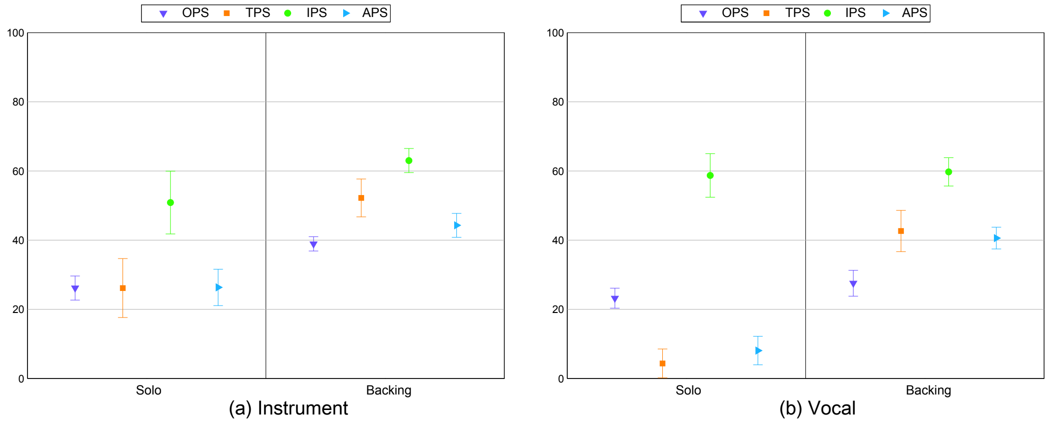


Figure 3.3.: Results obtained with the frame-based separation algorithm displayed independently for instrumental and vocal tracks. (a) Instrument (b) Vocal. Objective perceptual quality measures: Overall Perceptual Score (OPS), Target-Related Perceptual Score (TPS), Interference-Related Perceptual Score (IPS), Artifact-Related Perceptual Score (APS). Mean values with 95% confidence intervals are presented. Specially for the solo tracks, performance differences between the vocal and instrumental datasets can be clearly observed. Vocal solos show in general lower APS and TPS than the instrumental solos. Results for the backing tracks for the two datasets do not show such big differences in performance as the solo tracks.

for example, or strings and woodwinds, are not properly captured by the algorithm. In the case of vocal tracks (Figure 3.3b), confidence intervals are smaller and results are more consistent among the different signals.

Even when the evaluation using the set of objective perceptual measures is informative, very valuable observations can also be made by simply listening to the resulting separated tracks. Here, a series of observations made by informal listening tests are presented. It is very evident from the resulting backing tracks that the attacks of the tones are not being properly captured by the estimation of the solo signal. This causes, in some cases, very audible artifacts in the backing tracks where the remaining attacks that should have been assigned to the solo signal, are kept as information belonging to the accompaniment. A second observation refers to the sometimes very clear interference from percussion instruments in the solo tracks. In many cases and specially when percussion hits coincide in time with the beginning of a tone, very clear percussive sounds can be heard in the separated solo signal. Finally, backing tracks obtained for the vocal database often have audible noise-like sounds that belong to the unvoiced parts of the singing voice. Specially noticeable is a whisper-like sound with clear voice components, constantly present in the backing tracks.

3.1.6.4. Conclusions

The results presented in this section represent a compromise consciously taken between separation quality and algorithm efficiency. While performing a frame-based estimation allows very fast processing, only simple harmonic modeling constraints can be set.

> *The frame-based separation algorithm results in Overall Perceptual Scores (OPS) for the solo and backing track of 25.07 and 34.55, respectively. These results show that the goal of obtaining solo and backing tracks of similar quality is being reached. The algorithm has an average processing time of* $0.2 \cdot l$*, where* l *is the total length of the song, on a 2.66GHz computer, making it suitable for real world applications and real-time performance.*

With this in mind, important conclusions were drawn from the analysis:

- While the algorithm is very effective in the estimation of the solo instrument (evidenced by the high IPS obtained), results suggest that the algorithm could benefit from estimations that consider longer time intervals and allow the inclusion of more complex models of musical instrument sounds.
- The differences in performance between vocal and instrumental tracks are very evident. The algorithm fails to capture unvoiced sounds in vocal signals, and instrumental tracks show large confidence intervals that evidence variability in the results.

The natural question that comes to mind is: *How can the results of the algorithm be improved while maintaining the initial design goal of having a lightweight system that can handle different types of solo instruments?* In Section 3.2 several studies are presented which were conducted with the goal of better characterizing the spectral behavior of musical instrument sounds. Magnitude, frequency, and phase of musical instruments are studied to draw effective measures to improve separation results. In Section 3.3, the findings from these studies are applied to a new *Tone-based Separation Algorithm.*

3.2. Analysis of Spectral Parameters of Musical Instrument Signals

In Section 3.1, a frame-based separation algorithm was proposed to address the solo and accompaniment separation problem. Several important conclusions were drawn from the results obtained in that section; namely, the need to better characterize musical instrument spectra to obtain more accurate representations of the sources to be separated. In this section, several studies are conducted in the attempt to get a better understanding of the spectral characteristics of different musical instruments. The different studies and relevant datasets compiled for each analysis are described in the next sections.

3.2.1. Partial Tracking Method

In this section, a partial tracking method is proposed to extract magnitude, phase, and frequency of the harmonic components of a solo signal. The goal of this method is to extract accurate spectral information from isolated instrument signals, to be used in the analyses described in sections 3.2.2 and 3.2.3. The extraction process is composed of three stages: (1) Energy Detection, (2) Peak Detection & Pruning, and (3) Trajectory Formation. A block diagram of the proposed partial tracking method

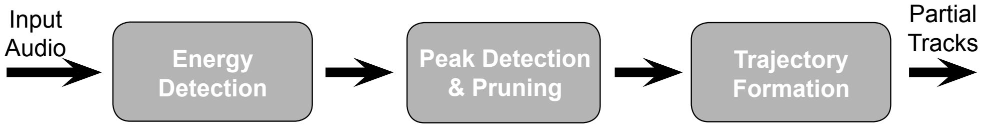


Figure 3.4.: Block diagram of the partial tracking method proposed. This method is used in the following sections to study the spectral parameters of musical instruments.

is shown in Figure 3.4 for reference. The main processing blocks of the algorithm are described in the following sections.

Energy Detection

At this stage, possible silent frames at the beginning and end of the signal are removed to make sure that the partial tracking algorithm only receives frames where the tone is active. To detect silent frames, the energy of the time domain signal $x(t)$ of length T is calculated:

$$\varepsilon = \sum_{t=0}^{T-1} |x(t)|^2 \tag{3.13}$$

To avoid excessive segmentation of the signal due to frames of low energy, a minimum silence length threshold of $\gamma = 80\,\text{ms}$ is defined. To avoid including clicks or bursts that might occur at the beginning or end of the signal, a minimum sound length threshold is also defined. Only segments of sound longer than $\rho = 200\,\text{ms}$ are used in the peak detection stage.

Let L be the number of samples of the minimum allowed silence $L = \gamma \cdot fs$. A moving average filter of length L is used to calculate the mean energy within segments of length γ. The mean energy value at time t is given by:

$$\bar{\varepsilon}(t) = \frac{1}{L} \sum_{j=0}^{L-1} \varepsilon(t-j) \tag{3.14}$$

The mean energy $\bar{\varepsilon}(t)$ is normalized to a [0 1] range and any segments whose energy falls below the energy threshold $\varphi_e = 0.015$ are discarded. Finally, only the remaining energy segments that pass the minimum segment length threshold ρ are kept.

The new boundaries of the signal which define the onsets and offsets of each of the tones, are then passed to the peak detection stage.

Peak Detection & Pruning

A STFT analysis with a 46 ms frame size and a 5.8 ms hop size is used as time-frequency representation. For each time frame n, the magnitude spectrogram is analyzed to extract relevant peaks that might correspond to partials of the given tone. Even when this method is meant to be used with isolated solo signals, care needs to be taken when building partial tracks. Some instruments such as the flute, present noise-like elements in their spectra that can appear as spectral peaks of different frequencies and amplitudes. Additionally, it is not uncommon for musical instruments to show subharmonic peaks. Even when these subharmonic components might exhibit particular characteristics of each instrument, they are not considered in these studies and thus, were excluded in the peak detection stage. Figure 3.5 shows the magnitude spectrogram of two example tones with visible peaks that do not correspond to harmonic components. In the figure, expected partial peaks are shown in blue while subharmonic components are shown in red.

To guide the peak detection algorithm and to avoid too many spurious peaks, an initial estimate of the fundamental frequency $f0$ of the tone is always provided to the peak detection algorithm. In the cases where the signal analyzed is a single note, the initial estimate of the fundamental frequency is taken from the metadata of the dataset. If metadata is not available or signals where the solo instrument plays a full melody are used (as in Section 3.2.2), an initial pitch detection stage is performed to obtain the corresponding $f0$ values. In such cases, the pitch detection algorithm

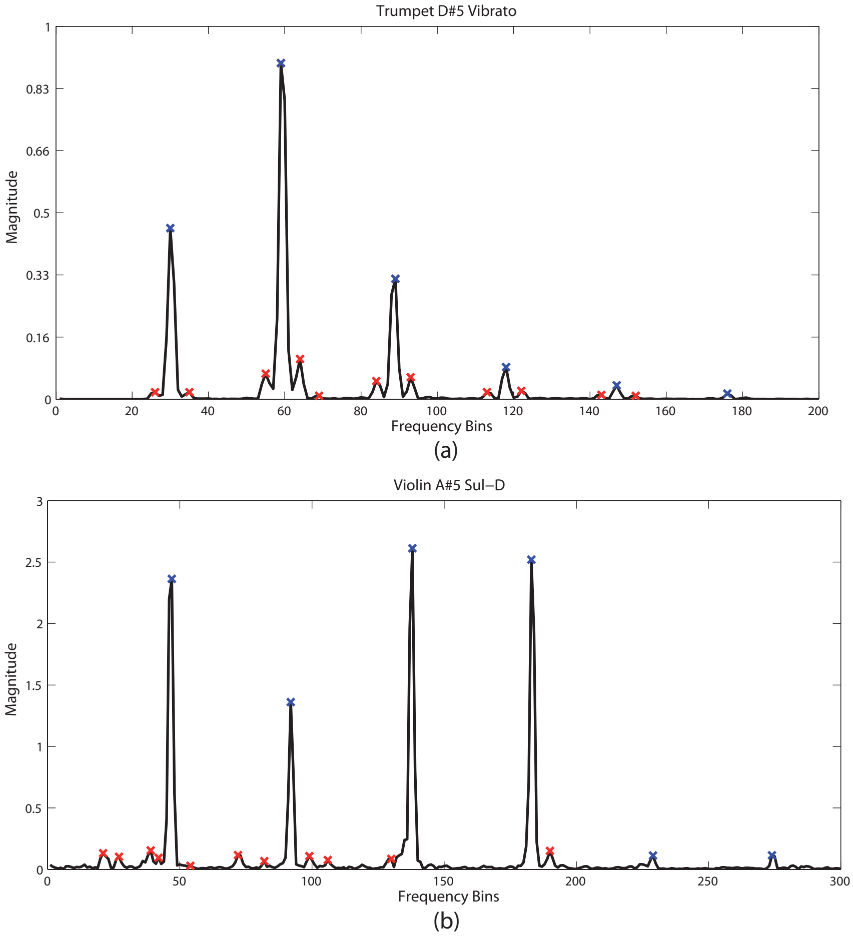


Figure 3.5.: Examples of magnitude spectrograms of instrument tones that exhibit subharmonic components. Subharmonic peaks are marked with red while harmonic peaks are marked with blue. (a) Trumpet D#5 Vibrato (b) Violin A#5 Sul-D (The notation Sul-D refers to the string where the note is played on the violin)

described in Section 3.1.1 is used. The goal of having a fundamental frequency value before peak detection is to restrict the search to frequency regions where partials are expected. The frequency range where the search is conducted is restricted to $[f0/2^{(100/1200)}, fs/2]$. The lower bound corresponds to a semitone lower than the

$f0$ defined in cents. The higher frequency bound is given by the Nyquist frequency, with fs the sampling frequency. In the defined frequency range, an initial peak detection stage is conducted where all the frequency bins k where local maxima in the magnitude spectrogram are detected, are saved.

To remove spurious peaks from the initial peak candidates, a frequency adaptive magnitude threshold φ_f is defined. A frequency delta $\Delta_f = 50\,\text{Hz}$ is defined such that the frequency band where φ_f is calculated, is given by: $[f_k - \Delta_f, f_k + \Delta_f]$. Here, f_k is the frequency in Hz of the bin k where the peak was detected. Given an analysis window of length N and a sampling frequency fs, low and high frequency bins where the threshold is defined are given by:

$$k_L = \lfloor \frac{N}{fs} \cdot (f_k - \Delta_f) \rceil \tag{3.15}$$

$$k_H = \lfloor \frac{N}{fs} \cdot (f_k + \Delta_f) \rceil \tag{3.16}$$

with $\lfloor x \rceil$ the round operator.

For a given time frame n, the magnitude spectrogram of the tone is denoted by $M(k,n)$. The frequency adaptive magnitude threshold is calculated as follows:

$$\varphi_f = \frac{1}{K} \sum_{k=k_L}^{k_H} M(k,n) \tag{3.17}$$

where K is the total number of frequency bins included in the $[k_L, k_H]$ band. Any peak within the band whose magnitude falls below the threshold is eliminated.

Finally, only a maximum number of peaks $p_{max} = 25$ are kept in each frame. In the cases where more than p_{max} peaks remain after the pruning stage, only the p_{max} peaks with the largest amplitude are kept.

The peak detection stage restricts the maximum number of harmonics that can be found; however, a minimum number is not imposed. This approach makes the detection stage more robust to spurious peaks and accounts for the fact that for higher fundamental frequency values, fewer harmonics are visible in the spectrogram.

For those tones where fewer peaks than p_{max} are found, the remainder of the buffer is filled with zeros. The resulting peaks correspond to rough, unordered peak candidates of the harmonic components of each frame in the tone.

Trajectory Formation

The peak detection stage delivers a set of peak candidates for all the time frames n of the signal. The peak candidates are delivered as a $p_{max} \times M$ matrix, with $p_{max} = 25$ partials, and M the total number of frames in the tone. It must be noted that zero entries in the peak candidate matrix indicate time frames where less than p_{max} peak candidates were detected. The goal at this stage is to organize the peak candidates into partial trajectories that represent the time evolution of each partial over the length of the tone.

To be able to assign peak candidates to a certain partial of a tone, it is necessary to first define a frequency range where the search for prominent peaks is conducted. A rough estimate of the frequency of the partials is given by their ideal harmonic locations, calculated as integer multiples of the fundamental frequency: $f_p = f0 \cdot p$, with $p = 1, \ldots, p_{max}$ the partial index. For each partial p, a frequency band centered at the ideal harmonic location is defined using the cents unit of measure: $[f_p/2^{(150/1200)}, f_p \cdot 2^{(150/1200)}]$. In each time frame n, the peak candidate matrix is searched for peaks that fall within the defined frequency band of each partial. If for a time frame n, more that one peak is found within the given frequency band of a partial, the peak that minimizes the frequency difference with the peak found in the previous frame $n-1$, is taken as the partial peak. This measure guarantees continuity of the partial tracks and reduces the possibility of selecting subharmonic components. If no peak fulfills the frequency bands, the partial track is set to zero.

At the end of this stage, frequency trajectories in time are delivered for each partial. To facilitate further analyses, results are delivered again as a $p_{max} \times M$ matrix where each row of the matrix corresponds to the temporal trajectory of one partial. In this particular case, frequency bin trajectories are delivered; namely, each of the entries in the matrix contains the frequency bin k where each partial was found in a given time frame.

Algorithm Performance

The performance of the partial tracking method described in this section was evaluated with a set of manually annotated instrument tones. A total of 20 instrument tones taken from the University of IOWA Musical Instrument Dataset [109] were manually annotated on a frame by frame basis for the first 15 observed partials. The instruments used in the evaluation were piano, violin, trumpet, flute, and alto saxophone. For the evaluation, a window length of 2048 samples and a hop size of 256 samples were used. All signals had a sampling frequency $fs = 44100\,\text{Hz}$. The dataset was processed with the partial tracking algorithm and Precision p, Recall r, and F-Measure F values were calculated. The proposed method resulted in $r = 0.7539$, $p = 0.9698$, and $F = 0.8398$ and showed to be suitable for robust detection of partial tracks from musical instrument recordings.

3.2.2. Contribution of Spectral Parameter in Separation Quality

During the many years of sound separation research, many techniques have been proposed to address the separation problem. As described in Chapter 2, many algorithms have focused on getting an accurate representation of the spectral magnitude of the individual sources. Some approaches have relied on assumptions such as smoothness of the spectral envelope, others have assumed a certain spectral shape for the different musical instruments, while others have enforced common amplitude modulation among harmonic components. When it comes to frequency information, the assumption of perfect harmonicity in musical tones has been made in many systems. Perfect harmonicty assumes that the frequencies of the partials of a tone are given by integer multiples of the fundamental frequency $f0$. Some systems have relaxed the perfect harmoncity assumption but to the author's knowledge, no structured study has been conducted within the separation community to better characterize inharmonicity patterns of musical instruments. When it comes to phase information, very few studies have addressed its effects on separation methods. It has in general been the case that phase information is left untouched and the original phase from the mixture is taken to re-synthesize the different sources. A notable exception is the work of [16] on spectrogram consistency.

Given the results obtained with the frame-base separation algorithm in Section 3.1, several open questions remained:

Which stage of the estimation process of the algorithm should be improved to optimize the quality of separation? Should the optimization focus on getting better estimates of the magnitude of the sources, or should more attention be given to frequency and phase information?

After many years of separation research, very little effort has been made to assess the impact of different spectral parameters such as phase, magnitude, and frequency location of harmonic components in the resulting quality of the extracted signals. The development of objective perceptual measures for separation quality assessment has considerably facilitated the systematic study of such impact. For a given separation approach, the effects of different algorithm parameters can be assessed in a very straightforward manner and guidelines for algorithm optimization can be drawn. In the next section, a study that attempts to describe the influence of phase, magnitude, and frequency location of harmonic components in the perceived quality of separated solo signals, is presented.

3.2.2.1. Experiments and Results

In this study, the effects of magnitude, phase, and frequency location of harmonic components on the perceived quality of separated tracks, was evaluated. The main goal of this study was to understand the impact on separation quality of having very accurate spectral information. This study evaluates independently the impact of the three spectral parameters: magnitude, frequency and phase.

To obtain the necessary spectral information for the study, solo signals from the chosen dataset were processed to extract frame-wise magnitude, phase, and frequency information of the harmonic components. The method for partial tracking described in Section 3.2.1. was used for this purpose.

To better assess the effect of each of the spectral parameters, different signal versions of the solo were created for each of tracks in the dataset. For each signal version, the spectral parameters were slightly modified before re-synthesis. Table 3.1 shows the characteristics of each signal version. The first signal in each table, i.e., *Name_FB*, is the solo signal obtained with the frame-based separation algorithm described in Section 3.1. The label *Original* was given to the parameters obtained from the original solo track from the multi-track recordings using the partial tracking method described in Section 3.2.1. The label *Estimate* was given to each of the parameters obtained with the frame-based separation algorithm. In should be noted that in this algorithm, no attempt was made to estimate the phase of the separated signals. The phase of the original mixture is conservatively taken as the phase of both the solo and accompaniment.

For the magnitude estimates, three additional variants were created: *exp. Estimate* which refers to the magnitude obtained with the frame-based separation algorithm additionally weighted with an exponential decay in frequency, *rand Estimate* which refers to the original magnitude modified by a random percentage within the [-1, 1] range, and *rand + exp Estimate* which refers to the original magnitude with the same random modification as the previous label with an additional exponential decay in frequency.

For each of the tracks in the dataset, ten signal versions of the solo instrument were synthesized using the parameters described in Table 3.1. For the evaluation, the available multi-track recording were used to create a solo, an accompaniment signal and the final mixture. The obtained mix was processed with the frame-based separation algorithm to obtain *estimates* of the spectral parameters. The clean solo signals from the multi-track recordings were used to extract the *original* spectral parameters.

Table 3.1.: Description of the spectral parameters used for each of the signal versions created.

Signal Name	Phase	Mag	Freq.
Name_FB	Estimate	Estimate	Estimate
Name_1	Original	Original	Estimate
Name_2	Original	Estimate	Estimate
Name_3	Original	exp. Estimate	Estimate
Name_4	Estimate	Original	Original
Name_5	Estimate	rand Estimate	Original
Name_6	Estimate	rand + exp. Estimate	Original
Name_7	Estimate	Original	Estimate
Name_8	Original	rand Estimate	Original
Name_9	Original	rand + exp. Estimate	Original

Dataset

A small dataset of three tracks was used for this experiment. This dataset is referred to as **SA_DS2**. As shown in Table 3.2, different solo instruments and genres were used for the study. Due to copyright restrictions, this dataset cannot be made publicly available; however, detailed information about the tracks is presented in Appendix B of this thesis. For all the signals, multi-track recordings were available.

Table 3.2.: Dataset **SA_DS2** description.

Track	Genre	Solo Instrument	Accompaniment
test1	Jazz	Alto saxophone	Drums, piano, bass
test2	Pop ballad	Male voice	Vocals, piano, bass
test3	Swing	Clarinet	Drums, piano, bass

Evaluation

The PEASS Toolkit was used to evaluate the quality of the synthesized solo signals. Resulting measures are presented for each of the signals in Table 3.3, Table 3.4, and Table 3.5, respectively. For each signal the Overall Perceptual Score (OPS), Target-related Perceptual Score (TPS), Interference-related Perceptual Score (IPS), and Artifacts-related Perceptual Score (APS), are presented.

Table 3.3.: Objective perceptual quality measures for signal test1: Overall Perceptual Score (OPS), Target-Related Perceptual Score (TPS), Interference-Related Perceptual Score (IPS), Artifact-Related Perceptual Score (APS). The highest scores are shown in red. The lowest scores are shown in blue.

Signal	**OPS /100**	**TPS /100**	**IPS /100**	**APS /100**
test1_FB	19.4461	25.5726	60.9986	15.5474
test1_1	55.328	40.7943	82.8943	52.1161
test1_2	**33.2821**	31.3511	**74.8165**	**28.2587**
test1_3	39.2252	**26.3295**	78.063	33.6584
test1_4	52.0785	51.8362	81.5127	50.3228
test1_5	50.6688	49.2281	80.7793	49.5745
test1_6	53.2174	39.3166	81.5976	51.9513
test1_7	47.1452	40.3439	79.752	45.1934
test1_8	60.1738	**60.248**	83.6985	58.7983
test1_9	**64.0188**	48.1209	**84.5643**	**62.2899**

Analysis of Results

For easier visualization of results, the lowest scores for each measure are shown in blue and the highest scores are shown in red in each of the tables. The scores obtained with the frame-based algorithm (*Name_FB*) were excluded from this ranking as they are always the lowest ones. The low scores obtained by the frame-based algorithm in comparison to the other signal versions are not surprising as it is the only signal version where all the synthesis parameters are estimates. The following remarks can be made by analyzing the results obtained:

- For all signals, the highest OPS was obtained by signal version *Name_9* which uses original phase and frequency of the harmonic components, and an exponentially weighted estimate of the magnitude.
- The lowest OPS scores were always obtained by *Name_2*, i.e., original phase, and estimate of the magnitude and frequency of the harmonics. For the three solo signals, *Name_2* also presents the lowest IPS and APS scores. It is important to point out that even with the lowest OPS scores, *Name_2* still shows a

Table 3.4.: Objective perceptual quality measures for signal test2: Overall Perceptual Score (OPS), Target-Related Perceptual Score (TPS), Interference-Related Perceptual Score (IPS), Artifact-Related Perceptual Score (APS). The highest scores are shown in red. The lowest scores are shown in blue.

Signal	**OPS /100**	**TPS /100**	**IPS /100**	**APS /100**
test2_FB	2.72197	8.46823	20.0746	0.87033
test2_1	16.2539	17.8368	62.3981	8.86316
test2_2	**7.28879**	9.51548	**46.4168**	**2.5511**
test2_3	7.6865	**9.40061**	46.9784	2.82705
test2_4	20.8346	48.5833	67.2165	13.6942
test2_5	20.1149	46.9775	66.5921	12.9841
test2_6	22.2277	36.1384	67.6809	15.3038
test2_7	16.2539	17.8368	62.3981	8.86316
test2_8	42.8365	**57.1865**	**80.5594**	35.4936
test2_9	**45.5591**	44.1082	80.4709	**40.033**

quality improvement in relation to *Name_FB* which supports the importance of refined phase information in the separation algorithm.

- The relationship between an accurate phase estimation and an accurate frequency location of the harmonic components appears to be very relevant to the quality of the solo signal as both *Name_8* and *Name_9* which are obtained with original phase and original frequency of the harmonic components, are always in the high end of the OPS ranking. These scores show the importance of fine frequency variations of the harmonic components in the perceived quality of a signal.

- To asses the impact of having very accurate estimates of the magnitude of the harmonic components, *Name_1*, *Name_4* and *Name_7* are analyzed as they are all obtained with the *Original* magnitude. *Name_7* always presents the lowest OPS of the three which accounts for the use of estimates both for the phase and harmonic locations. Additionally, the OPS of *Name_7* is always among the 3 lowest scores of each set which further points out the importance not only to concentrate on the spectral magnitude but to look at the interaction between

Table 3.5.: Objective perceptual quality measures for signal test3: Overall Perceptual Score (OPS), Target-Related Perceptual Score (TPS), Interference-Related Perceptual Score (IPS), Artifact-Related Perceptual Score (APS). The highest scores are shown in red. The lowest scores are shown in blue.

Signal	**OPS /100**	**TPS /100**	**IPS /100**	**APS /100**
test3_FB	15.756	33.4435	60.3932	9.46146
test3_1	47.8058	47.5321	80.7135	43.5196
test3_2	**35.0171**	**33.878**	**76.2494**	**28.6766**
test3_3	36.0682	34.6635	76.7815	29.7799
test3_4	55.1085	47.0971	81.2993	55.9969
test3_5	47.4624	**52.8007**	79.2465	46.8611
test3_6	51.5365	46.8204	80.3821	51.7322
test3_7	47.186	40.4	79.8656	44.2033
test3_8	53.6307	52.7405	81.3849	53.6137
test3_9	**60.0377**	48.3531	**83.0139**	**60.2893**

the different parameters. Both *Name_1* and *Name_4* show an improvement in OPS compared to *Name_7*; however, no consistent conclusion can be drawn from this test set regarding the difference in the impact of including original phase or original harmonic locations for synthesis.

These results suggest that even if the accuracy of the spectral magnitude estimates is comparable to the original magnitude, the OPS obtained will still be on the lower end of the scale. In contrast, *Name_8* and *Name_9*, which are both obtained with only estimates of the spectral magnitude, both obtained OPS scores in the highest end of the ranking. For the two types of magnitude estimates presented, the exponentially weighted outperforms in all cases the initial estimate, i.e., *Name_3* outperforms *Name_2*, *Name_6* outperforms *Name_5*, *Name_9* outperforms *Name_8*. Additionally, the use of exponential decay always results in lower IPS scores. Exponential decay by definition, lowers the contribution of high frequency components where estimation can be noisy and less accurate.

Conclusions

The analysis of results provided some insight on how to modify the proposed method to optimize quality of separated tracks. Even when this conclusions might specifically evidence the shortcomings of the frame-base separation algorithm evaluated, the conclusions can be used as rough guidelines for other separation approaches. The importance of refined estimates of phase and frequency of the harmonic components has shown to be necessary for improved separation quality, showing greater impact than the spectral magnitude. Having very accurate magnitude estimates showed little quality improvement; however, the use of exponential decay in frequency to weight magnitude estimates has proven to result in better separation quality. For a consistent improvement of the quality of resulting audio tracks, the development of sound separation algorithms should be directed to include and refine not only one of the given parameters, but to properly capture the interactions between them.

3.2.3. Analysis of Isolated Tones of Musical Instruments

The most part of the studies presented in this section were conducted from 03.2012-06.2012 during a research stay at the Center for Digital Music C4DM at Queen Mary University of London under the supervision of Prof. Mark Plumbley.

Results from the frame-based separation algorithm revealed that more accurate spectral estimates of the solo instrument were needed to improve quality of the separated tracks. The study conducted in Section 3.2 also evidenced the need not only to focus on more accurate magnitude estimations, but also to focus on possible refinements of the phase and frequency estimates of the harmonic components of the solo signal. To better understand the relationship of these spectral parameters, a study is proposed where isolated notes of different musical instruments are analyzed to characterize their behavior in the time-frequency domain. This study was designed with the goal of finding spectral characteristics of instrumental tones that can be used as guidelines for parameter estimation in the separation algorithm. The use of isolated notes naturally poses different acoustical and spectral conditions than the analysis of audio mixtures where several instruments overlap in time and frequency. Nevertheless, this study aims at understanding instrumental tones under ideal conditions to then analyze to what extent the same characteristics can be observed within an audio mixture.

The next sections present a description of the datasets and methods used for this analysis.

3.2.3.1. Dataset

For the magnitude and frequency studies (Sections 3.2.3.2 and 3.2.3.3) presented in this section, a selection of six musical instruments was made: 3 wind instruments —saxophone, Bb trumpet, and clarinet—, and 3 string instruments —violin, guitar, and piano—. All instrument samples were taken from the University of IOWA Musical Instrument Dataset [109]. For the alto saxophone, a total of 64 tones were processed, 32 with vibrato and 32 without vibrato. All the tones were played in *mf* and were in the Db3 - G#5 range. For the Bb trumpet, a total of 70 tones were processed, 35 with vibrato and 35 without vibrato. All the tones were played in *mf* and were in the E3 - D6 range. For the clarinet, a total of 35 tones played in *mf* without vibrato in the G3 - C7 range were processed. For the violin, a total of 88 tones in the G3 - B6 range were processed. All tones were played without vibrato in *mf* using the bow. Where possible, the different notes were played on different strings of the violin. For the guitar, a total of 117 tones in the E2 - B5 range were processed. All tones were played without vibrato in *mf*. As for the violin, the different notes were also played on different strings of the guitar. Finally, 33 notes were processed for the piano in the C3 - B5 range played in *mf*. This gives a total of 407 processed and analyzed tones all of them with a sampling frequency $fs = 44100\,\mathrm{Hz}$.

For the phase analysis (Sections 3.2.3.4), additional tones for the flute and the tenor trombone also taken from [109] were also processed. For the flute, a total of 76 *mf* tones were processed, 38 with vibrato and 38 without vibrato. All tones were in the B3 - Db7 range. For the trombone, a total of 33 *mf* tones in the E2 - C5 range were processed.

The method for partial tracking presented in Section 3.2.1 was used to extract magnitude, frequency, and phase information for all the tones in the dataset.

3.2.3.2. Magnitude Analysis

As described in Section 2.6.1, several studies have been conducted in the attempt to better characterize magnitude spectrograms of different musical instruments. However, results have shown that finding an accurate model that allows the prediction of the magnitude envelope of one partial based on the magnitude envelope of another partial of the same tone, or that allows the characterization of the shape of spectral and temporal envelopes of a given musical instrument, is a challenging task. To date, no robust solution has been proposed to address this estimation problem. Consequently, this study focuses on the characterization of the change in time of the temporal envelopes of musical instruments tones. The importance of statistically studying the change of magnitude over time, comes from the fact that it allows to characterize envelope smoothness of musical instrument tones. Envelope smoothness is an important parameter in the estimation of the spectral components of the target source in a separation context.

In this study, temporal envelopes $M_p(k, n)$ were obtained for each of the first 10 partials of the tones in the dataset. Here, p refers to the partial index, k and n are the frequency bin, and time index, respectively. To extract the envelopes, a window length of 2048 samples, and a hop size of 256 samples were used. All signals had a sampling frequency $fs = 44100\,\mathrm{Hz}$. The temporal envelopes were first normalized to the [0,1] range. Then, the first-order time derivative $\frac{d(M_p(k,n))}{dn}$ was obtained for each of the envelopes, and mean values and standard deviations were obtained for each of the partials.

Results of the analysis are presented in Figures 3.6 - 3.9. For the alto saxophone and trumpet where tones both played with and without vibrato were available, results are presented independently. For the clarinet, violin, piano, and guitar only tones without vibrato were studied. For each instrument, mean values and standard deviations for each partial over the entire instrument dataset are presented.

Analysis of results

Results of the magnitude analysis are presented in Figures 3.6 - 3.9. Mean frame-wise percentage magnitude changes are shown for each partial. The whiskers in the plots indicate the standard deviation.

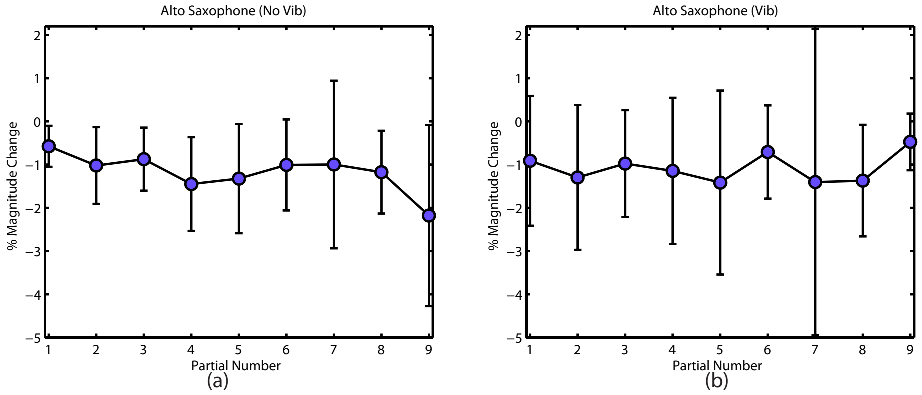


Figure 3.6.: Mean magnitude change of the temporal envelopes of the first 10 partials of saxophone tones. The standard deviation is indicated by the whiskers. (a) Alto sax without vibrato. (b) Alto sax with vibrato.

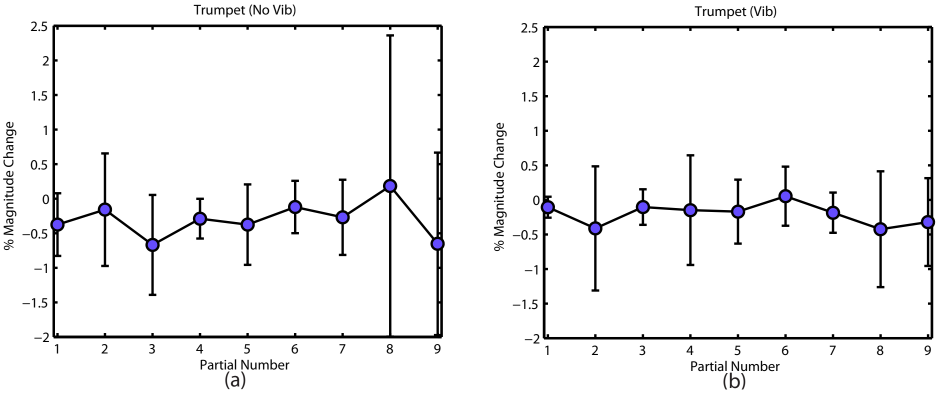


Figure 3.7.: Mean magnitude change of the temporal envelopes of the first 10 partials of trumpet tones. The standard deviation is indicated by the whiskers. (a) Trumpet without vibrato. (b) Trumpet with vibrato.

The saxophone (Figure 3.6) shows mean percentage magnitude changes about -1% both for tones with and without vibrato. The standard deviations differ slightly between the partials with values in the $[0.5, 2]\%$ range for tones without vibrato, and in the $[0.5, 3.5]\%$ range for tones with vibrato.

The trumpet (Figure 3.7) shows very stable mean magnitude envelopes both for

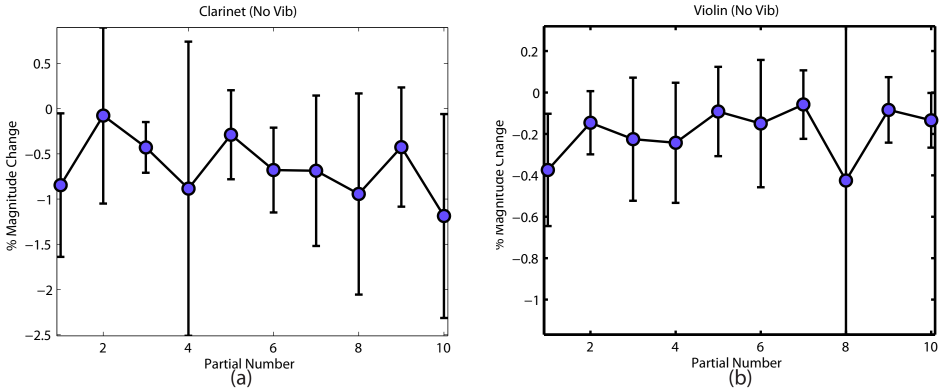


Figure 3.8.: Mean magnitude change of the temporal envelopes of the first 10 partials of clarinet and violin tones without vibrato. The standard deviation is indicated by the whiskers. (a) Clarinetwithout vibrato. (b) Violin without vibrato.

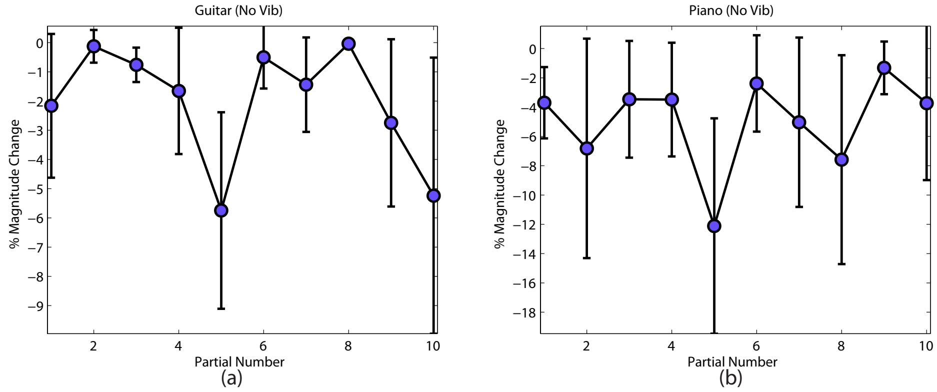


Figure 3.9.: Mean magnitude change of the temporal envelopes of the first 10 partials of guitar and piano tones without vibrato. The standard deviation is indicated by the whiskers. (a) Guitar without vibrato. (b) Piano.

tones with and without vibrato, with mean percentage magnitude changes around -0.3%. Standard deviations differ among partials with values lower than 1% in all cases but for partial 8 in the tones without vibrato.

The clarinet (Figure 3.8a) also shows very stable envelopes with mean percentage magnitude changes around -0.6% with standard deviations in the $[0.3, 1.5]\%$ range for the different partials. The violin (Figure 3.8b) shows mean percentage magnitude changes around -0.3% with standard deviations in the $[0.3, 0.6]\%$ range for the different partials.

The guitar and piano (Figure 3.9) show in general larger mean percentage magnitude changes than the rest of the instruments. The guitar shows mean values between -0.3% and -6% with standard deviations between 0.2% and 4% for the different partials. The piano shows mean values between -2% and -12% with standard deviations between 2% and 7% for the different partials.

Conclusions

The following general remarks can be made from the analysis conducted:

- For the saxophone, trumpet, clarinet, and violin mean percentage magnitude changes are never higher than -2%. This evidences the great degree of smoothness in the magnitude envelopes for an analysis time resolution of 5.8 ms.

- Results show clear differences between those instruments with sustained tones such as the clarinet, saxophone, trumpet, and violin (bowed), and the plucked string instruments such as the piano and the guitar whose tones immediately start decaying after the string has been plucked. The piano and guitar clearly show much larger mean percentage magnitude changes than the rest of the instruments (reaching percentage changes up to 12%). A common and very noticeable phenomenon is observed in the partial 5 of the guitar and piano, showing a very steep magnitude change in comparison to the other partials. The exact cause of this phenomenon is not known but appears to be common to plucked string instruments.

3.2.3.3. Frequency Analysis

In this section, inharmonicity of musical instrument tones as introduced in Section 2.6.2, is studied. In a separation context, the importance of characterizing inharmonicity of musical instruments comes from the fact that it allows more accurate estimation of the target source. Additionally, the study presented in Section 3.2.2 showed that more accurate localization of the sources in the frequency domain results in better quality of separation.

The full dataset was processed to obtain the observed frequency location $\hat{f}_p(n)$ of the first 10 partials of each tone. The observed frequency location was obtained with the partial tracking algorithm described in Section 3.2.1, by finding the spectral peaks corresponding to each of the partials of the tone. To obtain refined frequency locations of the partials, the method proposed in [110] was used to obtain an Instantaneous Frequency (IF) spectrogram. The IF spectrogram has proven to provide better frequency resolution than standard STFT and as such, is relevant for the present study. Frame-wise IF values in Hz were found for each of the partials of the tone.

Additionally, harmonic frequencies $f_p(n)$ in Hz were calculated for all partials. The harmonic frequency of each partial is calculated as the product of the fundamental frequency $f0(n)$ and the partial number p. That is, $f_p(n) = f0(n) \cdot p$, with $p = 2 \cdots 10$. To estimate the deviation of each partial from its harmonic location, the difference between the observed and the harmonic location of each partial was obtained. The deviation of each partial from its harmonic location is given by: $\Delta_p(n) = \hat{f}_p(n) - f_p(n)$.

Analysis of results

Inharmoncity values are presented in Figures 3.10 - 3.13. Results are shown for each partial as mean percentage deviation from its harmonic location. Negative values represent frequency deviations to lower frequencies than the calculated harmonic ones. Positive percentage deviations indicate observed frequencies higher than the calculated harmonic ones. The whiskers in the plots indicate the standard deviation.

For the saxophone and trumpet where tones both played with and without vibrato were available, results are presented independently. For the clarinet, violin, piano, and guitar only tones without vibrato were studied. This study considers inharmonicty from an objective point of view, and no attempt is made to study the perceptual aspects of inharmonicity.

Saxophone tones without vibrato (Figure 3.10a) show percentage frequency deviations in the $[-0.5, 0.5]\%$ range with standard deviation about 2%. For the saxophone tones with vibrato (Figure 3.10b), particularly noticeable are the very stable frequencies of the 3 lower partials, with frequency deviations close to zero and standard deviations about 0.1%. For the rest of the partials of the saxophone tones with vibrato, mean values between $[-0.5, 0.5]\%$ are observed with standard deviations of about 2.5%. The standard deviations of the tones with vibrato are slightly larger than the ones for tones without vibrato. Specially for the saxophone tones without vibrato, a slight tendency to deviations to lower frequencies than the harmonic ones is observed.

Particularly noticeable in the results for the trumpet are the large standard deviations obtained for partial 3 for tones both with and without vibrato (Figure 3.11) of approximately 3.5%. The third partial also showing a clear deviation to lower frequencies approximately 1% lower than the harmonic one. Trumpet tones without vibrato show percentage deviations in the $[-1, 1]\%$ range with standard deviation about 2% (excluding the aforementioned 3th partial). For the trumpet tones with vibrato, mean values in the $[-0.5, 0.5]\%$ range are observed with standard deviations approximately 2%. As opposed to the saxophone, no clear differences in the standard deviations of tones with and without vibrato are observed. For both tones with and without vibrato, a slight tendency to higher frequency deviations is observed.

The clarinet shows very stable behavior particularly in the lower partials where mean values are very close to zero (Figure 3.12a). The rest of the partials show mean percentage deviations of approximately 0.1%, showing in general, almost completely harmonic behavior. The standard deviations tend to increase with increasing partial number but remain lower than 0.5%. No clear tendency of deviations towards lower or higher frequencies than the harmonic ones can be observed.

For the violin (Figure 3.12b), mean percentage deviations in the $[0, 0.5]\%$ range are observed, showing a tendency to deviations to higher frequencies than the harmonic ones. Deviation tends to increase with increasing partial number. Standard deviations about 1.5% range are observed.

The guitar (Figure 3.13a) shows mean percentage deviations in the $[-0.5, 0.5]\%$ range exhibiting almost harmonic behavior with a slight tendency to lower frequency deviations. Standard deviations around 2% are observed. The piano (Figure 3.13b) shows mean percentage deviations in the $[-1, 1]\%$ range with a slight tendency to higher frequency deviations in the higher partials. Standard deviations around 2% are observed.

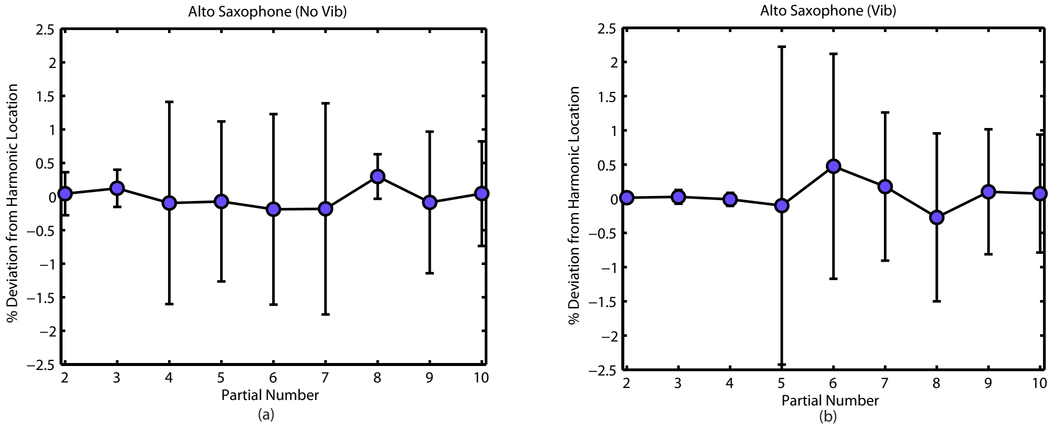


Figure 3.10.: Mean percentage deviation of the observed frequency of the first 10 partials of a tone with respect to calculated harmonic locations. (a) Alto saxophone without vibrato. (b) Alto saxophone with vibrato.

Conclusions

Even though results show clear differences between the different musical instruments, general observations can be made:

- None of the musical instruments shows a definite tendency to lower or higher frequency deviations than the harmonic ones. Some tendencies can be observed

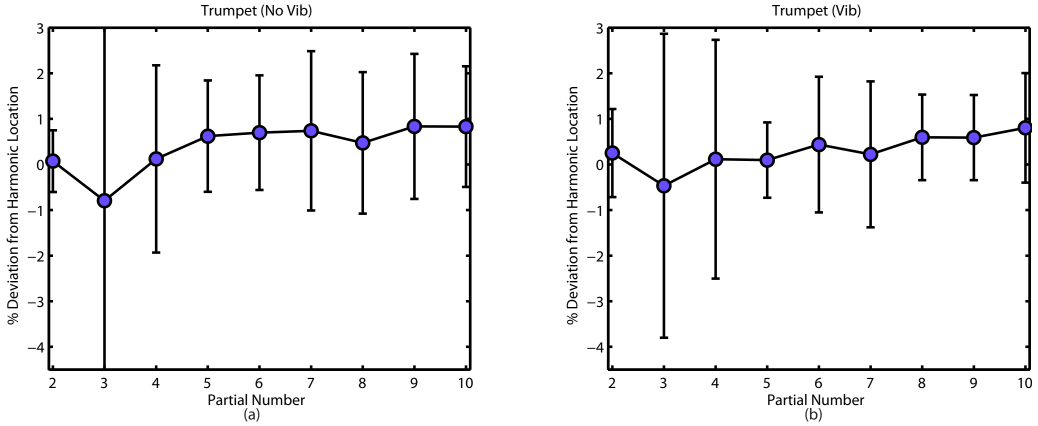


Figure 3.11.: Mean percentage deviation of the observed frequency of the first 10 partials of a tone with respect to calculated harmonic locations. (a) Trumpet without vibrato. (b) Trumpet with vibrato.

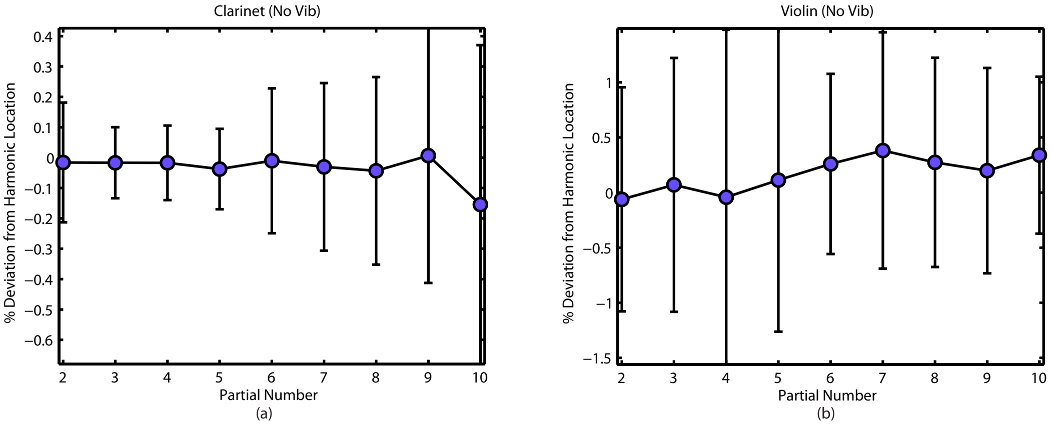


Figure 3.12.: Mean percentage deviation of the observed frequency of the first 10 partials of a tone with respect to calculated harmonic locations. (a) Clarinet without vibrato. (b) Violin without vibrato.

for some of the instruments, but in general, all the instruments studied can show both lower and higher deviations.

- Mean deviations and standard deviations tend to increase with increasing partial number.

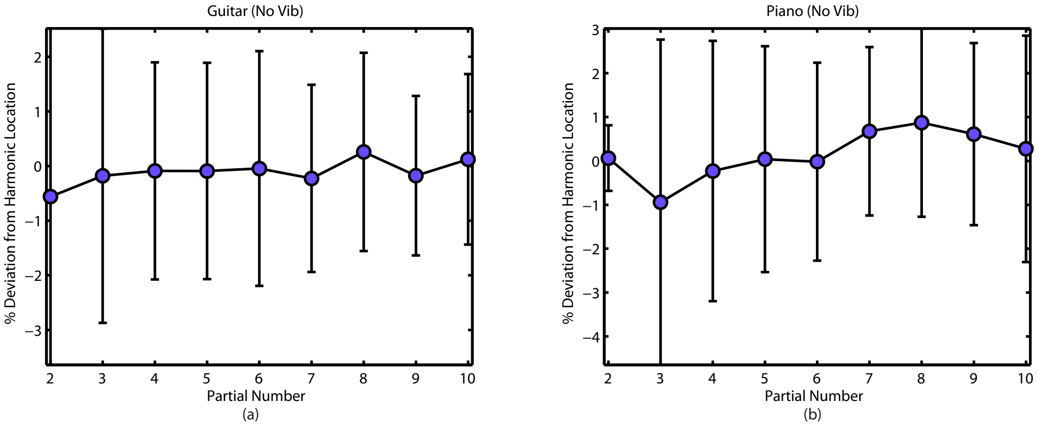


Figure 3.13.: Mean percentage deviation of the observed frequency of the first 10 partials of a tone with respect to calculated harmonic locations. (a) Guitar without vibrato. (b) Piano.

- Mode-locking mechanisms as described in Section 2.6.2 result in partial frequencies very close to harmonic ones. All the mean percentage frequency deviations observed are in the $[-1, 1]\%$ range.
- Even though mean values for all the instruments fall approximately in the same percentage deviation range, larger standard deviations were observed for the two plucked string instruments (piano and guitar) than for the other ones. The main difference between the two groups of instruments is that for the plucked ones, there is no mechanism available that allows to produce sustained tones. After the string has been plucked, the partials immediately start decaying. For the rest of the instruments, the performer can control the duration of the tone and use either the blowing pressure or the bow to sustain the tone as needed. This clearly sets a difference in the frequency characteristics of the two groups of instruments. Similar differences were observed in the magnitude analysis presented in Section 3.2.3.2.
- Given that this analysis was conceived within a separation context with the goal of drawing general guidelines for parameter estimation, no consideration was taken regarding the register of the notes analyzed. All the results

presented are mean values over the full range of the instruments considered. More specific inharmonicty tendencies (such as deviations to higher frequencies than the harmonic ones in the lower register of the guitar) could very likely be observed if pitch is considered in the analysis. However, as the class of the solo instrument is not known beforehand in the proposed method for solo and accompaniment separation, specific instrument distinctions for parameter estimation cannot be done and thus, this analysis was not conducted.

3.2.3.4. Phase Analysis

(1) Phase Correlation and Coupling

The principle of Common Fate states that different parts of the spectrum that change in the same way in time will probably belong to the same environmental sound [111]. This is a characteristic that can be clearly observed in the Instantaneous Frequency Distribution (IFD) of many musical instrument tones (see Section 2.6.3 for a detailed description of IFD). Figures 3.14 - 3.21 show the IFD of different musical instrument tones. All the IFDs where obtained using a window length of 4096 samples, a hop size of 256 samples, and a sampling frequency $fs = 44100\,\text{Hz}$. For visualization purposes only, and to avoid excessive overlapping between the different contours, the vertical axis of the plots was extended to the $[-2\pi, 2\pi]$ range; however, the original calculations use the standard $[-\pi, \pi]$ range for the phase. Additionally, to get a better understanding of the behavior of IFD across different musical instruments, plots for the saxophone, trumpet, violin, clarinet, flute, and trombone are presented.

For the saxophone (Figure 3.14), trumpet (Figure 3.15), and flute (Figure 3.16) tones with vibrato are presented. It can be seen that modulations in the IFD are common among the different partials of the tone and the principle of Common Fate can be clearly observed. Similar behavior as the one observed for the tones with vibrato, can be observed for the trombone (Figure 3.17), clarinet (Figure 3.18), and violin (Figure 3.19) tones played without vibrato. For these examples, common micro-modulations

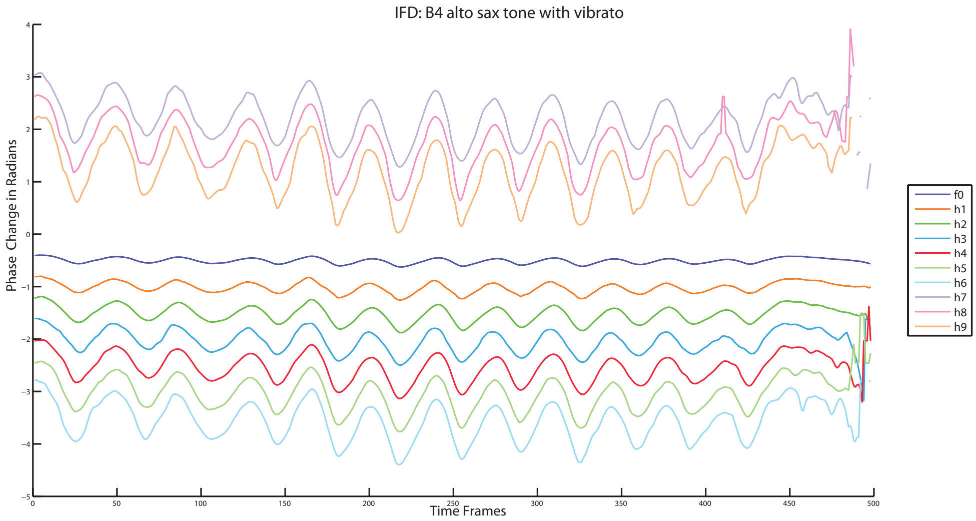


Figure 3.14.: Instantaneous Frequency Distribution (IFD): B4 alto saxophone tone with vibrato. The modulations are common to all partials of the tone.

can be seen among all the partials. For the violin, the sustained part of the tone is very stable, almost showing linear contours; however, common modulations can be more clearly seen in the beginning of the tone. For the guitar (Figure 3.20) and the piano (Figure 3.21), common modulations among the different partials cannot be observed, and IFD contours do not show clear correlations. Similar results were presented in [112, O3] and are most likely caused by non-linear interactions between the strings and the instrument's body. In the case of the piano, also the string-hammer interactions have shown to be non-linear and might have an influence in phase characteristics. Here again, results show that phase correlation characteristics differ among musical instruments.

Given that some musical instruments can exhibit clear common modulations in the IFD of the different harmonics, the natural question that follows is whether this information can be used to predict the phase of one harmonic component given the phase of another harmonic component of the same tone. The concept of phase coupling becomes relevant in this context. In this section, partials will be denoted by

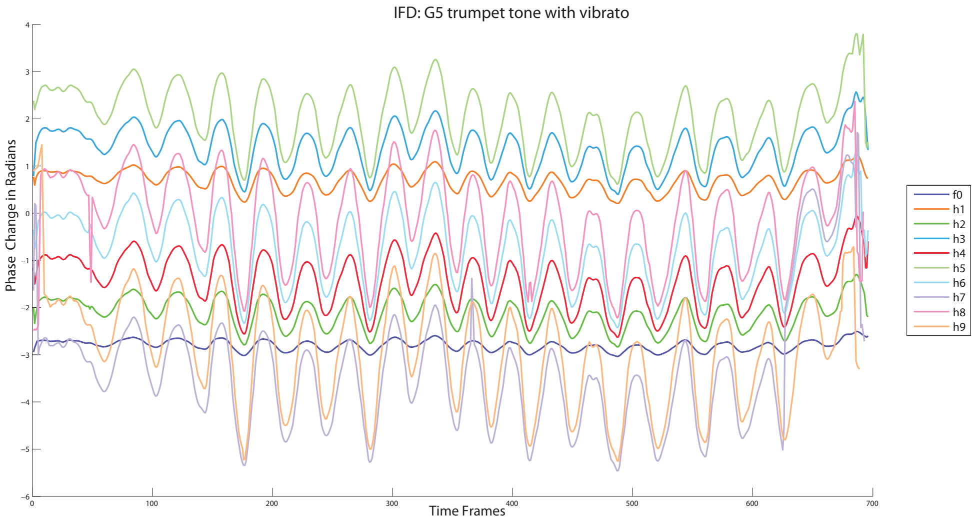


Figure 3.15.: Instantaneous Frequency Distribution (IFD): G5 trumpet tone with vibrato. Clear common modulations among the partials of the tone can be observed.

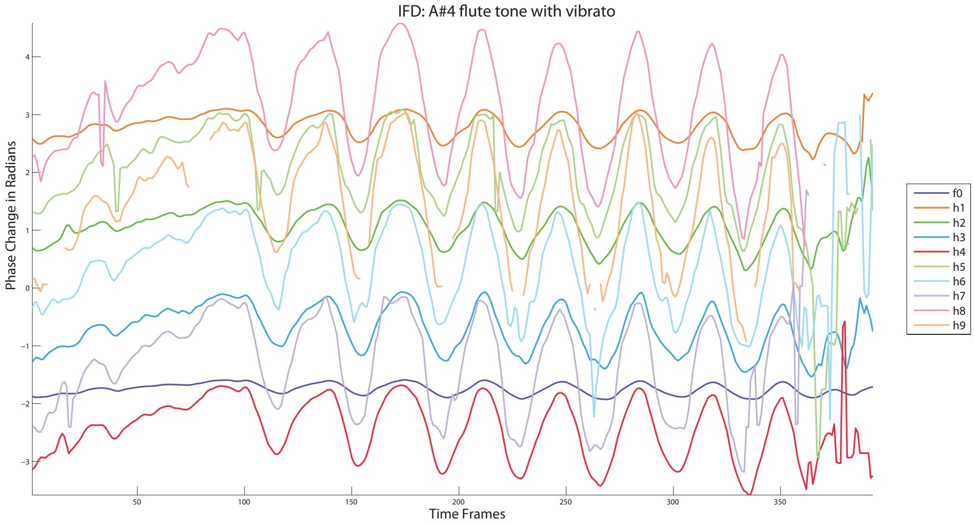


Figure 3.16.: Instantaneous Frequency Distribution (IFD): A#4 flute tone with vibrato. It can be observed that the contour of the 9th harmonic (h9) is not completely continuous and some segments are missing. This happens when for a given frame, the partial tracking method cannot find a peak for a given partial. These frames are marked as *missing information* in the analysis.

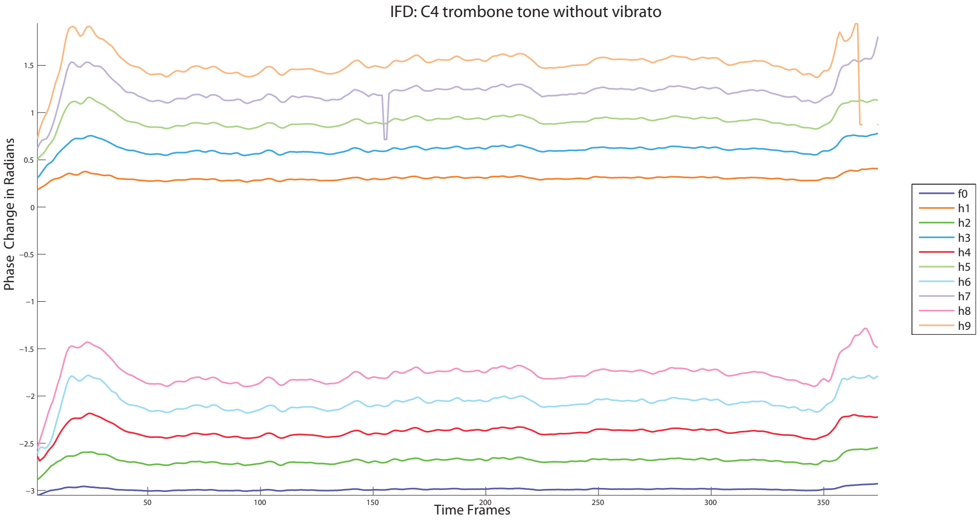


Figure 3.17.: Instantaneous Frequency Distribution (IFD): C4 trombone tone without vibrato. Common modulations can be observed in all the partials.

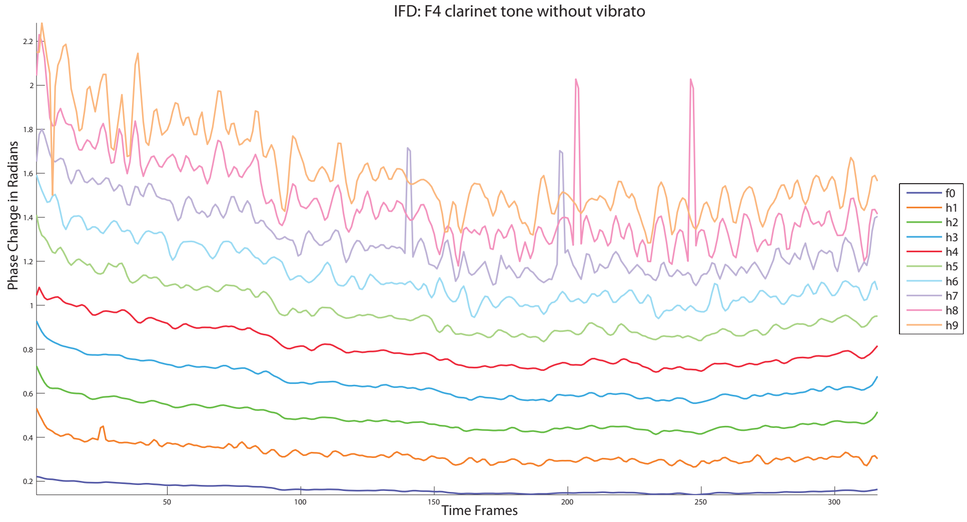


Figure 3.18.: Instantaneous Frequency Distribution (IFD): F4 clarinet tone without vibrato. It can be seen that for tones without vibrato, the common fate principle also applies.

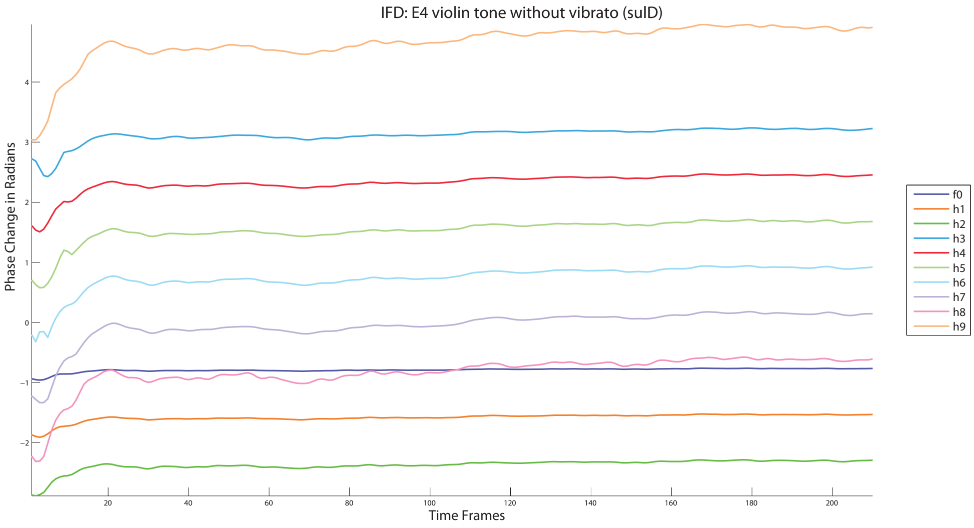


Figure 3.19.: Instantaneous Frequency Distribution (IFD): E4 violin tone without vibrato. In this case, extremely stable contours are observed, being the common modulations more clearly in the first frames of the tone.

p_x, with x the partial number. In general, phase coupling implies that for a triplet of harmonically related partials p_i, p_j, p_h, and with $p_h = p_i + p_j$, any deviations that occur in their respective phases $\phi_{p_i}(n)$, $\phi_{p_j}(n)$ will sum up to occur identically in $\phi_{p_h}(n)$ [112]. That is:

$$\phi_{p_i}(n) + \phi_{p_j}(n) - \phi_{p_h}(n) = 0 \tag{3.18}$$

Evaluation and Results

For phase coupling characteristics to be applicable within a sound separation context, one condition must be fulfilled: To be able to estimate information of partial p_h, it must be guaranteed that information of at least two partials p_j and p_i that fulfill the condition $p_h = p_j + p_i$, is available. The study presented in this section evaluates how accurately the phase information from one partial can be estimated

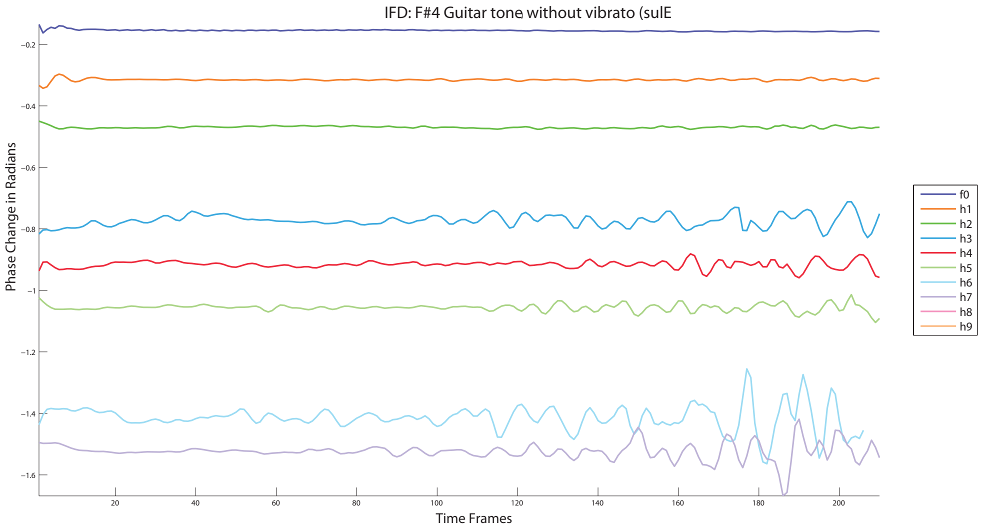


Figure 3.20.: Instantaneous Frequency Distribution (IFD): F#4 guitar tone without vibrato.

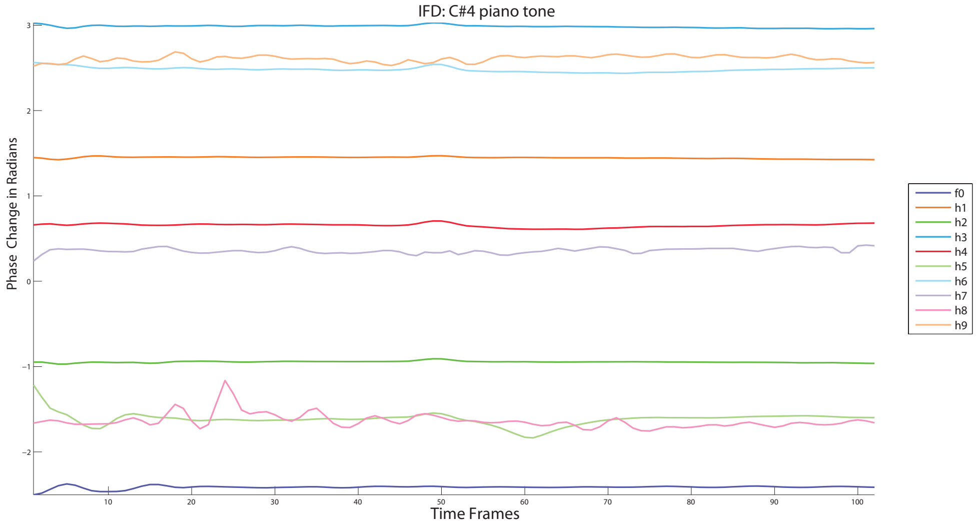


Figure 3.21.: Instantaneous Frequency Distribution (IFD): C#4 piano tone without vibrato.

using the phase information of two harmonically related partials. The first 10 partials of each tone were used, and estimation of partials p_3 to p_{10} were evaluated. The fundamental frequency $f0$ cannot be estimated under this condition but was used for the estimation of the higher partials. The different possibilities of harmonically related partial triplets were considered. For partial p_5 for example, the two combinations of partials that fulfill $p_h = p_i + p_j$ were considered, that is, $p_5(1,4)$ and $p_5(2,3)$. The notation $p_h(i,j)$ is used in this section to denote harmonically related partials.

For the saxophone, trumpet, and flute both tones with and without vibrato were considered. For the clarinet, trombone, violin, and piano only tones without vibrato were considered. Mean phase reconstruction errors in radians are shown in Figures 3.22 - 3.31 for partials 3 to 10. The different possibilities of harmonically related partial triplets are shown in each plot. For easier visualization, different background colors are used for consecutive partials, and the $p_h(i,j)$ notation is shown in the horizontal axis of each plot.

Results for the saxophone (Figures 3.22 and 3.23) show mean reconstruction errors between 0.2 and 0.4 radians with larger errors obtained for higher partials. For the trumpet, reconstruction errors for tones without vibrato (Figure 3.24) tend to be smaller when p_1 (fundamental frequency) is used for reconstruction in comparison to other partial combinations. Mean reconstruction errors between 0.2 and 0.4 radians are obtained. For the 3 lower reconstructed partials of trumpet tones with vibrato (Figure 3.25), reconstruction errors are very close to zero with standard deviations of approximately 0.1 radians. For flute tones without vibrato (Figure 3.26) errors between 0.2 and 0.4 radians are obtained. For flute tones with vibrato (Figure 3.27) mean errors greatly increase with increasing partial number, reaching values up to 0.8 radians. The clarinet (Figure 3.28) also shows more accurate estimations when p_1 (fundamental frequency) is used for reconstruction in comparison to other partial combinations. Very accurate lower partial reconstruction is obtained for the clarinet with reconstruction errors very close to zero. Larger standard deviations are observed for higher partials. The trombone (Figure 3.29) as well as the violin (Figure 3.30) obtain mean reconstruction errors between 0.2 and 0.4 radians. As expected

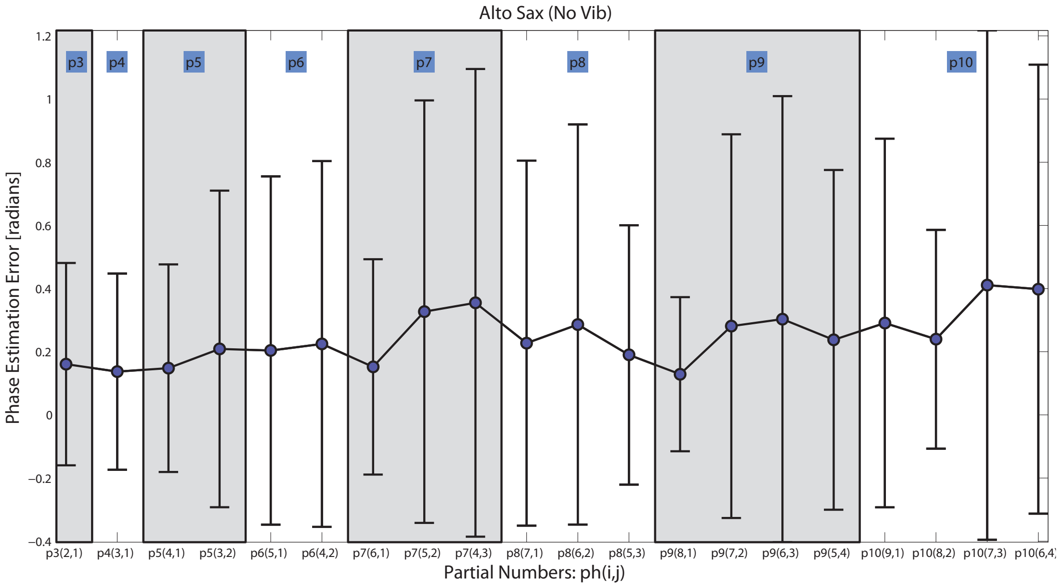


Figure 3.22.: Phase estimation error: Saxophone tones without vibrato. Estimation errors around 0.2% radians are observed.

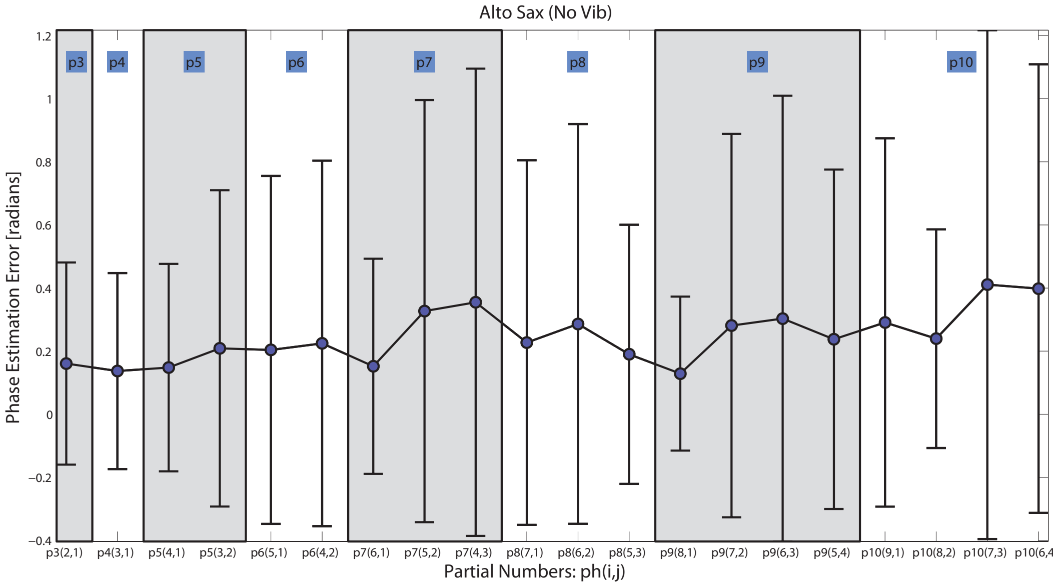


Figure 3.23.: Phase estimation error: Saxophone tones with vibrato. There is a slight increase in the estimation error of higher partials.

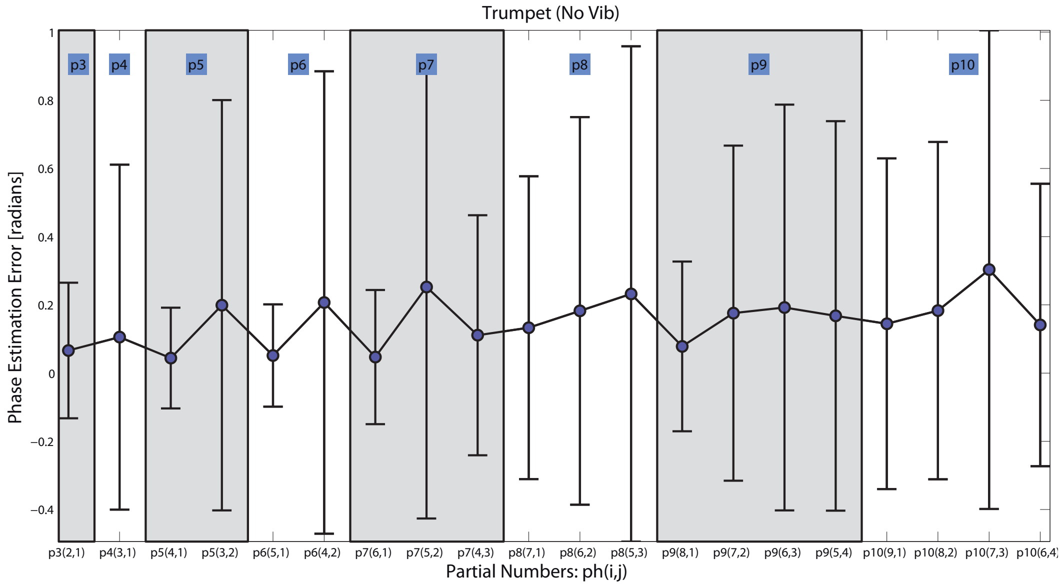


Figure 3.24.: Phase estimation error: Trumpet tones without vibrato. Phase reconstruction using p_1 results in smaller phase estimation errors than with other partial combinations.

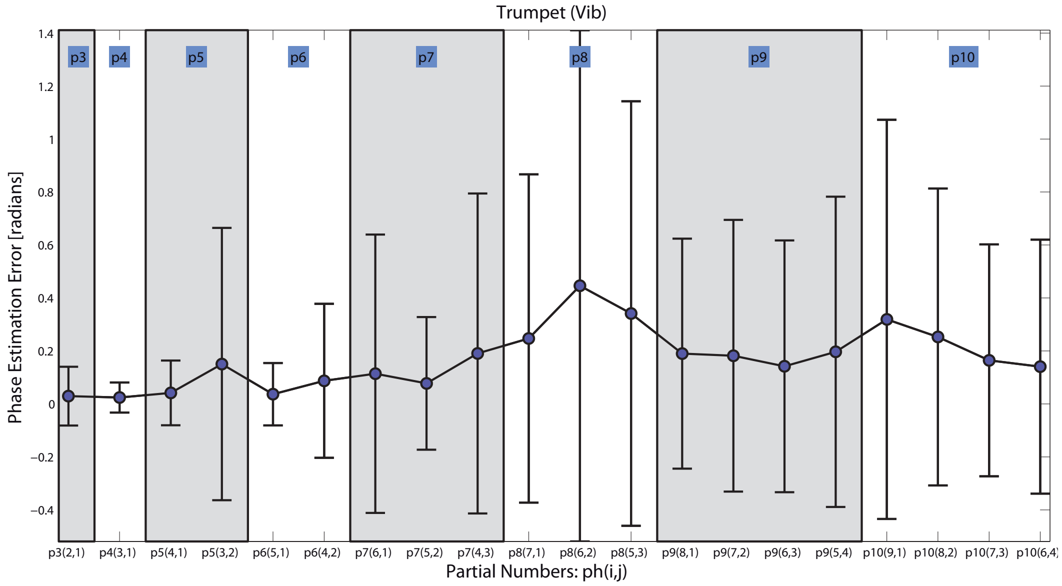


Figure 3.25.: Phase estimation error: Trumpet tones with vibrato. The first 3 reconstructed partials show reconstruction errors very close to zero.

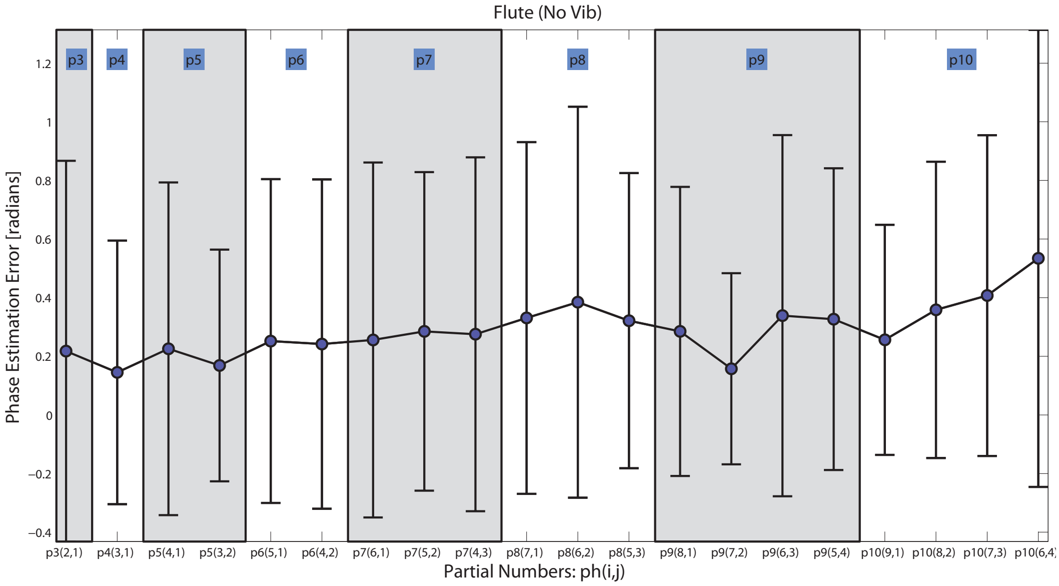


Figure 3.26.: Phase estimation error: Flute tones without vibrato. Estimation errors around 0.2% radians are observed.

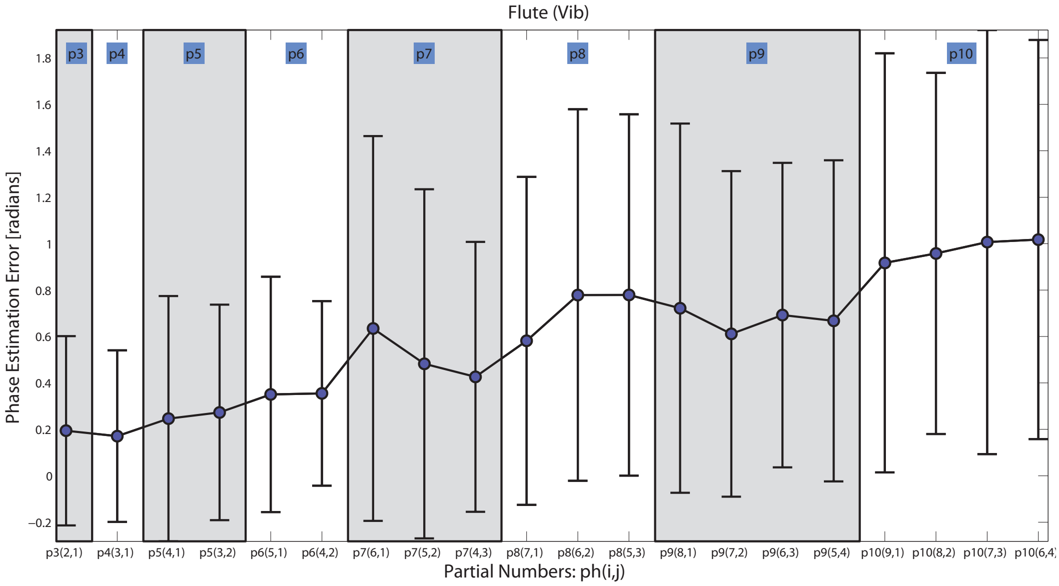


Figure 3.27.: Phase estimation error: Flute tones with vibrato. Phase estimation error greatly increases for higher partials.

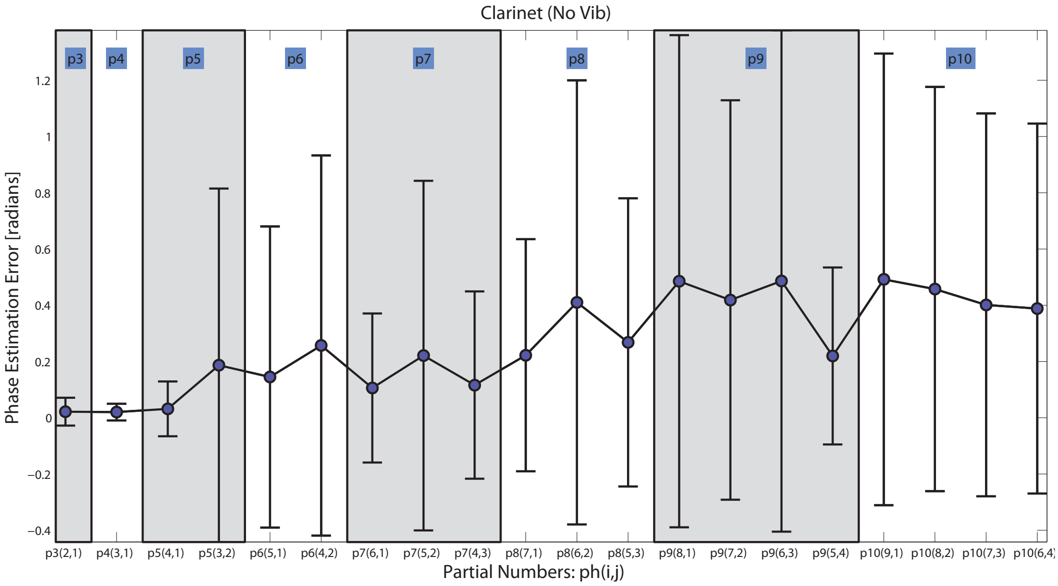


Figure 3.28.: Phase estimation error: Clarinet tones without vibrato. Phase estimation error increases with increasing partial number.

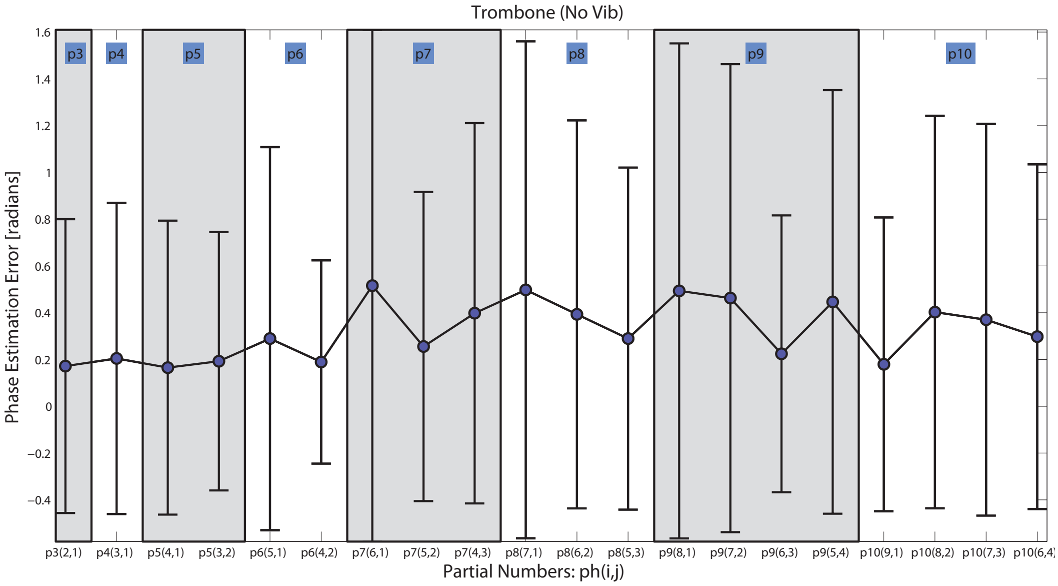


Figure 3.29.: Phase estimation error: Trombone tones without vibrato. Estimation errors around 0.2% radians are observed.

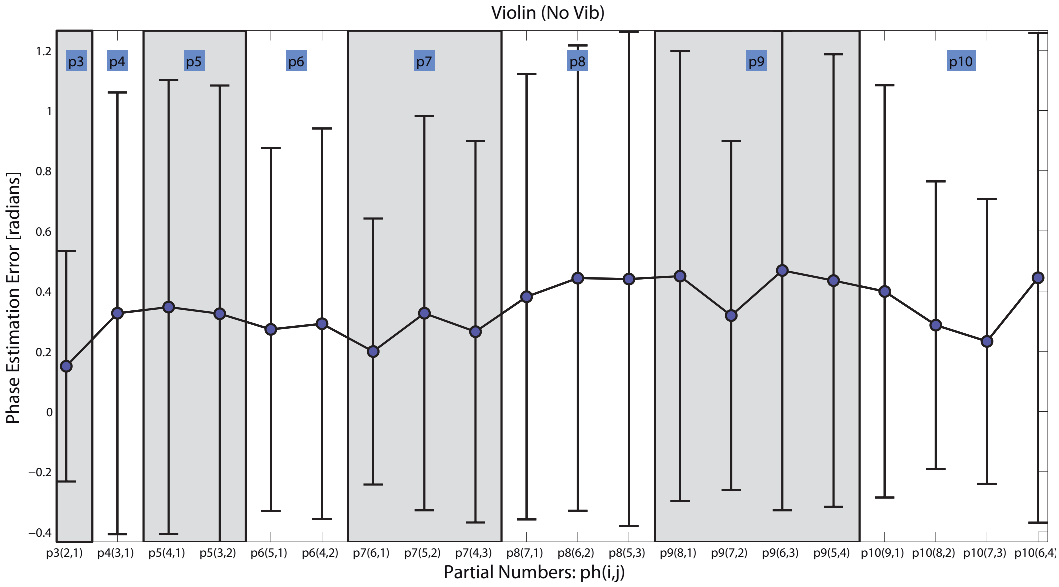


Figure 3.30.: Phase estimation error: Violin tones without vibrato. Estimation errors around 0.2% radians are observed.

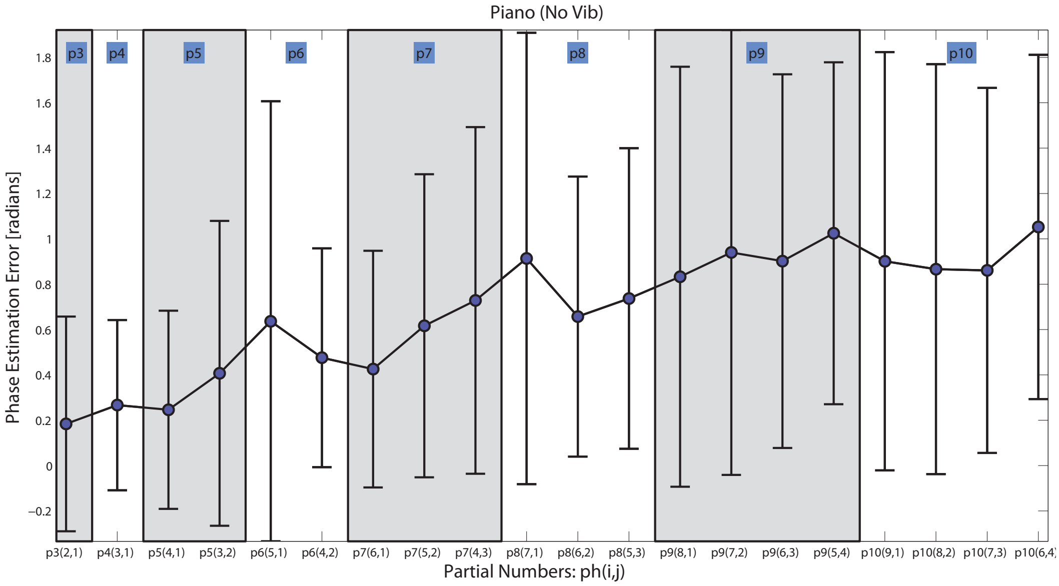


Figure 3.31.: Phase estimation error: Piano tones without vibrato. The piano shows the largest estimation errors of all instruments, with increasing error for higher partials.

given the uncorrelated IFDs of the piano in Figure 3.21, reconstruction errors for the piano (Figure 3.31) are considerably higher than for the rest of the instruments, with errors increasing with increasing partial number and reaching values up to 1 radian.

Conclusions

The following general observations can be made from the studies on phase reconstruction using phase coupling characteristics described in this section:

- Results suggest that estimation for lower partials tends to be more accurate than for higher partials, obtaining in general lower reconstruction errors for all the instruments.

- In the studies on phase presented in this section, the piano has proven to be the biggest challenge of all the instruments considered. Phase coupling characteristics are not observed in general for the piano, and reconstruction errors are considerably higher than for the rest of the instruments.

- It is particularly noticeable the large standard deviations obtained in general for all the instruments. Only the clarinet and the trumpet tones with vibrato show standard deviations around 0.1 radians, and only for the lower partials. These large standard deviation values show the difficulty of obtaining robust phase estimations under all conditions. It has to be noted that these studies were conducted with isolated instrument tones where the analysis conditions are somehow ideal. Within a separation scenario, spectral analysis and partial tracking is a much more complex task and estimations can only decrease in accuracy. For phase reconstruction based on coupling to be possible, it has to be guaranteed that clean phase information from at least two harmonically related partials can be extracted. This is in itself, another challenge to be added to the estimation task.

- To give a rough idea of the perceptual impact of phase reconstruction errors as the ones observed, all the tones for the clarinet where reconstructed using estimated phase for partials p_3 to p_{10}. All partials were reconstructed using p_1 (fundamental frequency). That is, for partial p_3 for example, the combination $p_3(p_1, p_2)$ was used. The clarinet was chosen as it showed the most accurate estimations of all instruments and as such, can give a good performance reference. The first partial p_1 was used for reconstruction because for the clarinet in particular, it proved to result in lower reconstruction errors. The Perceptual Evaluation of Audio Quality (PEAQ) model [113] was used to calculate the perceptual quality of the reconstructed audio tracks with respect to the original instrumental tone. PEAQ returns for each signal an Objective Difference Grade (ODG) which is an impairment scale with the following meaning: 0 imperceptible, -1 perceptible but not annoying, -2 slightly annoying, -3 annoying, and -4 very annoying. The clarinet dataset obtained a mean ODG of -1.18 showing distortions that according to the scale, are slightly annoying. This evaluation only gives a very rough reference of the reach of this approach. All the signals were synthesized with 7 partials with reconstructed phase. This is naturally a hypothetical scenario and more studies need to be conducted under different characteristics to better assess the perceptual impact of phase reconstruction.

(2) Phase Expectation: Phase-based Harmonic/Percussive separation

In this section phase expectation is explored in the context of sound source separation. The main idea behind phase expectation is that the frame-wise change in phase of a harmonic source, can be predicted given the pitch of the source and the hop size H in samples of the time-frequency transform. For a given harmonic source appearing in frequency bin k of the time-frequency representation, the phase change in radians from time frame n to time frame $n+1$ is given by:

$$\Delta\phi_k(n) = \frac{2\pi f_k \cdot H}{fs} \tag{3.19}$$

with fs the sampling frequency, and f_k the center frequency of bin k in Hz.

In a given time-frequency transform where a window of length N has been used, each frequency bin k covers a band of frequencies of size fs/N. Equation (3.19) can be used to calculate the range of expected phase changes for the frequency band covered by each bin k. These values of expected phase changes can be used to predict whether the energy falling in a given time-frequency bin belongs to a harmonic source or not. If the change in phase between two consecutive time frames falls in the expected radian ranges, the source can be assumed harmonic. If on the contrary, the phase change falls outside the expected ranges, the source exhibits transient or noise-like characteristics.

Spectral leaking plays a critical role in this type of analysis. For a certain harmonic source that exhibits a peak in frequency bin k, it is likely that at least the adjacent frequency bins, $k-1$ and $k+1$, are also affected by the presence of the harmonic source. With this in mind, it is to be expected that phase expectation characteristics of the frequency bins adjacent to a harmonic peak are also affected by spectral leaking. Figure 3.32 shows two example plots where phase expectation of the frequency bin k (where the peak is observed), and its two adjacent bins, $k-1$ and $k+1$, is displayed. Figure 3.32a shows a phase expectation plot extracted from the sustained part of an isolated A4 trumpet tone without vibrato using the partial tracking method described in Section 3.2.1. Figure 3.32b shows the phase expectation plot of two consecutive A4 saxophone tones. The saxophone tones were estimated using the tone-based separation algorithm proposed in Section 3.3 from an audio mixture of a jazz quartet containing saxophone, piano, bass, and drums. In both plots, the horizontal axis represents the time in frames. Additional black bars displaying the length of the tones are shown for reference (A4 trumpet tone and A4 saxophone tone). The vertical axis in both plots displays three frequency bins for each of the ten first partials of the tones; namely, the lower adjacent bin $k_p - 1$, the bin where the peak is observed k_p, and the higher adjacent bin $k_p + 1$. The natural order of the frequency bins was preserved in the plot, being $k_p - 1$ the lower bin shown, k_p the middle bin shown, and $k_p + 1$ the higher bin shown for each partial. The color convention used is displayed on the right side of the figure. The time-frequency bins

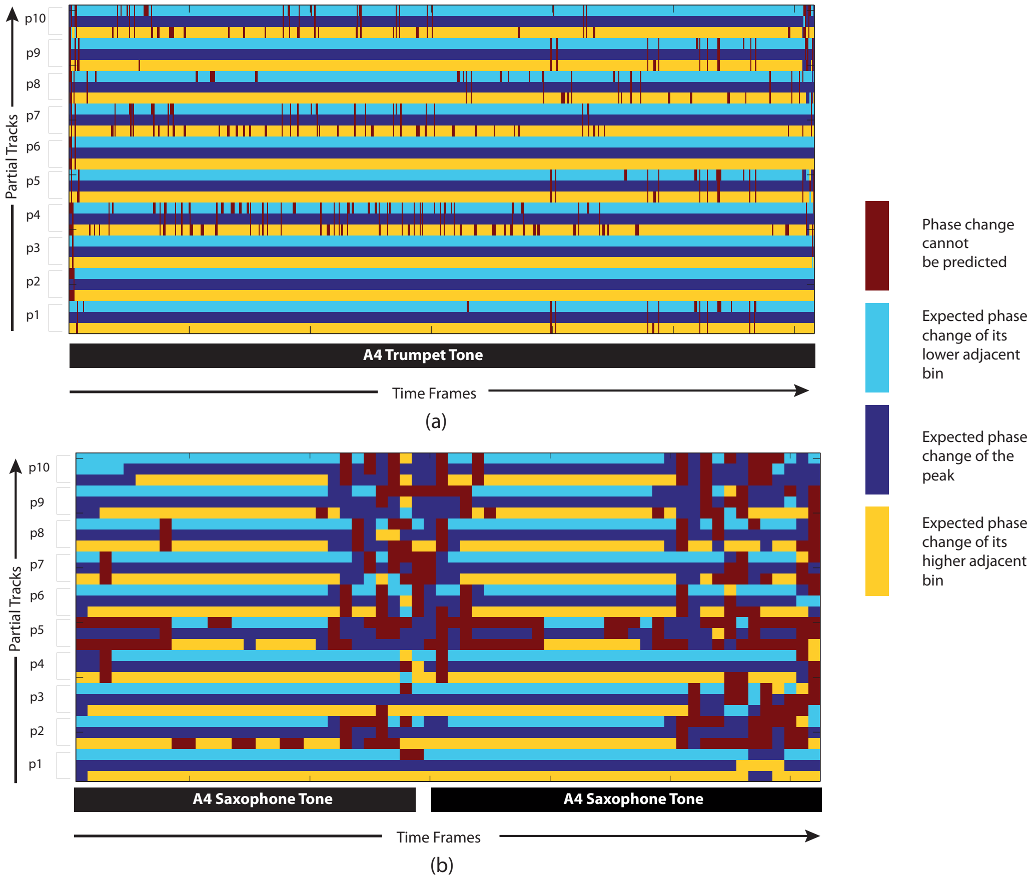


Figure 3.32.: Phase expectation plots for the first 10 partials of musical instrument tones. For each partial p_i, the bin corresponding to the main peak and its two adjacent frequency bins are shown. The dark blue color shows time-frequency bins with expected phase changes. The dark red color shows time-frequency bins whose phase cannot be predicted. (a) A4 Trumpet tone (b) Two consecutive A4 saxophone tones.

whose phase change falls in the expected radian ranges as calculated with (3.19), are shown in dark blue. Those frequency bins whose expected phase change falls in the radian ranges of their higher adjacent bins are shown in yellow. Those frequency bins whose expected phase change falls in the radian ranges of their lower adjacent bins are shown in light blue. Those frequency bins whose phase cannot be explained or predicted based on radian ranges calculated with (3.19), are shown in dark red.

The plots clearly show a structured behavior for the phase expectation of each partial and its adjacent bins. It can be observed that the frequency bin k_p where each partial p is observed, presents phase change values mostly in the predicted ranges, showing very clear dark blue horizontal trajectories. Phase expectation of $k_p + 1$ and $k_p - 1$ clearly follow the behavior of k_p, $k_p - 1$ being mostly pulled to higher phase changes (closer to k_p) than the ones given by (3.19). This can be observed by the clear yellow horizontal trajectories in the plots. Phase expectation of $k_p + 1$ is also clearly affected by the main peak k_p, showing phase change values lower to the ones predicted and mainly being pulled down to be closer to k_p. This can be observed by the horizontal light blue trajectories in the plots. As a result, the three bins corresponding to each partial show a very structured behavior mostly in a – yellow - dark blue - light blue – configuration. A very important observation to be made, is the behavior of phase expectation when the higher partials of the tones in Figure 3.32b decay. It can be seen that as the partials decay at the end of each of the two tones, their phase changes cannot longer be accurately predicted, mainly showing dark red color in the plot. In Figure 3.32a, only the sustained part of the tone is shown and thus, most frames follow the expected behavior. This clear difference in the behavior of phase expectation between harmonic sources and non-harmonic ones will be exploited in the separation context.

The concept of phase expectation is used to address the harmonic/percussive separation task. Preliminary results on this direction were published in [O3] where the proposed method outperformed the method proposed in [114]. In this thesis, these results are extended to a larger dataset and results are compared to a more recent method proposed in [115] where the authors claim to obtain similar results to the ones obtained in [114]. Both in [114] and [115], the fact that percussive instruments mostly appear as vertical events in the magnitude spectrogram is used to differentiate harmonic from percussive components. To better understand this concept, Figure 3.33 shows two example spectrograms of percussive instruments. In both examples, clear vertical events mark the location of percussive hits.

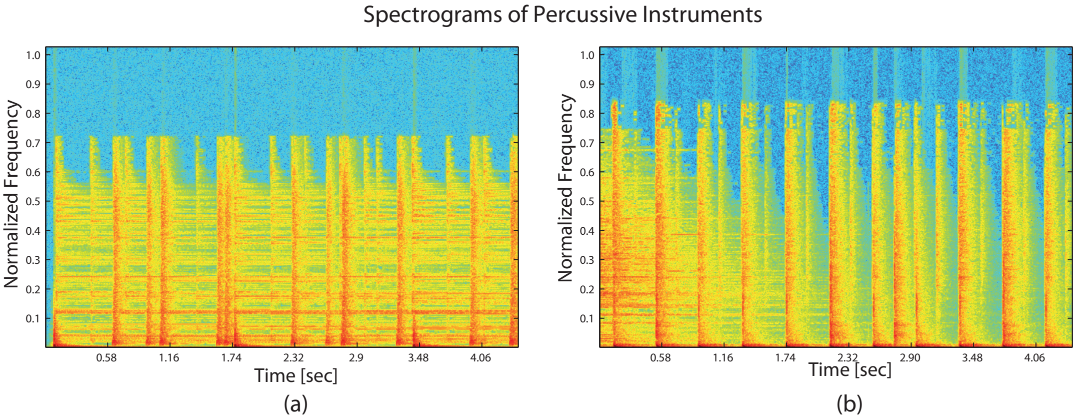


Figure 3.33.: Example spectrograms of percussive instruments. In both figures, clear verticals events in the spectrogram mark the percussive hits.

The system proposed in [114] attempts to differentiate between horizontal and vertical components in the spectrogram by minimizing an objective function based on the quadrature form of the spectrogram gradients. Similarly, [115] proposes a method for harmonic percussive separation based on the use of median filtering in the horizontal (time domain—to detect harmonic instruments), and in the vertical (frequency domain—to detect percussive instruments) directions of the spectrogram.

As opposed to these two methods, phase expectation is exploited here to perform the separation task. In the approach here proposed, the fact that for a certain frequency bin k, phase values of tonal components will fall within a radian range determined by the frequency band covered by k, and the hop size H of the time-frequency transform, is exploited. Phase values outside the calculated range are assumed non-harmonic and classified as percussive components. The method works as follows: (1) The magnitude and phase spectrograms of the input audio signal are obtained by means of the STFT. (2) For each frequency bin k in the time-frequency transform, the minimum and maximum radian changes are calculated using (3.19). (3) The main spectral peaks in the power spectrogram in each frame are detected using the peak detection algorithm described in Section 3.2.1. A spectral mask is created where for each time frame n, the frequency bins where clear peaks are observed are marked with 1s and the rest are marked with 0s. (4) The phase spectrogram is masked using the calculated binary mask. (5) The Instantaneous Frequency Distribution (IFD) as described in Section 2.6.3 is calculated for the masked phase spectrogram. (6) Binary spectral masks for the percussive and harmonic components are created. For every time frame n and bin k, phase values that fall within the calculated radian ranges are assigned to the harmonic mask (marked with 1). The remaining time-frequency bins are classified as percussive and marked with 1 in the percussive mask. It has been observed that for those time-frequency bins where both harmonic and percussive components overlap, phase characteristics of percussive instruments tend to prevail. These time-frequency bins are classified as percussive and no attempt is made to estimate the underlying harmonic component. (7) Spectral leakage is considered by including the adjacent frequency bins $k_p + 1$ and $k_p - 1$ (whose phase changes follow the ones of the main peak k_p) in the harmonic mask. In Figure 3.32, that means that all the yellow and light blue time-frequency bins are also included in the harmonic mask. (8) The complex spectrogram of the original mixture is masked to obtain estimates of the harmonic and percussive sources. (9) Audio signals for the percussive and harmonic components are obtained by means of the Inverse Short Time Fourier Transform (ISTFT).

Dataset

The dataset SA_DS1a (copyright free) was used to evaluate the harmonic/percussive separation method. Those signals that do not contain percussive instruments were removed from the evaluation (signals 2, 8, 11, 12, 14, 16, and 17 in Table B.2). That gives a total of 10 signals for which percussive and harmonic signals were mixed from the multi-track recordings. The dataset was processed with the proposed method and with the algorithm proposed in [115] for comparison (reference algorithm).

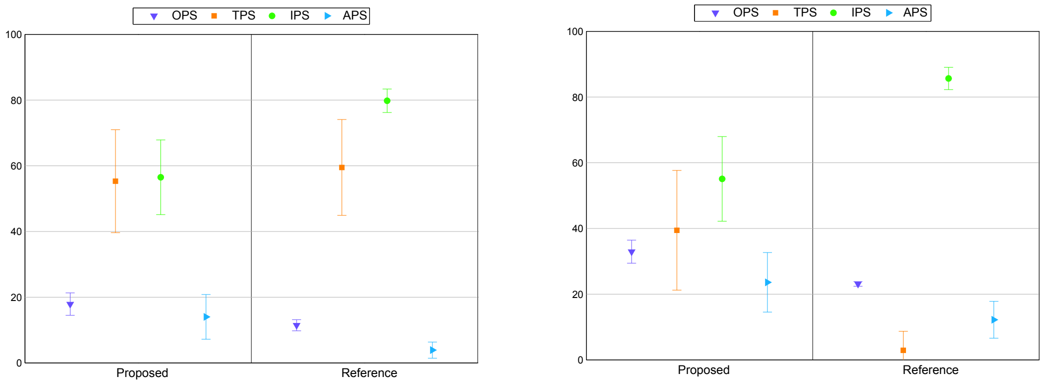
Implementation Details

For the proposed method, the following algorithm parameters were used: a window length $N = 2048$ with a hop size $H = 128$ was used. All tracks in the dataset had a sampling frequency $fs = 44100\,\text{Hz}$. For the peak detection algorithm the frequency adaptive magnitude threshold φ_f was calculated using a frequency delta $\Delta_f = 50\,\text{Hz}$.

For the reference algorithm [115] the processing parameters recommended by the author for best performance were used: window length $N = 4096$, hop size $H = 1024$, filter length $L = 17$, and spectral compression parameter $p = 2$.

Evaluation and Results

Perceptual quality measures for both algorithms were obtained using the PEASS Toolkit. Mean values and 95% confidence intervals are presented in Figure 3.34. The full table of results is also presented in Appendix C for reference.



(a) Harmonic

(b) Percussive

Figure 3.34.: Results obtained with the proposed harmonic/percussive separation algorithm and the algorithm proposed by Fitzgerald [115] (Reference). Overall Perceptual Score (OPS), Target-Related Perceptual Score (TPS), Interference-Related Perceptual Score (IPS), Artifact-Related Perceptual Score (APS). Mean values with 95% confidence intervals are presented. (a) Harmonic. (b) Percussive. It can be observed that the proposed algorithm outperforms the reference method in terms of OPS and APS. For the percussive tracks the proposed method also outperforms the reference method in TPS. The reference algorithm always results in higher IPS than the proposed method.

The proposed algorithm outperforms the reference method in terms of the Overall Perceptual Score (OPS) both for the percussive and harmonic components. Particularly noticeable is the performance improvement obtained with the proposed method for the percussive components, the proposed method obtaining a mean OPS score of 32.93, and the reference method a mean OPS score of 23.11 over the entire dataset.

For the harmonic components (Figure 3.34a), the proposed method obtains slightly higher APS scores than the reference algorithm, but slightly lower TPS scores.

For the percussive components (Figure 3.34b) the proposed method outperforms the reference algorithm in three of the perceptual scores, that is, OPS, TPS, and APS.

It is to be noted the particularly high IPS scores obtained by the reference algorithm for the both harmonic and percussive components, outperforming the proposed method in both cases. Additionally, the reference algorithm shows in general smaller confidence intervals than the proposed method.

Informal listening tests showed that for vocal tracks, the proposed method assigns more of the fricative and plosive sounds to the percussive signal than the reference algorithm. For the instrumental tracks, more information from the attacks of the instruments is assigned by the proposed algorithm to the percussive signal than the reference algorithm. These two observations explain the clear difference in IPS scores between the two methods. Due to the transient-like characteristics of fricatives, plosives, and attacks, their corresponding phase exhibits non-harmonic characteristics and consequently, are assigned to the percussive components in the separation.

3.3. Tone-based Solo and Accompaniment Separation

In Section 3.1 of this thesis, a frame-based solo and accompaniment separation algorithm was proposed. Following the analysis of results of the frame-based algorithm, several studies were conducted in Section 3.2 to better characterize the behavior of musical instruments in the spectral domain. In this section, a new tone-based solo and accompaniment separation algorithm is proposed where instead of estimating the spectral parameters of the solo instrument based solely on the information of one frame, all the frames that belong to a given tone are accumulated for processing. Tone-based parameter estimation presents several benefits compared to the originally proposed frame-based processing. Firstly, it allows a meaningful segmentation of the audio signal. Results from the frame-based separation algorithm showed that parameter estimation based solely on the information of one frame, only allowed basic modeling of the solo instrument. Having information from several frames allows the inclusion of more complex temporal modeling as will be explained in the following sections. With this in mind, one might be led to think that any kind of temporal segmentation of the audio signal would allow the same kind of processing. However, specifically segmenting the signal into tones allows one to take advantage of known characteristics of musical tones: It allows to directly address attack sections of the tone, to better model temporal envelopes as for example with the use of Common Amplitude Modulation (CAM), and to incorporate post-processing strategies to remove interferences from other sources. Theses processing stages will be explained in detail throughout this section. Additionally, this kind of segmentation still allows efficient processing with minimal memory requirements as the only spectral information saved in memory is the one that corresponds to the current tone.

It should be noted that extracting tone information before performing separation, brings Automatic Music Transcription (AMT) and sound separation a step closer. Even when these are considered separate fields of research, they greatly overlap in the parameter estimation stage.

As shown in Figure 3.35, parameter estimation is now composed of three processing blocks. After an initial pitch detection stage, the tone formation stage creates tone objects from the delivered $f0(n)$ sequences. A Harmonic Refinement stage follows where an estimate of the spectrum of the solo instrument is obtained on a tone by tone basis. A new post-processing stage is included where the above mentioned tone-based refinements (attacks, CAM, and transient interferences) are conducted. Finally, the original mix is masked and solo and accompaniment signals are obtained.

In the following sections, the different processing stages included in the tone-based separation method are described.

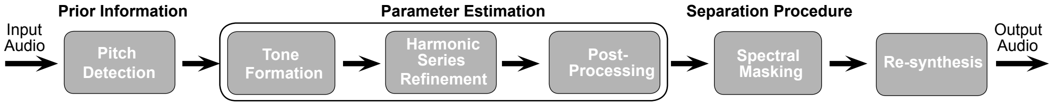


Figure 3.35.: Block diagram of the proposed tone-based solo and accompaniment separation algorithm. It has to be noted that an additional Tone Formation block has been included in the parameter estimation stage if compared to the frame-based separation method in Figure 3.1.

3.3.1. Notation

In this section, a slightly different notation is used for simplicity and readability's sake. The spectral masks created for the solo and accompaniment are denoted here $M_S(k,n)$ and $M_A(k,n)$, respectively. The index k will always denote frequency bins while the index n will denote time frames in the time-frequency representation. The window length will always be denoted by N, and the sampling frequency by fs.

3.3.2. Pitch Detection

As in the frame-based separation algorithm, the method for main melody detection presented in [45] was used as pitch detection front-end. For reference purposes the main characteristics of this method are described again: the algorithm is based on a multi-resolution FFT used to extract pitch candidates based on a pair-wise evaluation of spectral peaks. The algorithm attempts to build voices from the pitch candidates found, by defining a frequency range and a mean magnitude value for each voice. Peaks in the magnitude spectrogram are either assigned to an existing voice—if the peak passes the voice's magnitude threshold and falls in its frequency range—or starts a new voice. The most salient voice is selected as the main melody. The reader is referred to Section 2.3.1 for a detailed description of the method.

Following the notation introduced in Section 3.1, the pitch detection algorithm delivers frame-wise fundamental frequency sequences of the main instrument $f0(n)$. In frames where no solo instrument is detected, the algorithm delivers $f0(n) = 0\,\text{Hz}$.

3.3.3. Tone Formation

The goal of the tone formation stage is to create tone objects from the $f0(n)$ sequence delivered by the pitch detection stage. In this case, instead of estimating parameters on a frame by frame basis, the proposed method accumulates all the spectral frames that correspond to one tone, before performing parameter estimation. In this work, a *tone* is defined as a sound with distinct pitch and duration and it is characterized by its onset frame, offset frame, and frame-wise instantaneous frequency (IF) values.

The raw $f0$ estimates from the pitch detection stage are analyzed over time to create tone objects. When no melody is detected, the pitch detection stage delivers $f0(n) = 0\,\mathrm{Hz}$. A new tone is only started when an $f0$ value in the [65 Hz, 2000 Hz] range is found. This range roughly corresponds to 5 octaves between C2 and B6. After the start of a tone has been detected, a moving average filter of length $L = 3$ frames is used to calculate the mean frequency value $\bar{f}0(n)$ in the time interval defined by the filter length L. That is:

$$\bar{f}0(n) = \frac{1}{L}\sum_{j=0}^{L-1} f0(n-j) \tag{3.20}$$

The end of a tone is defined either by a new $f0(n) = 0$ Hz (no tone was detected) or by a mean frequency variation larger than a semitone (a new tone has started). Low and high semitone intervals from $\bar{f}0(n)$ are calculated using the cent units of measure. The interval is then given by $[\bar{f}0(n)/2^{(100/1200)}$, $\bar{f}0(n) \cdot 2^{(100/1200)}]$. To remove any spurious tones, a minimum tone length of 100 ms which is roughly a 16th note at 140 BPM, is defined. After this stage, each tone object is defined by its initial frame n_i, final frame n_f, and an instantaneous frequency (IF) value for each of the frames in the tone. Capturing frame-wise IF values for each tone allows minor pitch variations that can occur within a tone.

3.3.4. Harmonic Series Refinement

The goal of this stage is to obtain an estimate of the harmonic structure of each tone of the solo instrument. Following the results obtained in the frequency analysis presented in Section 3.2.3.3, each harmonic component is allowed to have an *independent* deviation from the calculated ideal location of the harmonic, i.e., multiple integer of the fundamental frequency. As opposed to the frame-based separation algorithm, no restriction is set to deviations to higher or lower frequencies than the ideal harmonic ones.

In this section, p denotes the partial index, with $p = 1$ representing the fundamental frequency and $p = p_{max}$ representing the highest partial considered in each tone. Additionally, $k_p(n)$ is defined as the frequency bin of the ideal partial location of partial p (calculated as integer multiple of the fundamental frequency). Finally, δ_{max} is defined as the maximum frequency deviation that each partial p is allowed to have from its ideal harmonic location.

For each time frame n in the range defined by $[n_i, n_f]$, where n_i is the initial detected frame of the tone, and n_f is the final frame of the tone, a frequency band given by $[k_p(n) - \delta_{max}, k_p(n) + \delta_{max}]$ is defined where a search for the observed partial location is conducted. An iterative search in the vicinity of the ideal partial location $k_p(n)$ is performed for all partials with partial index $p = 2, \ldots, p_{max}$. For each partial, the search returns the frequency bin $\hat{k}_p(n)$ where the observed harmonic with the largest amplitude is detected. A simple smoothness constraint δ_k is set in the attempt to avoid sudden frequency bin jumps in the harmonic estimation. This constraint is supported by the smoothness observed in the IFDs of different musical instruments described in Section 3.2.3.4. When the refinement stage finds a frequency bin $\hat{k}_p(n+1)$ that implies a larger bin jump than δ_k with respect to $\hat{k}_p(n)$, the smoothness constraint is enforced and the harmonic component is located in $\hat{k}_p(n) \pm \delta_k$. (the sign in the expression is determined by whether the deviation is to higher or lower frequency bins). A detection mask $D(k, n)$, where the observed harmonics are marked with 1 for each frame, is defined for k in the $[1, N/2]$ range:

$$D(k,n) = \begin{cases} 1 & \text{if } k \in \{\hat{k}_p(n), p = 1, \ldots, p_{max}\} \\ 0 & \text{otherwise} \end{cases} \tag{3.21}$$

3.3.5. Spectral Masking

After the complete harmonic series has been estimated, initial binary spectral masks for the solo $M_S(k,n)$ and accompaniment $M_A(k,n)$ are created. At this stage, each time-frequency bin is defined either as part of the solo instrument or part of the accompaniment. To compensate for spectral leakage in the time frequency transform, a tolerance band Δ centered at the observed partial location $\hat{k}_p(n)$, is included in the masking procedure. Thus, for a frequency range $\hat{k}_p(n) - \Delta \leq k \leq \hat{k}_p(n) + \Delta$, and time frame $n \,\epsilon\, [n_i, n_f]$ the masks are defined as follows:

$$(M_S(k,n), M_A(k,n)) = \begin{cases} (1,0) & \forall k, n \,\text{with}\, D(k,n) = 1 \\ (0,1) & \text{otherwise} \end{cases} \tag{3.22}$$

3.3.6. Attack Correction

The pitch detection front-end requires clear peaks in the magnitude spectrogram to accurately detect pitch candidates and create melodic voices. Due to the noise-like characteristics of attacks, clear peaks can usually only be observed in the sustained part of the tone. This causes the pitch detection algorithm to deliver a valid $f0$ only after the attack portion of the tone has passed. This is naturally problematic in a separation context as the attack portions of the solo instrument will always be estimated as part of the accompaniment, creating audible and disturbing artifacts.

Durations of attack transients of musical instruments were studied in [116]. In this study, the authors measured durations of attack transients of different musical instruments under different conditions. Their study showed that attack durations vary from instrument to instrument, from note to note, and from player to player. Average attack durations range from 14 ms to 85 ms with the flute and string instruments having slower attacks ($> 50\,\text{ms}$), and the clarinet, double reeds, and brass instruments faster attacks ($< 50\,\text{ms}$). Results also showed that attack durations tend to decrease with increasing pitch and can vary from one player to the other. Their study also concluded that attack durations are independent of dynamic, length of the note played, and presence or not of vibrato in the note played.

Following the findings of [116] and taking into consideration that this separation algorithm deals with different kinds of solo instruments, an attack duration of 70 ms was used for processing. This value represents a compromise between long and short attacks, and proved to deliver good separation results. Modifying longer regions resulted in audible interference from other sources in the solo track, and shorter regions resulted in audible artifacts in the backing track.

The strategy used to capture the attacks replicates the observed harmonic structure in frame n_i of $M_S(k, n_i)$ in all the frames 70 ms before n_i.

This is a simple but effective solution to better capture attacks of the solo instrument. Figure 3.36 shows four example spectrograms of tones played with different musical instruments. Special care was taken to clearly display the attack portions of each tone. It can be seen that all musical instruments have attacks with very different spectral characteristics. The piano and trumpet show more broadband characteristics than the flute and the saxophone for example. The saxophone shows clearly marked subharmonic components and the flute in this case, shows a very harmonic structure with minimal noise-like components. The great diversity of spectral characteristics of attack sections added to the differences in attack durations for all instruments, makes finding an unified approach for attack estimation, very challenging. This is the reason why a conservative approach was taken where only the observed harmonic structure in time frame n_i is replicated and no attempt is made to better represent the transient-like characteristics of attacks. The approach of creating a noise spectrum in the attack frames for example, resulted in audible interferences in the solo signal.

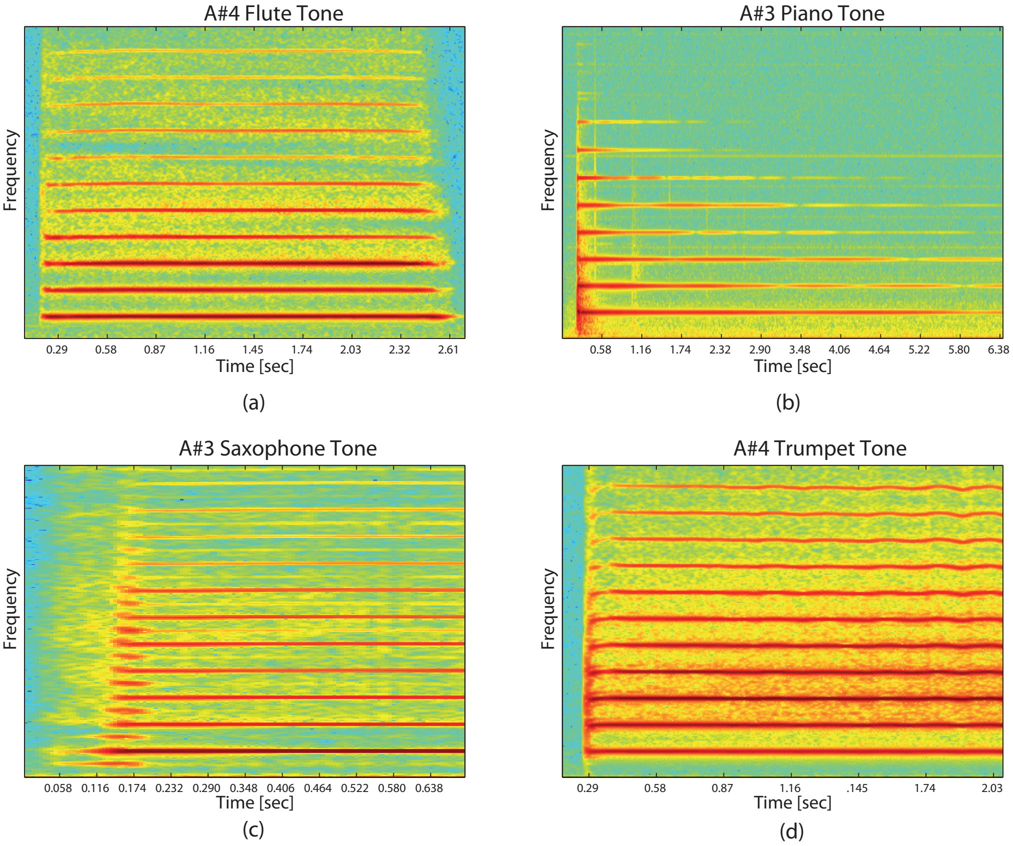


Figure 3.36.: Example spectrograms showing attack sections from different musical instruments. (a) A#4 flute tone. (b) A#3 piano tone. (c) A#3 saxophone tone. (d) A#4 trumpet tone. The great differences between the attack sections of different musical instruments can be clearly observed.

A final note has to be added about the importance of tone-based processing for the attack refinement stage. To be able to perform attack refinement, clear information about the estimated start frame of each tone n_i, is needed. This is only possible when segmentation based on tone objects is used in the separation approach.

3.3.7. Post-processing of Tone Objects

In a polyphonic music context, sounds from different musical instruments merge into a single piece of music to produce an unique sonority and texture. In Western music in particular, where equal temperament and harmony based on major and minor chords is used, it is particularly likely that musical sources overlap in the frequency domain. This is naturally very evident in a time-frequency representation where energy from different sources often falls in the same frequency bands. In such cases, telling one sound source from the other is not a straight forward task.

Many studies have been conducted that address the problem of overlapping of spectral content from different sources. Source separation is in itself a matter of resolving overlapping components from different sources. Systems like [108, 89] have attempted to resolve overlapped components by using pitch information from all the sources in the mix. If the pitch of each instrument is known at all times, it can be easily predicted where spectral collisions are likely to occur. Once the collisions have been detected, clean information from the temporal and spectral envelopes or trained systems with instrument specific information are used to resolve the collisions.

A much bigger challenge is encountered by systems where pitch information from all the sources is not available before processing. To date, performance of multipitch detection systems is not accurate enough to allow robust separation with a wide range of music signals [117]. Therefore, unless musical scores from the signal to separate are available, having reliable pitch information from all the sources is a difficult task. An additional challenge is posed by percussive instruments that also overlap with harmonic instruments in the spectral domain. The transient and noise-like characteristics of percussive instruments in the spectral domain, add a higher difficulty level to the separation problem.

The separation method proposed in this thesis extracts pitch information from the solo instrument only. Thus, predicting time-frequency regions where collision can occur is not possible. To be able to detect and reduce interference from other sources in the estimation of the solo instrument, strategies that can accurately perform without the use of pitch information from all the sources, are required.

In this post-processing stage, two strategies to reduce interferences from other sources in the solo estimation are proposed. (1) The first one addresses transient and percussive interferences that might have been erroneously estimated as part of the solo instrument. (2) The second strategy, addresses interferences from other harmonic sources. (3) Additionally, a strategy to better estimate non-harmonic elements that can occur in musical instrument tones such as fricatives in the singing voice, is proposed based on phase expectation.

The most important feature of the post-processing stage is that it is completely based on modeling of musical instrument tones. As will be explained, only known information about the temporal evolution of musical instrument will be used at this stage. Here, the importance of the tone-based separation proposed in this thesis becomes once again very clear.

(1) Removing transient and percussive interferences from the solo signal

As explained in Section 3.2.3.4 and shown in Figure 3.33, studies on percussive instruments have shown that percussion onsets are evident in the spectrogram as vertical events occurring in a short time interval. This broadband characteristic of percussive events naturally results in a large degree of overlapping of spectral components from the solo instrument and the percussive ones. In Figure 3.37 an example of the estimation of a saxophone tone after Harmonic Series Refinement is shown. In the figure, interferences from percussive instruments can be clearly observed. The red arrows indicate two percussive hits that were initially estimated as part of the solo signal. The mentioned vertical characteristics of percussive events can be clearly observed. Even when the magnitudes of these events are not particularly large in comparison to the lower partials of the tone, the perceptual impact of such events is considerable, being in most cases clearly audible and disturbing. It can be observed that these events are common to all harmonics and occur in a short interval of time.

To detect these transients, an approach based on the method for harmonic/percussive separation described in [115], is proposed. The main idea is to detect vertical events

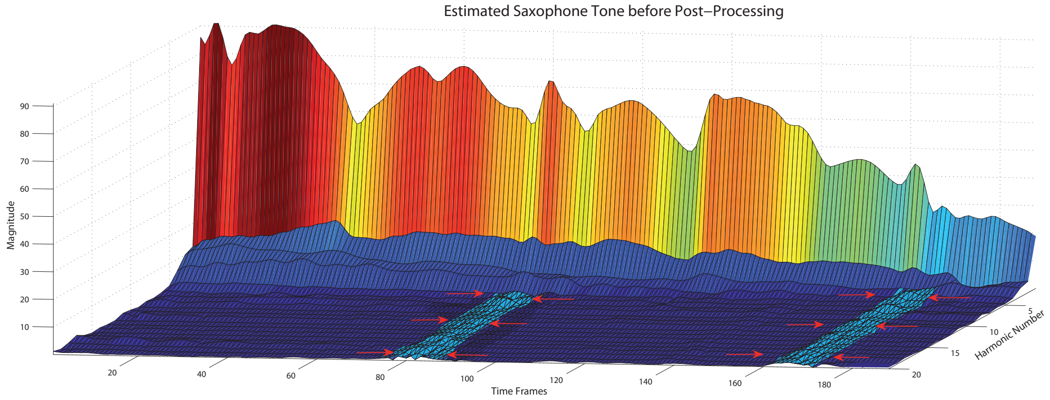


Figure 3.37.: Estimated saxophone tone before post-processing. The red arrows indicate the places where elements from two percussive events are mistakenly estimated as part of the solo instrument.

in the spectrogram by detecting large amplitude variations common to several partials. The approach works as follows: First, the temporal envelope of each partial is smoothed with a median filter of length L. Each smoothed envelope is then normalized to the $[0, 1]$ range. Second, a magnitude threshold γ_L is defined and all the time frames where the normalized smoothed magnitude envelopes have amplitudes larger than γ_L, are detected. To guarantee that the detected events are indeed percussive transients, the events should simultaneously happen in several partial (if not in all). The minimum number of partials min_p where the event should occur is defined, and only the detected time frames that are common to at least min_p partials are kept as possible transients. As the perceptual impact of such events is stronger for higher frequencies, only the magnitude envelopes of partials with partial index $p > p_{low}$, are modified. For partials with partial index $p < p_{low}$, magnitude values are in general considerably large (as lower partials are always stronger). For this reason, percussive transients are better masked by the magnitudes of the partials of the solo instrument. Tests conducted showed that modifying the temporal envelopes of partials lower than p_{low} was detrimental to the perceptual quality of the solo instrument and thus, these envelopes are kept untouched.

To remove the detected transients, the value of the solo spectral mask $M_S(k,n)$ in the time frames where the transient was detected is modified. The mean magnitude value of the normalized smoothed magnitude envelope in the L time frames before the transient was detected, is obtained for each partial. Bearing in mind that the initial mask $M_S(k,n)$ is binary, the initial 1 in the binary mask is replaced by the mean magnitude value in the $[0,1]$ range obtained. This introduces a smoothness constraint in the temporal envelopes of the partials. The number of time frames that determine the smoothness constraint of the temporal envelopes is given by the filter length L. The accompaniment mask is also recalculated as $M_A(k,n) = 1 - M_S(k,n)$. It should be noted that the new post-processed spectral masks are no longer binary.

Another important processing consideration taken was to remove from the transient detection stage, the attack frames considered in the Attack Correction stage (Section 3.3.6) . As already mentioned, attack portions of instrument tones often show transient characteristics. By removing the attack frames from post-processing, the transient-like characteristics of the attacks that might have been captured in the Attack Correction stage, are preserved. This approach has the disadvantage of failing to remove percussion hits in the solo signal that coincide with the attacks of the solo instrument. Nevertheless, it presents a good compromise between removing percussive interferences from the sustained part of the tones (where they are most audible), preserving attack portions of the tone, and minimizing the interference from the solo instrument in the backing track.

The effect of the transient removal stage can be observed in Figure 3.38 where the same saxophone tone from Figure 3.37 is shown after post-processing. The location of the detected percussive events are marked with red arrows and with a slightly lighter blue color in both plots. It can be seen that this processing stage guarantees a degree of smoothness in the temporal envelopes of the tone.

The transient removal stage is a lightweight but effective approach to remove transients from the solo signal. It has the benefit of only being performed in the time frames where the solo instrument has been detected and removes the need of performing a previous harmonic/percussive separation to avoid percussive interference in the solo signal.

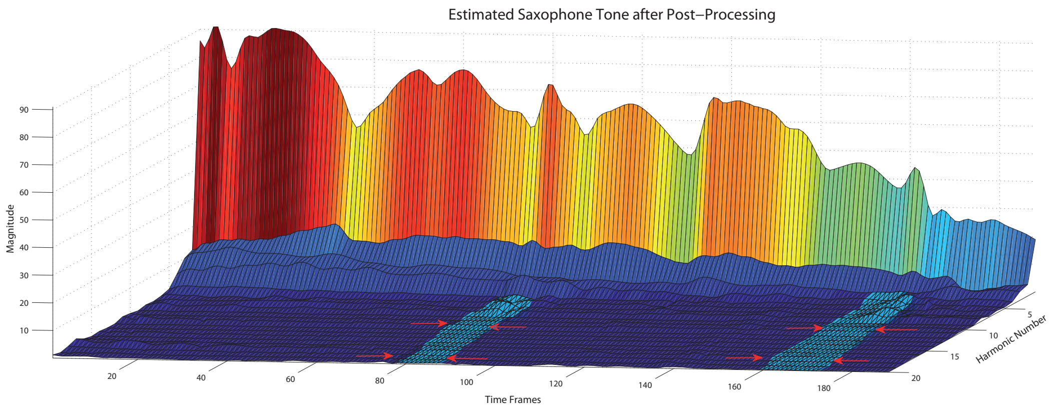


Figure 3.38.: This figure displays the saxophone tone shown in Figure 3.37 after the post-processing stage has been applied. The red arrows indicate the places where elements from two percussive events had been originally assigned to the solo instrument. As can be observed, the post-processing stage greatly reduces the interference from percussive hits in the solo signal.

(2) Minimizing effects of interference from other harmonic instruments in the solo signal: Data-driven Common Amplitude Modulation (CAM)

Amplitude envelopes of musical instruments (both temporal and spectral envelopes) have proven to be difficult to model. Finding a reference curve that could serve as a template envelope for a given musical instrument has proven to be unreliable and inaccurate [108]. There are many factors that contribute to the shape of the envelopes including instrument model, performer, and dynamics among others. However, even when finding a reference curve that accurately represents a musical instrument is not possible, harmonic components of the same source exhibit similar temporal envelopes that can be highly correlated. In other words, even when accurately predicting the shape of the magnitude envelopes of a source is difficult, this shape is expected to be common to all the harmonic components of the musical instrument tone. This is known as Common Amplitude Modulation (CAM) and it is an important cue in human auditory perception [111]. With CAM being an observed characteristic in musical instrument spectra, some separation approaches have attempted to include CAM in their processing chains. In [46] for example, CAM is used as a means to resolve overlapped harmonic components in a least squares estimation framework. In [89], the authors propose a spectro-temporal modeling of harmonic magnitudes and test their method on isolated instrument notes. They also test their estimation algorithm in the separation context by creating random mixes of a maximum of 6 instrumental tones. Both [46] and [89] require prior information of the pitch of all the sources in the mix to be able to detect time-frequency sections where sources are likely to overlap.

To impose CAM in the estimation of solo signals, it is necessary to first obtain a reference temporal envelope that all the harmonic components of the tone should follow. However, as opposed to [46] and [89] where prior knowledge of the $f0s$ of all the sources allowed the differentiation between clean and overlapped envelopes, this prior information is not available in the method proposed in this thesis. In the solo/accompaniment separation method here proposed, determining where harmonic components overlap is not plausible without having a good idea of the spectral content of the other sources. Similarly, extracting clean envelope information from at least one of the harmonics is not straight-forward either as the presence of other sources is impossible to predict. Consequently, the use of CAM is proposed in a different way. To introduce CAM, the magnitude envelope of the partial which is most similar to the envelopes of the other partials is found, and used as a reference to impose CAM in the spectral estimation. The method works as follows: (1) The temporal magnitude envelopes of all partials are estimated as described in Section 3.3.4. (2) As estimation of lower partials is more robust than of higher ones (as lower partials are always stronger), only the first p_{CAM} partials are used at this stage. The goal is to find the partial p among the first p_{CAM} partials, whose temporal envelope has the highest mean cross-correlation with the other $p_{CAM}-1$ envelopes. For this matter, the cross-correlation r_{ij} between the temporal envelopes of the *i-th* and *j-th* partials for all $i, j \leq p_{CAM}$ and $i \neq j$, are calculated. (3) The mean cross-correlation $\bar{r}_p$ for each partial is calculated by averaging the $p_{CAM}-1$ cross-correlation coefficients r_{ij} obtained for each partial. The partial with the maximum mean cross-correlation $\bar{r}_p$ is taken as the reference. That is, $p_{ref} = \underset{p}{argmax}\,(\bar{r}_p)$. (4) The temporal envelope of the reference partial is normalized to the $[0, 1]$ range and used as a weighting function for all the other partial envelopes. Even when only the first p_{CAM} partials are used to obtain the reference envelope, the temporal envelopes of all partials $p = 1 \cdots p_{max}$ are weighted with the reference curve.

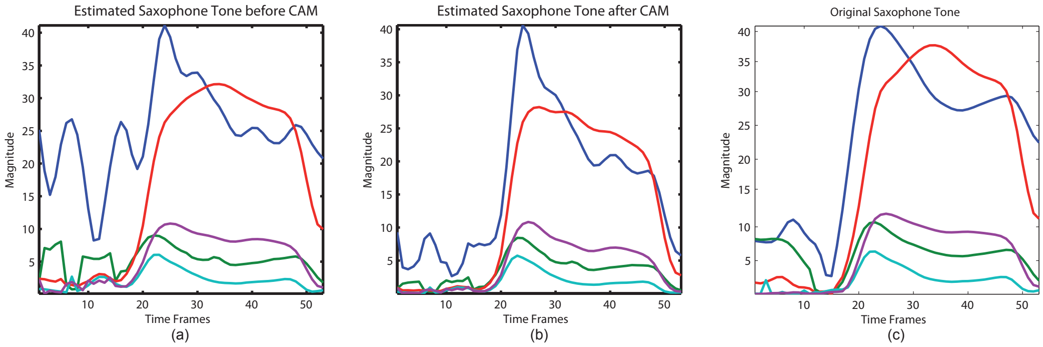


Figure 3.39.: Effects of common amplitude modulation (CAM) on the estimation of a saxophone tone. The three plots show the temporal envelopes of the 5 first partials of the tone: F0 (blue), p1 (red), p2 (magenta), p3 (green), p4 (cyan). (a) Initial estimation of the tone before CAM. (b) Estimated tone after CAM has been applied (c) Original saxophone tone extracted from the original saxophone recording (ground-truth). The greater similarity between the envelopes in (b) and (c), especially in the early time frames, can be observed.

The effects of imposing CAM in the spectral estimation of a saxophone tone are shown in Figure 3.39. For visualization purposes, only the first five partials of the tone are shown. In Figure 3.39a, the estimated tone before CAM is displayed. In Figure 3.39b, the estimated tone after CAM is shown. Finally, in Figure 3.39c, the original tone taken from the original saxophone recording (ground-truth) is shown for reference. It can be seen that the use CAM results in temporal envelopes closer to the original ones. Particularly noticeable is its effect on the $f0$ envelope (blue curve) where the estimation of the initial time frames of the tone are clearly affected by overlapping of spectral components of other sources. This causes the initial estimation to show considerable differences with the original tone. However, the use of CAM reduces the impact of initial mis-estimations and results in solo signals with reduced interference from other sources.

(3) Capturing non-harmonic elements of the tones using phase expectation:

The concept of phase expectation as described in Section 3.2.3.4, is used to detect time frames where phase values show non-harmonic behavior in the estimated spectral content of the tone. Detecting these events is particularly important for vocal sounds as it allows for example, to better capture characteristic fricative and plosive sounds.

The phase spectrum of all the partials p with frequencies $f_p > 3000\,\mathrm{Hz}$ is evaluated using Eq. (3.19). For all the partials considered, the time frames whose phase cannot be explained with Eq. (3.19), and that do not exhibit phase values characteristic of any of its adjacent bins, are detected. As described in Section 3.2.3.4, due to spectral leakage in time-frequency representations, the frequency bins adjacent to a main spectral peak tend to exhibit phase values characteristic of the main peak and as such, expectation values fall outside the calculated radian ranges. This phenomenon can be easily observed in Figure 3.32. These bins are not used in this processing stage and only phase values that cannot be explained under any circumstances are detected. Only partials with frequencies $f_p > 3000\,\mathrm{Hz}$ are considered as initial tests showed that taking lower partials can result in increased interference in solo tracks. In all the time frames where at least NH_p partials show non-harmonic behavior, a noise spectrum starting at 3000 Hz and ending at the highest partial frequency is created in the solo mask $M_S(k,n)$. All the frequency bins between 3000 Hz and the highest partial are replaced by a random number in the [0,1] range. This creates a noise-like spectrum in the detected time-frames.

3.3.8. Experiments and Results

3.3.8.1. Analysis of Pitch Detection Algorithms

In Section 3.2.3, an analysis of musical instrument tones was conducted with the goal of better characterizing the behavior of musical instruments in the spectral domain. With the findings from this study and the preliminary results obtained with the frame-based separation algorithm, a tone-based method for solo and accompaniment separation was proposed. Going back to the general block diagram of an informed sound separation method (shown again in Figure 3.40 for reference), prior information plays a very important role in the final separation results.

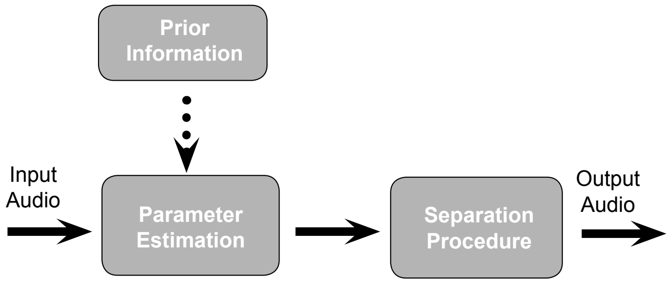


Figure 3.40.: Block diagram of the general structure of an informed sound source separation algorithm. Three main stages compose the entire separation process: prior information, parameter estimation, and the final separation procedure. Pitch detection resides inside the Prior Information block.

In the particular case of a pitch-informed separation algorithm as the one described in this thesis, the performance of the pitch detection front-end plays a very critical role. As shown in the block diagram of the tone-based separation algorithm displayed in Figure 3.35, with such an algorithm structure, any errors in the pitch estimation will directly propagate to the separation stage. It is naturally of great importance to guarantee that the pitch information delivered to the parameter estimation stage, is as accurate and robust as possible.

In this section, the performance of the pitch detection stage of the proposed method is addressed by comparing two pitch detection front-ends. These two algorithms were chosen as they have shown superior performance in the Music Information Retrieval Evaluation eXchange (MIREX) of previous years. The reader is referred to [12] for a thorough comparison of the performance of pitch detection algorithms in past MIREX campaigns. The two compared methods are: (1) Pitch Estimation by Pair-Wise Evaluation of Spectral Peaks [45], and (2) Melody Extraction using Pitch Contour Characterization [12]. The two algorithms were already described in Section 2.3.1 of this thesis and here, only their main characteristics are provided. These two algorithms will be referred to as *Alg1* and *Alg2*, respectively. Table 3.6 presents in a comparative manner the main processing stages of the two methods. It should also be noted that *Alg1* is the method used in both the Frame-based and in the Tone-based separation algorithms presented in Sections 3.1 and 3.3, respectively.

Table 3.6.: Comparative table of the two pitch detection algorithms evaluated in this section. The main processing stages of each algorithm are listed in a comparative manner.

Processing Step	**Alg1: Dressler**	**Alg2: Salamon**
Spectral Representation	Multi-resolution FFT	STFT
Peak Characterization	IF and Magnitude	IF and Magnitude
Freq. Range	55Hz - 5kHz	55Hz - 1.76kHz
Pitch Candidate Selection	Pair-wise peak evaluation	Salience function
Time Evolution of Pitch Candidates	Tone formation & Voice Formation	Pitch Contours
Extra Processing	–	Voice/unvoiced & Octave error
Main Melody Selection	Salience	Salience

The relevance of this evaluation of the pitch detection algorithms is two-fold. On the one hand, the study evaluates the two algorithms directly in the separation context. Namely, the evaluation attempts to determine which of the two algorithms results in better separation quality. With this in mind, it is important to note that a thorough evaluation of pitch detection accuracy (using ground-truth pitch information and frame-wise evaluation) goes beyond the scope of this study and only their performance as front-ends of the separation method is evaluated. On the other hand, this evaluation was designed in a way that performance differences between vocal and instrumental tracks can be addressed. This particular distinction in the evaluation was motivated by the fact that musical instruments and the singing voice show very different spectral characteristics. Previous results has shown that these differences affect the performance of pitch detection algorithms and as such, this study is of great relevance.

Implementation Details

The evaluation of pitch detection performance was conducted using the Tone-based Solo and Accompaniment separation method presented in Section 3.3. The tone-based separation algorithm used for this evaluation did not include the non-harmonic element detection stage described in the Section 3.3.7 and included as slightly different Harmonic Refinement stage. For *Alg1*, a C++ implementation was used that delivers $f0(n)$ sequences as output. The resulting sequences are then used as input to the separation scheme. For *Alg2* the available *VAMP* plug-in for *Sonic Visualiser* [118, 119] was used, and annotations were used as inputs to the separation algorithm. For *Alg2*, the following processing parameters were used: given a sampling frequency $fs = 44100\,\mathrm{Hz}$, an analysis frame of 46 ms with a hop size of 2.9 ms was used. The Voicing Tolerance parameter was set to 0.2.

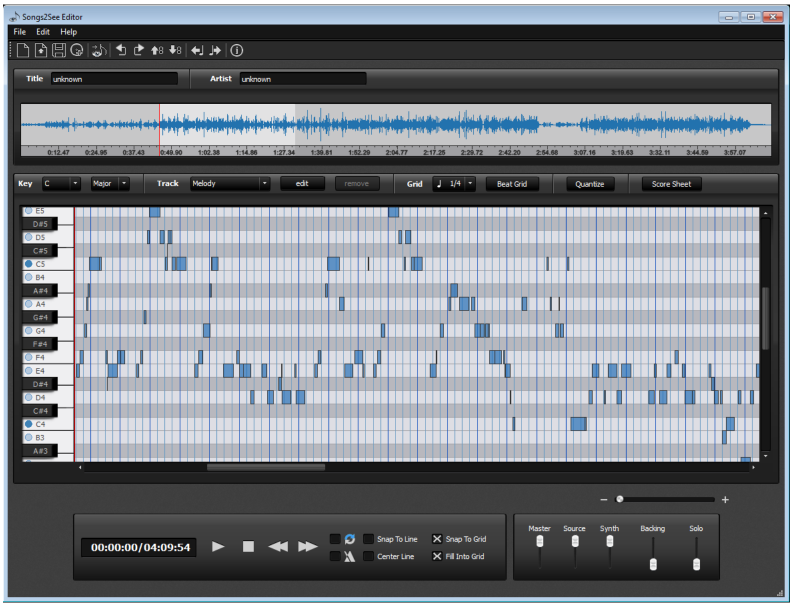
The processing parameters used in the separation algorithm for this evaluation were: $p_{max} = 20$, $\Delta = 1$, $p_{low} = 9$, $min_p = 6$, $L = 5$, $\delta_k = 2$, and $\gamma_L = 0.6$.

Dataset

For this evaluation, the dataset SA_DS1a first introduced in Section 3.1.6 was used. It is composed of 17 copyright free tracks, 10 of which are vocal and 7 instrumental. A full description of this dataset is presented in Appendix B of this thesis.

Evaluation

The SA_DS1a dataset was processed with both pitch detection algorithms and the resulting $f0(n)$ sequences were used as input to the tone-based separation algorithm. Additionally and with the goal of having a clear performance boundary, ground-truth pitch information was obtained for the full dataset using the Song2See Editor presented in [O2] and briefly described in Section 4.3.1 of this thesis. The Songs2See interface allows the manual refinement of the results provided by the automatic pitch detection feature in the software by deleting, inserting, and correcting pitch and time information of all note objects. In Figure 3.41, a screenshot of the Songs2See Editor is presented. The automatic pitch detection functionality in the software delivers pitch estimates of the solo instrument in the form of note objects displayed as blue horizontal bars. The user is allowed to delete notes, create new ones, or modify existing ones in duration and pitch.



Title: unknown
Artist: unknown
Key: C Major
Track: Melody
Grid: 1/4
00:00:00/04:09:54

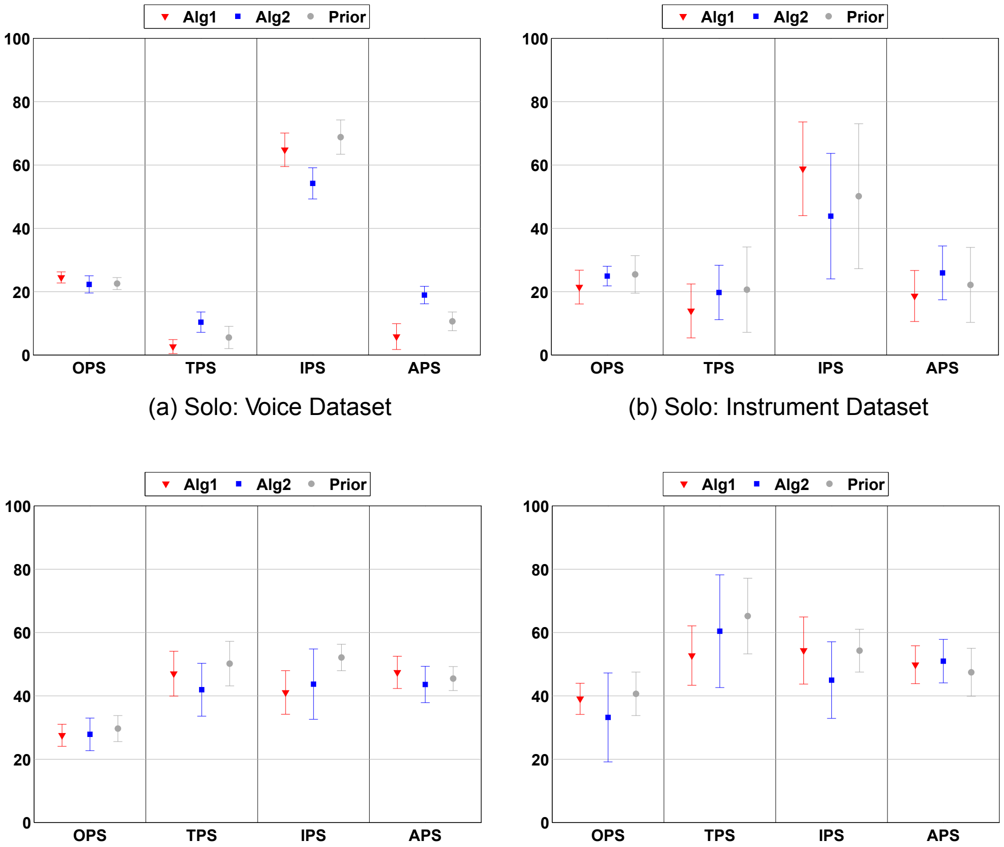
Figure 3.41.: Screenshot of the Songs2See Editor. The Songs2See Editor was used to manually create ground-truth pitch information. Each blue bar in the piano-roll representation is a note played by the solo instrument.

The full dataset was processed with the Songs2See Editor and results were refined and corrected by musical experts. The refined pitch sequences were used as ground-truth pitch information in this study.

To evaluate separation results the PEASS Toolkit was used and objective perceptual measures were calculated. Figure 3.42 presents mean values and 95 % confidence intervals of the results obtained for *Alg1*, *Alg2*, and with ground-truth pitch information (referred to as *Prior* in the figure). In this section, the terms accompaniment and backing tracks are used interchangeably. To better understand performance differences for the solo and backing tracks, results are presented independently for each of them. Furthermore, to assess the differences between vocal and instrumental performance, results are also presented separately for vocal and instrumental tracks.

Analysis of Results

With respect to general separation quality, results show minor overall performance



(a) Solo: Voice Dataset

(b) Solo: Instrument Dataset

(c) Backing: Voice Dataset

(d) Backing: Instrument Dataset

Figure 3.42.: Overall Perceptual Score (OPS), Target-Related Perceptual Score (TPS), Interference-Related Perceptual Score (IPS), Artifact-Related Perceptual Score (APS). Results with the two pitch detection algorithms (Alg1, Alg2) and with ground-truth pitch information (Prior) are presented independently for the vocal and instrumental datasets. Mean values with 95% confidence intervals are presented. (a) Solo: vocal dataset. (b) Solo: instrument dataset. (c) Backing: vocal dataset. (d) Backing: instrument dataset. The OPS scores obtained by both algorithms are very similar.

differences between the two algorithms, obtaining in general comparable Overall Perceptual Scores (OPS). For both algorithms, scores obtained for the instrument dataset are higher than the ones obtained with the voice dataset. These results suggest that independently of the pitch detection method used, the proposed separation method can better handle instrumental signals than vocal ones. However, it should also be noted that larger confidence intervals are also obtained for the Instrument dataset, suggesting that some instruments can be better handled than others.

A few important differences between the two algorithms can also be outlined:

- *Alg2* shows in general larger confidence intervals than *Alg1*. These results suggest that the performance of *Alg1* is more consistent across signals with different acoustical and timbral characteristics.

- For the task here addressed, the IPS of the backing tracks can be considered a rough indicator of the quality of pitch detection. With the solo instrument being the only source of interference for the backing track, high IPS scores for the backing indicate that very little content from the solo signal remained in the backing track. *Alg1* obtained a slightly higher IPS score for the backing tracks with the instrument dataset. *Alg2* obtained a higher IPS score for the backing tracks with the voice dataset. Both of these IPS scores are slightly lower than the ones obtained with *Prior* (ground-truth information), which represents the performance boundary for the proposed separation scheme.

- It is particularly noticeable that in some cases, *Alg2* results in higher APS and TPS than *Alg1*, sometimes even higher than the ones obtained with *Prior* (ground-truth information). These results might seem surprising but careful analysis of the extracted audio tracks show that *Alg2* tends to benefit more continuous pitch contours. This allows the spectral estimation to better characterize each of the tones and to capture more accurately their attacks and releases. This comes at the cost of slightly lower IPS scores for the solo track.

- Analysis of the resulting signals suggest that *Alg2* can discriminate more accurately voiced from unvoiced segments in the vocal tracks but octave errors occur more often.

Conclusions

In this section, a comparative evaluation of two pitch detection methods within a separation context has been presented. The two algorithms perform very similarly; however, slight differences might benefit the use of one over the other for certain applications. The fact that each algorithm maximizes different quality measures, makes them optimal in different scenarios. *Alg1* shows very robust performance under different types of signals, slightly favoring instrumental tracks and resulting in reduced interference from other sources. In contrast, *Alg2* slightly outperforms when dealing with vocal signals. It results in reduced artifacts (in comparison to *Alg1*) but with slightly more interference from other sources.

3.3.8.2. Tone-Based Algorithm Performance

Implementation Details

For the evaluation of the tone-base separation algorithm, the following processing parameters were used: Algorithm *Alg1* was chosen as a pitch detection front-end as it showed to be more robust to different types of signals, genres, and instrumentations. The total number of estimated partials per tone was set to $p_{max} = 20$. A tolerance band was set to $\Delta = 1$. Larger values of Δ would result in perceptible interference from other sources in the target source. A $\delta_k = 2$ was set. For the post-processing stage, $p_{low} = 9$ was selected as higher values were shown to be too restrictive and failed to remove certain percussive interferences. The minimum number of partials used for transient detection was set to $min_p = 6$. A filter length value $L = 5$ with $\gamma_L = 0.6$ were used as they showed to be a good balance between proper handling of spectral leakage and magnitude variations in magnitude envelopes. A value of $NH_p = 4$ was used for non-harmonic element detection. An analysis frame of 46 ms is used in conjunction with a hopsize of 5.8 ms.

Dataset

For the evaluation of the algorithm, the **SA_DS1** (Solo and Accompaniment DataSet

1) was used for evaluation. A full description of all the tracks, as well as the copyright and availability information is provided in Appendix B. Here, only a general description is provided. The dataset has a total of 27 track and is composed of two parts: (1) SA_DS1a with copyright free tracks. A total of 17 tracks, 10 with voice as main instrument, and 7 instrumental tracks are included. (2) SA_DS1b composed of 10 commercial instrumental tracks with saxophone as the solo instrument. This dataset was also used for the evaluation of the frame-based separation algorithm presented in Section 3.1.6.

All the tracks in SA_DS1a were mixed from the available multi-track recordings into a solo signal, an accompaniment track, and a final mixture. Solo and mixed accompaniment tracks were already available in the commercial distribution of SA_DS1b.

Evaluation

The PEASS Toolkit was used for evaluation of separation results. Mean values with 95% confidence intervals for the entire dataset are presented in Figure 3.43. For reference purposes and to allow future comparison of results, the full table of results with independent scores for each track are presented in Appendix C of this thesis.

For the solo tracks, an OPS of 24.78 was obtained. For the backing tracks, slightly better results were obtained with an OPS of 34.68. Similar to the results obtained with the frame-based separation algorithm (see Section 3.1.6.3), particularly high IPS scores were obtained both for the solo and the backing tracks. These results suggest an effective estimation of the spectral information of the solo instrument.

Particularly noticeable is the improvement in the TPS and APS scores obtained for the solo instrument in comparison to the frame-based separation algorithm. The TPS of the solo shows an improvement from 18.07 in the frame-based algorithm, to 23.91 in the tone-based separation algorithm. This evidences that the strategies included in the harmonic series estimation and in the post-processing stage were effective in improving the estimation of the spectral content of the solo signal. The improved TPS and APS scores for the solo come at the cost of a slightly lower IPS for the solo (from 53.77 to 47.46). However, it is in general desirable to

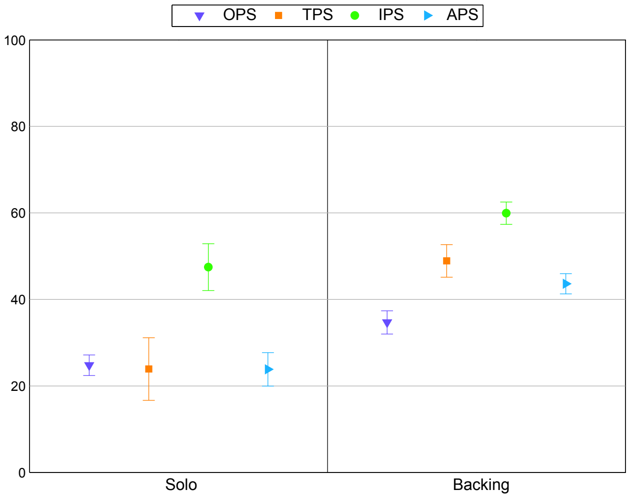


Figure 3.43.: Results obtained with the tone-base separation algorithm for the entire dataset. Overall Perceptual Score (OPS), Target-Related Perceptual Score (TPS), Interference-Related Perceptual Score (IPS), Artifact-Related Perceptual Score (APS). Mean values with 95% confidence intervals are presented. It can be observed that the algorithm results in particularly high IPS scores for both the solo and the backing tracks, indicating good isolation of the solo instrument.

obtain an homogeneous set of perceptual measures as it is an indication of a well-balanced algorithm that does not emphasize too hardly one measure over the others. The processing stages included in the tone-based separation algorithm resulted in a more homogeneous set of measures for the solo in comparison to the frame-based algorithm.

All the confidence intervals obtained with the tone-based algorithm are smaller than the ones obtained with the frame-based method. This proves that robustness in the estimation was improved and the algorithm can more uniformly handle the different musical instruments.

To better assess the performance differences between vocal and instrumental tracks, independent results for the two datasets are shown in Figure 3.44. The most notable improvement with respect to the frame-based algorithm are the TPS and APS scores

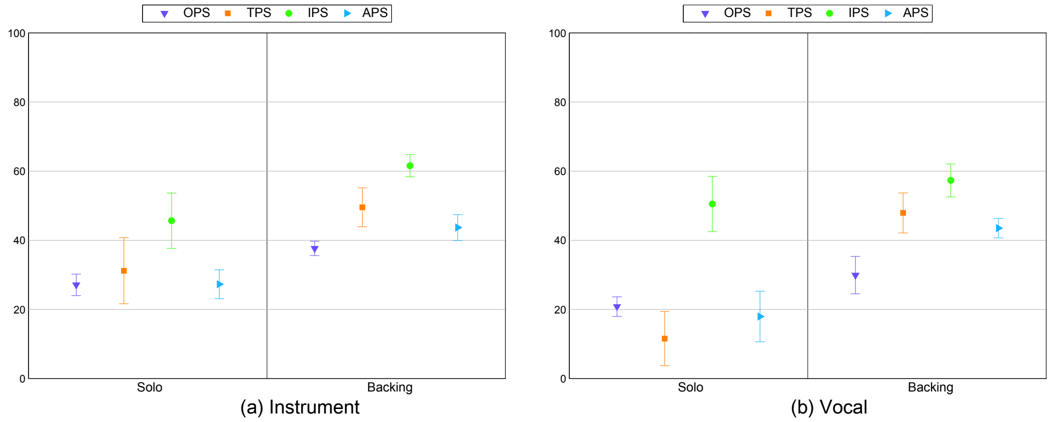


Figure 3.44.: Results obtained with the tone-based separation algorithm displayed independently for instrumental and vocal tracks. (a) Instrument (b) Vocal. Overall Perceptual Score (OPS), Target-Related Perceptual Score (TPS), Interference-Related Perceptual Score (IPS), Artifact-Related Perceptual Score (APS). Mean values with 95% confidence intervals are presented. Results for the backing tracks in both datasets are slightly better than for the solo tracks. Vocal solo extraction obtains slightly lower scores than instrument solo extraction, particularly noticeable in the APS and TPS scores.

obtained for the vocal solo. The TPS for the vocal solo shows an improvement from 4.34 to to 11.56, and an APS improvement from 8.07 to 17.97. This results show that the non-harmonic element detection stage included in the tone-based separation algorithm is very effective in capturing fricatives and plosives of the vocal tracks which results in increased TPS scores. The backing tracks obtained for the vocal dataset show improved OPS, TPS, and APS scores with respect to the frame-based algorithm. The instrumental dataset also shows improvement in the OPS, TPS, and APS scores of the solo tracks with respect to the frame-based separation algorithm. The achieved improvement in the APS scores for both the vocal and instrumental datasets shows that the transient detection stage is effective in removing erroneous estimations in the solo instrument. The general improvement in most of the measures comes at a cost of slightly lower IPS score for both the vocal and instrumental solo and backing tracks.

The effects of the different processing blocks on the perceptual quality of separated tracks were studied in [O5]. Results indicate that the largest perceptual quality gain is achieved during the post-processing stage of the proposed algorithm.

Conclusions

The results presented in this section show that the processing stages included in the tone-based separation algorithm have contributed to the improvement of perceptual quality of separated tracks while still allowing efficient performance with minimized processing delays:

> *The tone-based separation algorithm results in Overall Perceptual Scores (OPS) for the solo and backing track of 27.12 and 37.36, respectively. Results show clear improvement with respect to the frame-based separation algorithm, particularly in terms of the TPS and APS scores of the solo signals. The goal of obtaining solo and backing tracks of similar quality is being reached with the additional benefit of having obtained more homogeneous sets of measures without any notorious differences between the perceptual scores. Smaller confidence intervals were obtained with the tone-based algorithm which shows improved robustness to different kinds of signals. The algorithm has an average processing time of* $0.25 \cdot l$*, where* l *is the total length of the song, on a 2.66GHz computer, making it suitable for real world applications.*

3.3.9. Proposed methods vs state-of-the-art algorithms

In this section, a performance comparison between different versions of the proposed method for solo and accompaniment separation and other state-of-the-art algorithms is presented. In particular, results from the two Signal Separation Evaluation Campaigns (SiSECs) that took place during the development of this work are presented. The SiSEC campaign takes place every two years and in 2011 and 2013 intermediate versions of the algorithm described in this thesis were submitted. A brief description of the details of the algorithm versions submitted for each campaign is presented in the corresponding sections. Additionally, a performance comparison between the tone-based separation algorithm presented in Section 3.3 and the algorithm proposed in [29] is presented.

3.3.9.1. SiSEC 2011

The frame-based separation algorithm presented in Section 3.1 of this thesis with an additional post-processing section to reduce transient artifacts as described in [O5] was submitted to the *Professionally Produced Music Recordings* task in SiSEC 2011

For this campaign, a total of seven algorithms were submitted and evaluated under a common dataset. In this section, only the results from the proposed method and the algorithm submitted by Durrieu [65] are presented. The Durrieu algorithm was chosen for comparison as it works under the same processing conditions as the proposed method: The two methods attempt to solve the single-channel solo/accompaniment separation problem using only pitch as prior information.

The full table of results including other methods submitted that work under different processing conditions can be found in the campaign's website [120]. As described in Section 2.3.1, the algorithm proposed by Durrieu uses an instantaneous signal model which represents the audio signal as the sum of a signal of interest, i.e., the lead instrument, and a residual, i.e., accompaniment. A source-filter model is used to represent the signal of interest. Information from the source is related to the pitch of the lead instrument and information from the filter is related to the timbre of the instrument. The residual is modeled using Non-negative Matrix Factorization (NMF).

In Table 3.7, the resulting scores for SiSEC 2011 of the two algorithms are presented. It has to be noted that the dataset used for this campaign was entirely composed of commercial vocal tracks. As the campaign only evaluates the quality of single sources, the backing tracks obtained were not evaluated. Consequently, the results presented in the table are for solo extraction on the test dataset only.

It can be seen that the proposed method presents OPS values comparable but slightly lower than those obtained by Durrieu. Both algorithms show in general high IPS values that suggest a successful isolation of the main melody. The proposed method outperforms the method proposed by Durrieu in terms of IPS but obtains slightly lower values for the TPS and APS.

Table 3.7.: Results from the SiSEC11 Evaluation Campaign on the test dataset. Overall Perceptual Score (OPS), Target-Related Perceptual Score (TPS), Interference-Related Perceptual Score (IPS), Artifact-Related Perceptual Score (APS)

	Cano	**Durrieu**
OPS	24.1	26.3
TPS	30.4	54.2
IPS	59.1	46.7
APS	27.9	44.3

While the algorithm proposed by Durrieu reports an average processing time of 600 sec per excerpt, the proposed method requires an average of 8 sec per excerpt. This means that the proposed method is 75 times faster than the Durrieu algorithm. Results from this campaign show that the proposed method accomplished its design goal of finding a balance between performance and algorithm efficiency.

3.3.9.2. SiSEC 2013

The performance of the proposed tone-based separation algorithm (without Non-harmonic Element detection and a slightly different Harmonic Series Refinement stage) was compared to state-of-the-art approaches under the *Signal Separation Evaluation Campaign (SiSEC 2013)* in the *Professionally Produced Music Recordings* task. A total of 15 algorithms were submitted and evaluated under a common dataset. In Table 3.8 the results obtained with the proposed method and with three other algorithms designed for separation of solo instruments (or specifically singing voice) from music accompaniment are presented for reference. The algorithm Marxer2 [121] is an NMF-based approach which extends the work of [25] to specifically address the problem of singing voice extraction and fricative modeling in the separation scheme. The REPET algorithm presented in [29] takes advantage of the repetitive structure of most commercial songs to separate singing voice from music accompaniment in single-channel mixtures. Additionally, results from the Marxer1 algorithm are also presented for reference. As opposed to the other approaches presented in the table, the Marxer1 algorithm works in stereo mixtures and uses panning information to perform separation. Even when a fair direct comparison cannot be done with this algorithm, results are presented to give the reader a general overview of the state-of-the-art performance in solo/accompaniment separation. The algorithm Marxer1 is a low-latency main instrument separation approach for **stereo mixtures** presented in [48]. The method uses a probabilistic pitch extraction approach in conjunction with panning masks to perform separation.

The full table of results for all the algorithms submitted can be found at the campaign's website [122]. As in the 2011 campaign, the dataset used in 2013 was entirely composed of commercial vocal tracks and consequently, only results for voice extraction on the test dataset are presented in the table.

Table 3.8.: Results from the SiSEC13 Evaluation Campaign for vocal extraction on the test dataset.

	Cano	**Marxer2**	**REPET**	**Marxer1**
OPS	19.5	20	22.8	22.0
TPS	5.0	18.2	54.6	49.3
IPS	62.0	64.1	35.7	49.5
APS	8.7	16.5	49.4	29.3

The proposed method obtained comparable OPS scores to the other approaches, exhibiting particularly high IPS scores at the expense of lower APS and TPS scores. The proposed method has a processing time of 0.25 sec for 1 second of audio on a 2.6 GHz computer, allowing real-time processing. The algorithm Marxer2 has a performance time of approximately 3 times the length of the audio segment on a 3.2GHz. This means that the proposed method is approximately (without taking computer speed into consideration) 12 times faster than Marxer2. The authors report that algorithm Marxer1 allows real-time processing on a 3.2 GHz computer.

The good balance between performance and efficiency of the REPET algorithm is particularly interesting. The authors report processing times of 0.04 sec for 1 second of audio on a 3.4 GHz computer. However, it is to be noted that the assumptions behind the REPET algorithm are particularly strong as the algorithm assumes that a repetitive structure will always be present in the accompaniment. This assumption holds true for most commercial music; however, the veracity of the assumption for other types of music is still questionable. Further tests with the REPET algorithm are presented in Section 3.3.9.3.

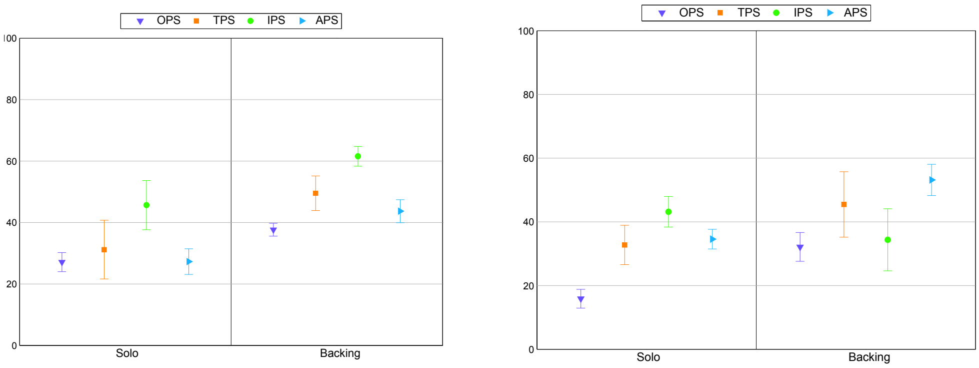
Results show that the proposed method is an efficient solution for solo/accompaniment separation, obtaining comparable OPS to other state-of-the-art algorithms without making any strong assumptions about the signals to be separated.

3.3.9.3. Other comparisons

Given that the REPET algorithm [29] obtained good results in the SiSEC 2013 campaign (see Section 3.3.9.2), the algorithm is used in this section for performance comparison with the proposed tone-based separation algorithm (including all the processing stages described in Section 3.3). The available Matlab implementation of the REPET algorithm [123] was used for the analysis. The instrumental dataset in **SA_DS1** (see Appendix B for a description of the dataset) was processed with the REPET algorithm. This dataset was chosen as it shows a great diversity of musical genres and instrumentations. Results obtained for the REPET algorithm are shown in Figure 3.45b. Results with the proposed algorithm with the instrumental dataset are displayed again in Figure 3.45a for reference and easy comparison.

The proposed tone-based separation algorithm clearly outperforms the reference method both in terms of Overall Perceptual Score (OPS) and Interference-related Perceptual Score (IPS) for the solo and backing tracks. Additionally, the proposed method obtains slightly higher Target-related Perceptual Score (TPS) (49.63) for the backing than the reference algorithm (45.46). However, the reference method obtains slightly higher Artifacts-related Perceptual Score (APS) than the proposed method for both the solo and backing.

As described in Section 3.3.9.2, both the proposed algorithm and the REPET algorithm allow real-time processing. Results presented in this section show that for the



(a) Proposed

(b) Reference

Figure 3.45.: Results obtained with the proposed tone-based separation algorithm (proposed) and the REPET algorithm (reference) for the instrumental dataset. (a) Proposed (b) Reference. Overall Perceptual Score (OPS), Target-Related Perceptual Score (TPS), Interference-Related Perceptual Score (IPS), Artifact-Related Perceptual Score (APS). Mean values with 95% confidence intervals are presented. It can be observed that the proposed method outperforms the reference method in terms of OPS and IPS both for the solo and backing tracks

instrumental dataset evaluated, the proposed method outperforms REPET in 5 of the 8 perceptual measures.

4. Sound Source Separation in Music Education

One application that can directly benefit from sound source separation research is music education. The possibility to extract a desired sound source (or group of sources) and use them as practice material, is very powerful. Different use-cases of separation technologies in music education can be depicted : (1) Music teachers in schools and music academies often use audio recordings and YouTube videos when teaching a new piece of music to a group of students [124]. Sound separation would allow the creation of accompaniment tracks for the students to play along with. (2) Amateur and self-taught musicians can benefit from separation technologies by creating accompaniment tracks to use in practice time and performances, and by using solo tracks for reference when learning a new musical piece. (3) Semi-professional musicians and college music students, specially in the classical realm, often encounter the difficulty of limited rehearsal time with accompanying ensembles when preparing for solo concerts. It is often the case that only one or two general rehearsals are scheduled before a given concert. The possibility of creating backing tracks from existing recordings of a given piece would allow musicians to make themselves familiar with accompanying parts, and to have unlimited practice time with a reference accompaniment before the real rehearsal.

Even when the possibilities and potential are immense, the usage of music technologies in music education is an ongoing process: on the one hand, it completely relies on the accomplishments of the scientific community where robustness and performance of many methods still requires a lot of work; on the other hand, it is a process that requires a progressive change of mentality in a community where many processes and techniques still remain very traditional. The use of Music Information Retrieval (MIR) technologies in the development of music education systems faces many challenges: (1) Development of music technologies robust and efficient enough to be delivered to the final user. (2) Bridging the gap between two communities—music education and music technology—that have completely different ways of working and mentalities. (3) Design of appealing and entertaining systems capable of creating interest while developing real musical skills.

This chapter gives a brief overview of the use of MIR and separation technologies in music education. Additionally, a comprehensive list of music education applications and tools that use music technologies is presented in Appendix A for reference. Furthermore, to place this research in the context of music education, the usability of sound separation technologies was evaluated through a listening test procedure. The listening test was developed with the goal of understanding the quality requirements posed by music education to separation technologies. Finally, as an example of a commercial application that includes solo and accompaniment separation as a feature, the Songs2See application is described. Songs2See uses the frame-based separation method presented in this thesis for solo and accompaniment separation.

4.1. Music Technologies in Music Education: A Historical Overview

The rapid development of music technology in the past decades has dramatically changed the way people interact with music today. It was only a natural consequence that the potential of developing more advanced tools for music education was also recognized. An automatic system that could give instructions, performance feedback, and guide music students through practice sessions, could become a very powerful teaching and learning tool. However, between the 1980s and the early 2000s, automatic methods for pitch detection, music transcription, and sound separation among others, were still in very preliminary stages. Consequently, initial systems for music education, even though innovative and creative, had many restrictions and mainly relied on the possibilities offered by recording studios.

Starting in the 1980s, play-along CDs became popular as an alternative way to practice an instrument. Play-along CDs consist of specially recorded versions of popular musical pieces, where the user plays along to the recorded accompaniment. The main advantages of play-alongs are that users can directly practice with their musical instrument, time for practice and rehearsal is unlimited, and getting familiar with accompaniment parts becomes much easier. The downside of these types of practice content is that the amount of available practice material is limited by the high production costs of recording sessions: In many cases, large ensembles and long recording sessions are needed for the production of one track. Consequently, play-alongs are mainly available for very popular songs and for some representative concerts of the instrumental repertoire.

Additionally, instructional videos came out as an educational tool where renowned musicians addressed particular topics—playing techniques, improvisation, warm-up exercises—and offered hints and instructions to help users improve their skills. With time, the catalog of instructional videos grew both in size and diversity, featuring not only famous musicians, but also different playing techniques, learning methods, and the very famous self-teaching videos.

The main weakness of both play-alongs and instructional videos is that there is no direct feedback for the user in terms of performance evaluation. Users completely rely on their own perception and assessment, which in case of beginners, can be challenging. However, these types of learning material have played a very important role as they offer an alternative way to practice at home, helping to keep motivation for learning, and offering the flexibility of practicing on your own time, pace, and schedule.

Later on, the video game community introduced rhythm games to the market with titles such as the well-known Guitar Hero series. In rhythm games, the user is required to repeat patterns of fingering gestures on special hardware controllers while the audio track plays on the background. The use of game controllers that resemble real musical instruments greatly simplified the signal processing involved in capturing user renditions. Entertainment, more than music education, was clearly the target of these games. However, they are still considered in this section as they were and still are very successful in creating interest in music performance and thus, play an educational role.

From a music education point of view, video games of these characteristics have two main weaknesses: (1) They often fail to develop musical skills that can be directly transfered to real musical instruments [125]. This is due to the fact that game controllers cannot capture the complexities and intricacies of musical instruments. (2) Content is entirely limited to a set of songs delivered with each game or to titles offered by publishers for download.

As the importance of having more engaging means for education was recognized, interactive applications and web services that offered a more formal approach to music education were also developed. Different topics such as music history, musical instruments, and band practice have been covered by applications of this kind. Some of these systems are used in music schools and universities as part of their class work; others are targeted for home users that wish to approach music in a more individual way.

A very strong trend in recent years propelled by the high processing power of portable devices, is the development of platform specific apps. Music learning has not been the exception and many music-related applications are now available for the Android and iOS markets.

All the above mentioned types of applications have benefited by the developments of the MIR community. Applications of different kinds that already include music content analysis have emerged in the last couple of years. The reader is referred to Appendix A of this thesis for more detailed information about music education applications that use MIR technologies. In the Appendix, a comprehensive list of play-alongs, instructional videos, music video games, music education software, online systems, mobile applications, and research projects is presented.

4.2. Sound Separation and Music Education

As already mentioned in Section 2.4 (where quality assessment in sound separation is introduced), different applications pose different quality requirements in terms of allowed artifacts, interference from other sources, or target related distortions. A separation algorithm used as an intermediate step in an Automatic Music Transcription (AMT) system would probably have strict requirements in terms of the allowed interference from other sources. Minimizing interference will allow a more accurate extraction of melodic and rhythmic information. However, the overall perceptual quality of the separated track falls into second place as the separated tracks are not meant to be heard by the final user. Similarly, if the goal is to remix a given recording, target-related distortions become more important and audible artifacts should be kept to a minimum.

In the case of music education, sound separation is in itself, the final goal of processing. A given audio mixture is separated and the resulting tracks are used directly during practice time. Ideally, all the individual quality metrics would be optimized to guarantee separated tracks suitable for the music education context, and for that matter, for any other context. However, reality shows that the state-of-the-art is far from reaching that optimal point. This is however no reason to discard the possibility of using separation results for music education. The questions that this study addressed is:

(1) Which are the quality requirements expected from separation algorithms for them to be suited for Music Education and practice applications?.

To address this question, a listening test procedure that puts music practice and separation research together, was designed and conducted. The characteristics of this test are presented in the following section.

4.2.1. Listening Test Procedures

Laboratory Set-up

All the tests described in this section were conducted using the same laboratory set-up. The music practice room at Fraunhofer IDMT was used to conduct the listening tests. Subjects were asked to play short pieces of music featuring their musical instrument. To minimize the impact of musical knowledge in the results of the listening tests, users were given several performance options to choose from. Participants could choose between using a traditional printed musical score, a scrolling piano-roll view, and a tablature representation for guitar and bass. For the scrolling piano-roll and the tablature representation, the Songs2see Game interface was used [O6]. The Songs2See interface is shown in Figures 4.1a-c, and is briefly described in Section 4.3. With this, the goal was to resemble as closely as possible a real practice session for the participants. All the audio material was played through a pair of AKG K701 semi-open headphones and the subjects were allowed to modify both the playback level of the tracks and of their instruments to their own personal taste. The choice of headphones over speakers was made based on the fact that in real practice scenarios, commonly used playback devices are portable audio players, tablets, and cell phones in combination with a pair of headphones.

4.2.1.1. Listening Test 1: Is sound separation a desired functionality in music education?

This research has always followed the lead of the music education community when it comes to the relevance of developing algorithms capable of separating audio recordings into its solo and accompaniment components. Given that a great number of music education methods and literature have been published where the use of solo and accompaniment tracks is proposed for practice (as was described in Section 4.1 and published in [O7]), the assumption that sound separation is a desired functionality for music education and practice was always made. However, to verify this initial assumption, a simple listening test was conducted to assess users' preferences when both separated tracks and mixtures are available within a practice environment. In the listening test, subjects were asked to compare playing to the original mix, with playing to the original solo and accompaniment tracks obtained from the multi-track recordings. This is of course a hypothetical scenario that tries to asses the usability of solo and accompaniment tracks in Music Education applications given that very high quality separation can be achieved. In the listening test, the solo and backing track were not evaluated independently; instead, subjects were allowed to play with either the solo, the backing, or with a mixture of the two. The mixer options within the Songs2See Game were used to allow subjects to get the desired balance between the tracks. Having the option to freely mix the solo and backing tracks is a functionality that is available to the users when sound separation is performed. In Figure 4.1c, the Songs2See game interface with its mixer options are displayed. By moving the sliders for the solo and backing playback levels, the desired balance can be obtained.

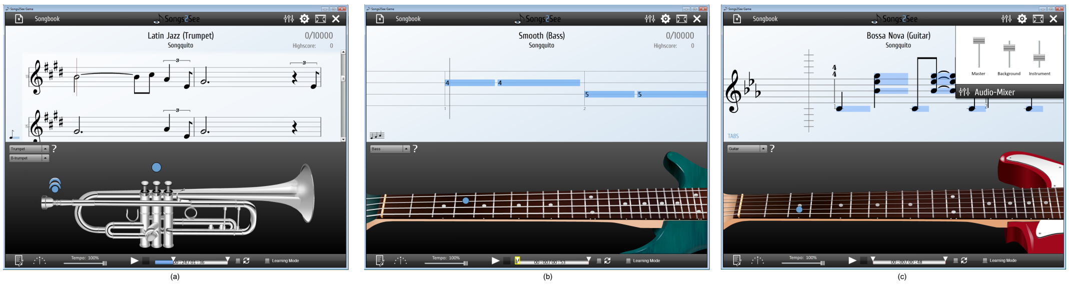


Figure 4.1.: Songs2See Game interface. (a) Score sheet view (b) Tablature view (c) Scrolling piano-roll view and mixer menu. The mixer menu can be used to modify the solo/backing track playback volume. The different visualization options were offered to the subjects of the listening test for convenience.

A total of 10 subjects conducted the listening test. The subjects were all beginner to advanced musicians between 27 and 34 years old: 3 bass players, 1 trumpet player, 3 guitar players, 2 piano players, and 1 saxophonist. All the subjects were asked to play a piece of music (previously unknown to them) featuring their musical instrument. Each subject was given 60 minutes of practice time where they could freely use the mix or the separated tracks to assess their preference. Subjects were asked to rate how comfortable they felt playing with the mix and how comfortable they felt playing with the separated tracks. All ratings in the listening test were performed in a continuous scale from 0 to 100 where additional descriptive hints were given: *Bad* [0-20], *Poor* [20-40], *Fair* [40-60], *Good* [60-80] and *Excellent* [80-100] [70].

Listening Test Results

Results from the listening test are displayed in Figure 4.2. Following the recommendations in [70], mean values with 95% confidence intervals are presented. In the figure, results obtained with the original mix are shown as *Mixture*, while the results obtained with the solo and accompaniment (or a mix of the two), are shown as *Separated*. Results clearly show that subjects prefer having the possibility of playing with separated tracks than with the original mix. Performing separation allows users to control the playback levels depending on their skills and preferences: completely mute the solo, add a little of the solo to the backing track for reference, or combine the tracks again to get the original mix. The separated tracks obtained a mean value of 95.22 and a confidence interval of only 3.3. The mixture obtained a mean value of 62.88 with a slightly large confidence interval of 14.87. This listening test confirms the somehow expected results that the use of sound separation in music education applications brings beneficial functionalities for the practice sessions.

4.2.1.2. Listening Test 2: What are the quality requirements for sound separation to be suitable for music education?

After having verified the initial assumption that sound separation is a desired functionality for music education, this listening test addressed the question of quality expectations for separation results within a music education context: Which types of signal distortions are acceptable in a music practice scenario? Are quality requirements the same for the solo and accompaniment? How can separation algorithms be optimized to be suitable for music education?

A group of participants were asked to play along to different versions of solo and accompaniment tracks specifically created to resemble common distortions in separation algorithms: artifact distortions, interference distortions, and target distortions. All the participants played with their own musical instruments a piece unknown to

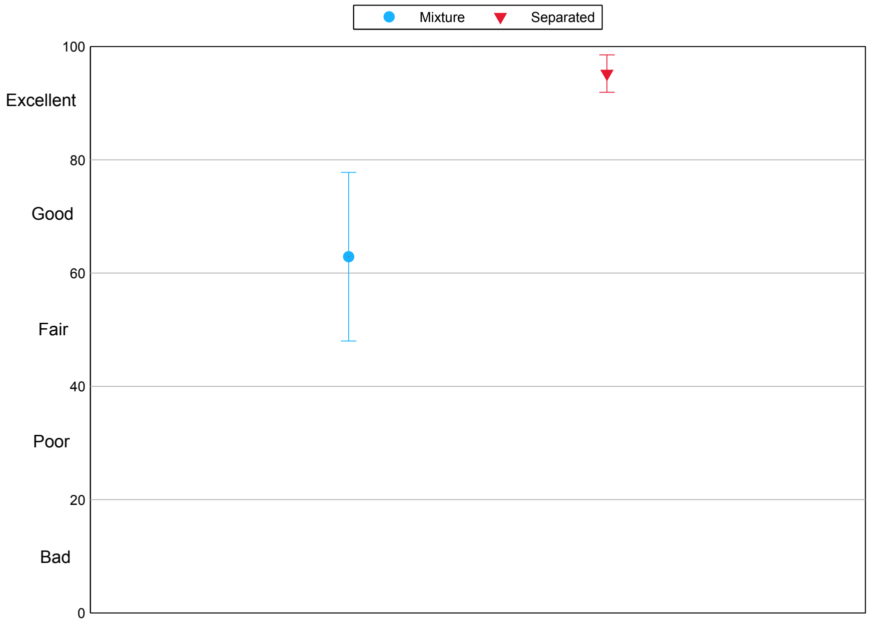


Figure 4.2.: Listening test results. Mean values with 95% confidence intervals are shown for the original mix and the separated tracks. It can be clearly seen that subjects prefer to play with separated tracks than with the original mix.

them before the test. Participants were asked to rate how comfortable they felt playing with each of the tracks and the degree to which each of the tracks contributed to making it easier to play the newly presented musical piece.

A total of 12 subjects conducted the listening test which again took place in the music practice room at Fraunhofer IDMT. The subjects were all intermediate to advanced musicians from 15 to 34 years old: 4 guitar players, 3 bass players, 3 piano players, 1 trumpet player, and 1 saxophonist. For each instrument, commercial multi-track recordings were used to create the test material. Due to copyright restrictions, this dataset cannot be made publicly available.

The listening test consisted of a *Training Phase* and an *Evaluation Phase*. During the training phase, the subjects were given a short explanation of the listening test procedure, its goals, and the evaluation procedure. The subjects were also presented with test material so they could make themselves familiar with the types of signals and distortions in the evaluation.

The evaluation stage was divided in two sections: (1) Solo Track Evaluation, and (2) Backing Track Evaluation. In the two evaluation sections, four versions of solo and backing tracks, as well as the original recording (mix) were used. This gives a total of five versions of each signal that the subjects were asked to compare. Three of the signal versions (v1, v2, v3) were created so that each one specifically described one of the signal distortions (interference, artifacts, target). The fourth version (v4) was obtained with the separation algorithm presented in Section 3.3. The original recording (mix) was always used as a comparison as in most music practice scenarios, this track is the only one available to the users. To create v1, v2 and v3 the approach proposed in [71] was used:

1. Artifacts Signal: This version was obtained as the sum of the original target signal and an artifacts signal. The artifacts signal was created by randomly taking 1 % of the time-frequency coefficients of the target source (and thus setting 99 % of the time-frequency coefficients to zero) and synthesizing this very sparse signal. The loudness of the artifacts signal was adjusted to that of the target. This artifacts signal is then added to the original target signal (clean) to artificially create a signal with artifact distortions. Randomly taking 1 % of the time-frequency coefficients results in a very sparse time-frequency representation that sounds like clicks, breaks, and musical noise when re-synthesized.

2. Target Signal: This version was created by low-pass filtering the original source signal to a 3.5 kHz cut-off frequency and by randomly setting 20 % of the time-frequency coefficients to zero.

3. Interference Signal: This version was obtained as the sum of the original source signal and an interference signal. In the case of solo and accompaniment

separation, the interference signal for the solo is always the accompaniment and the interference for the accompaniment is the solo signal. The loudness of the interference signal was adjusted to that of the target.

The order in which the signals were presented to each user in the two sections of the test was randomized. The subjects were asked to rate how comfortable they felt practicing the musical piece with each of the different signal versions. In the Solo Track Evaluation section, the subjects were asked to practice the musical piece and play it as fluidly as possible with the aid of the solo track versions. In the Backing Track Evaluation section, the users were asked to play the given melody with the accompaniment of the backing tracks. Subjects were then asked to evaluate the provided tracks. All ratings in the listening test were performed in a continuous scale from 0 to 100 where additional descriptive hints were given: *Bad* [0-20], *Poor* [20-40], *Fair* [40-60], *Good* [60-80] and *Excellent* [80-100] [70]. Users were also allowed to submit any comments that they found relevant about their experience in the listening test.

Listening Test Results

Following [70], the results from the listening test presented in Figure 4.3 show mean values and 95% confidence intervals for each of the four signal version and the mix. Results for the solo signals are presented in the left pane of the figure and results for the backing on the right.

In the Solo Track Evaluation the highest score was obtained by the *Interference* signal closely followed by the mix. Subjects found the *Artifacts* signal the most disturbing type of distortion during practice time. This evidences the importance of preserving the signal's quality with a minimum of introduced artifacts, regardless of the fact that traces of the other source are still present. Subjects comments after the listening test emphasized the fact that artifacts are distracting and make it more difficult to keep rhythm. The lowest rating for the solo signals was obtained by the proposed algorithm with fairly large variance between subjects. Two possible explanations for these results can be envisioned. On the one hand, it is possible that due to the special characteristics of each instrument's sound and the types of melody lines that each instrument usually plays, different instruments pose different quality requirements. A trumpet for example, an instrument with a powerful sound and distinctive timbre, might pose different requirements than a bass guitar which has a less distinctive timbre and a less powerful sound. It is plausible that signal distortions are perceptually more disturbing for the latter. On the other hand, the fact that the proposed separation algorithm handles different instruments differently might have an important effect on the ratings. If the quality of the extracted solo is lower for certain instruments than for others, it is to be expected that this is evidenced by the ratings. Further experiments need to be conducted to better understand the variance of the ratings.

In the Backing Track Evaluation the highest score was also obtained by the *Interference* signal, the mix also following very closely. In this case, the *Target* signal obtained the lowest overall rating, making it the most disturbing of the all signal distortions. The importance of a clear bass to follow was mentioned by the users and due to the somehow smoothed (low-pass filtered) *Target* versions, onsets and bass notes were no longer so clear. The backing signals obtained with the proposed algorithm obtained in this case, scores superior to those of the *Target* and *Artifacts* signals.

The high mean values obtained for the reference mix might be due to the fact that the reference mix is the only version that entirely preserves the quality of the signal. Furthermore, a familiarity factor might play a role here as in most cases, the mix is the only version available to the users and they might feel most familiar with it. However, results from the Listening Test 1 where subjects clearly showed their preference to play with separated tracks than with the mix, support the fact that high mean values obtained for the mix in the second listening test are greatly due to the fact that signal quality is entirely preserved. In a hypothetical scenario where quality of separated tracks reaches its maximum (as in Listening Test 1), user preference is clearly towards separated tracks. This naturally poses a challenge for separation research as the performance boundary that separates users' preference for the mix from users' preference of separated tracks has not entirely been reached.

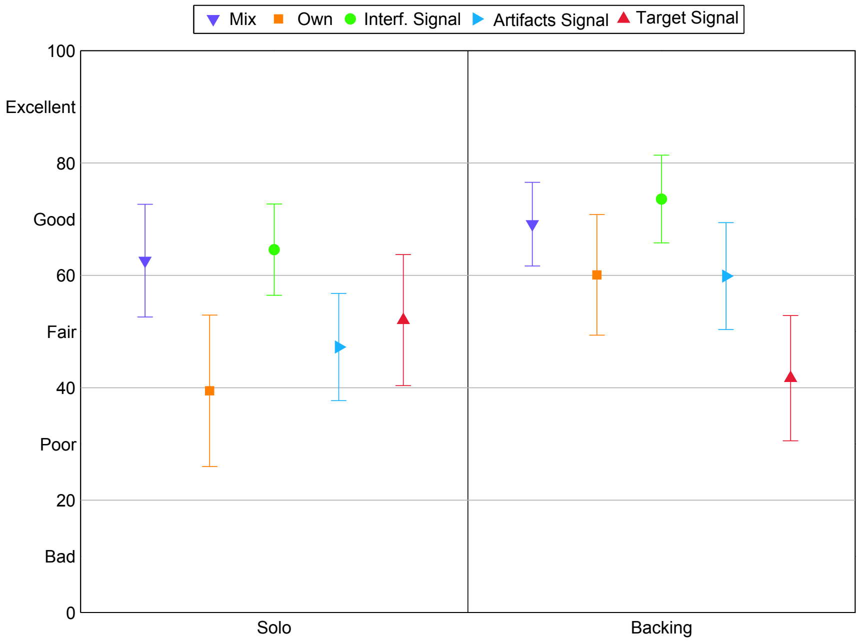


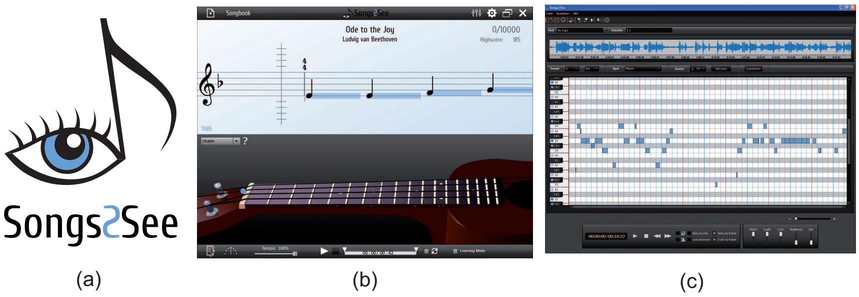
Figure 4.3.: Listening Tests Results: (left) Solo tracks, (right) Backing tracks. Mean values with 95% confidence intervals are shown. Ratings for the original recording (mix), signals obtained with the proposed algorithm (own), Interference Signal (Interference), Artifacts Signal (Artifacts), and Target Signal (Target) are presented. It can be observed that from the distortion signals, the lowest rating was obtained by the artifacts signal in the solo evaluation and by the target signal in the backing evaluation.

4.3. Songs2See: Music Learning Game

This section briefly describes a commercial application for music practice that incorporates solo and accompaniment separation as one of its features. The release of this application and the inclusion of a separation algorithm in it, sets an important precedent for the potential and usability of separation algorithms in music education applications.

Songs2See is an application developed at the Fraunhofer Institute for Digital Media Technology IDMT with the goal of placing music practice in a gaming environment.

Songs2See offers, among many other features, the possibility to use real musical instruments and receive real-time performance feedback. Songs2See is composed of two different applications depicted in Figure 4.4: the *Songs2See Game* used at practice time, and the *Songs2See Editor* used to create content for the game.



(a) (b) (c)

Figure 4.4.: Song2See application. The Songs2See application uses the frame-based separation algorithm proposed in this thesis for practice content creation. (a) Songs2See logo (b) Songs2See Game Interface, (c) Songs2See Editor Interface.

Song2See currently supports eight musical instruments: saxophone, trumpet, clarinet, ukulele, guitar, piano, and bass. For all the musical instruments, automatic fingering animations are displayed to guide users through the performance. Furthermore, users' performance is rated through a real-time pitch detection functionality. As shown in Figure 4.1, Songs2See offers different visualizations of the main melody: scrolling score, tablature for guitar, bass, and ukulele, and traditional music score. A feature for tempo reduction is also available in the game, allowing users to slow down the track to a tempo where they feel comfortable.

One important feature in Songs2See is the possibility to create personal content using the Songs2See Editor. The main idea is to give users flexibility to play any piece of their taste and provide the corresponding processing tools that allow the extraction of all the necessary performance information. Besides performing the transcription process for the main melody, Songs2See also includes the possibility to separate the solo instrument from the musical accompaniment. Users can play to the created backing tracks which resemble real performance environments, or use the solo tracks for performance analysis and reference. A mixing control, as displayed in

Figure 4.4c, allows users to adjust the playback level of the solo and accompaniment tracks to their personal taste and needs. Also within the Songs2See Editor, features for automatic tempo extraction and key analysis are available.

Songs2See was released to the market on March 2012.

4.3.1. Sound Separation in Songs2See

The main goal of including a solo and accompaniment separation algorithm within the Songs2See application was to give users flexibility to create practice content for the game from any musical recording of their taste. For this matter, two possibilities of performing solo and accompaniment separation are included in Songs2See: (1) fully automatic separation. This feature uses the separation algorithm presented in Section 3.1. (2) User-assisted separation.

As the details of the automatic separation algorithm have already been explained in Section 3.1, only the details of the user-assisted separation as described in publications [O5, O7], are described in this section.

Bearing in mind that both pitch detection and separation results are highly dependent on the complexity of the audio mixture, a user-assisted version of the algorithm was included such that the user could correct and guide the application in the separation process. After an initial pitch detection stage, the sequences of $f0$ values are transformed into note objects, in a similar approach as the one used in Section 3.3.3. A specially designed user interface was included to display the initial estimate of the main melody in a piano-roll representation. In Figure 4.5 an example piece is displayed in the Songs2See piano-roll representation.

Each one of the tone objects can be modified both in pitch and length. Furthermore, new tone objects can be included and erroneous ones can be deleted. These flexible processing options allow the user to obtain a very accurate representation of the main instrument in the track. With the refined pitch estimate, the separation algorithm is run again (by-passing the pitch detection stage), and new solo and accompaniment tracks are calculated.

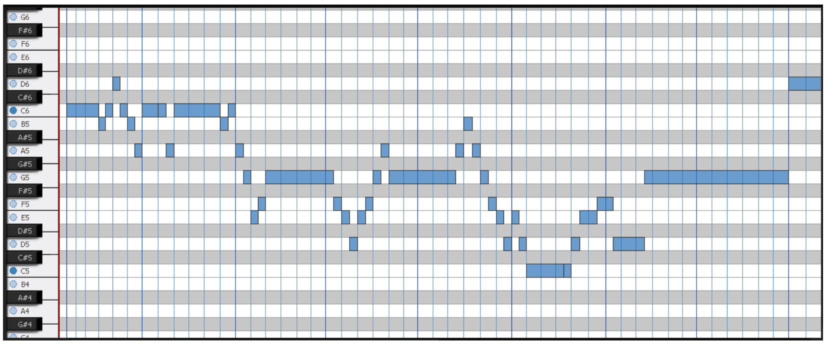


Figure 4.5.: Songs2See Editor piano-roll view: The $f0$ sequences delivered by the pitch detection stage are converted into tone objects and displayed to the user as blue bars covering the length of the tone. Pitch is given by the piano keyboard on the left side of the figure.

Two additional options for user-assisted separation are also included in Songs2See: (1) Audio + MusicXML import, and (2) Audio + MIDI import. These two options allow the user to import not only the audio track to be separated, but a MIDI or a MusicXML file containing the transcription of the melody of the main instrument. These options were motivated by the great number of MIDI files available online and by the widespread use of music notation software such as Finale and Sibelius that support the MusicXML format. The separation process in these cases is equivalent to a score-informed separation approach where the user is directly in charge of aligning the melody representation with the given audio track.

4.3.2. Conclusions

This chapter presented a general overview of sound separation in music education applications. After a historical overview of the use of MIR technologies in music education, results from two listening tests were presented where quality requirements of separation results in a music education context were evaluated. Lastly, an example of a commercial application for music education where the separation algorithm proposed in this thesis is featured, was described. Several conclusions can be drawn from the studies presented in this chapter:

- The listening tests have confirmed the initial assumption that solo and accompaniment separation is a desired functionality in music education and practice applications. However, results also showed that separation quality still needs to be improved for users to prefer the use of separated tracks over the original mix.
- The second listening test showed that solo and accompaniment tracks should be addressed differently in terms of quality, with artifact distortions being most relevant for solo signals, and target distortions most relevant for the backing tracks.
- Songs2See, as well as applications like Riffstation (See Appendix A), have set a precedent for commercial application for music education that offer sound separation as a feature.

5. Concluding Remarks

This thesis addressed the development of a system for pitch-informed solo and accompaniment separation capable of separating main instruments from music accompaniment, regardless of the type of solo instrument used, musical genre of the track, or type of music accompaniment. For the solo instrument, only pitched monophonic instruments were considered in a single-channel scenario where no panning or spatial location information is available. The algorithm was kept lightweight and processing times were minimized allowing its use in real-world applications.

The different stages of the separation process—prior information, parameter extraction, and the final separation procedure—were studied independently to better assess their contributions to separation quality, their main weaknesses and strengths, and possible ways to improve their performance.

For the prior information stage, two methods for main melody extraction were studied and compared in the separation context. Even though each algorithm proved to benefit separation in different ways, experimental results suggest that more accurate pitch detection will not bring considerable quality improvements unless more complex modeling strategies for the target sources are used. The use of ground-truth pitch information in pitch-informed separation algorithms has made clear that the performance boundary of current methods will not reach maximal quality scores. The main limiting factor of separation algorithms is parameter estimation which has shown to restrict separation quality more than accuracy of pitch as prior information. Similar results have been presented in [51], [65], and [O8].

In the parameter estimation stage, a tone-based separation approach was proposed to address the solo/accompaniment separation problem. Parameter estimation is based on a sinusoidal modeling approach where the harmonic components of each tone are estimated using known characteristics of musical instrument tones. The proposed approach is entirely based on the understanding of the temporal and spectral characteristics of musical instruments and has proven to be both an efficient and flexible strategy.

To improve parameter estimation, the spectral parameters of musical instrument tones were studied both to understand their importance in the perceived quality of separated tracks, and to better model them within the separation scheme. General tendencies for the magnitude variations of spectral envelopes, inharmonicity, as well as phase expectation and coupling were studied and applied in the context of sound separation. For the instruments studied, magnitude envelopes showed mean frame-wise percentage changes of approx. ±5 %. Slightly larger magnitude variations were observed for musical instruments such as the piano and guitar whose notes do not have a sustained part and only decay in time after the string has been plucked. Inharmonicity varies slightly between musical instruments; however, an average percent frequency from harmonic location of ± 1% was observed for the first 9 harmonics of the instruments considered. The study of the Instantaneous Frequency Distribution (IFD) of the different partials of a tone showed that for most instruments, phase contours of the different partials are highly correlated and common micro-modulations can be observed among the partials. A clear exception to this observation was the piano where no correlation was observed between the IFDs of the different partials. Phase coupling was also studied as a mean of estimating phase of a given partial using information from other partials of the tone. Phase estimation errors of approx. ±0.3 radians were observed, with errors increasing with increasing partial index. A preliminary study on the perceptual quality of instrumental sounds with reconstructed phase based on coupling was conducted for the clarinet. Results suggest that good perceptual audio quality can be achieved under certain conditions; however, the approach also proved to be very sensitive to minor estimations errors. Reconstructed audio for the clarinet dataset obtained a mean PEAQ score of -1.18 which stands for a *"slightly annoying"* degradation of quality. Finally, phase expectation was studied and a novel approach for harmonic/percussive separation was proposed. The proposed harmonic/percussive separation method outperformed the reference state-of-the-art algorithm [115] obtaining a mean Overall Perceptual Score (OPS) of 32.93 for percussive extraction, with the reference algorithm obtaining a mean OPS of 23.11.

Within the separation scheme, the proposed tone-based processing allowed the inclusion of novel processing stages to better estimate attack sections of the tones, to reduce transient interferences in the solo signal, and to apply the concept of Common Amplitude Modulation (CAM) to reduce interferences from other harmonic sources in the solo estimation. Additionally, the concept of phase expectation was used within the separation scheme to capture noise-like characteristics of musical tones such as fricative sounds in the singing voice. The tone-based solo and accompaniment separation method proposed achieved comparable performance to state-of-the-art algorithms under the Signal Separation Evaluation Campaign (SiSEC) 2013. It also outperformed the reference state-of-the-art algorithm [29] with the instrumental dataset obtaining a mean OPS for the solo of 27.12, with the reference algorithm obtaining a mean OPS of 15.86.

Additionally, solo and accompaniment separation was evaluated in the Music Education context. Two listening tests were conducted with the goal of understanding ways of optimizing separation algorithms for music education. An initial listening test clearly proved that instrumental practice can benefit from solo and accompaniment separation, giving users flexibility and better tools to learn a new musical piece. The second listening test showed that different quality requirements are posed for the solo and for the accompaniment tracks in a practice scenario. While artifact distortions are most relevant for the solo signal, target distortions proved to be very important for the accompaniment tracks. The use of solo and accompaniment separation in the Songs2See application was also described. The frame-based separation algorithm proposed in this thesis is used in the application to create practice content to be used in the Songs2See Game. Algorithm robustness and efficiency were critical factors in the inclusion of the separation method in the Songs2See application. Songs2See was awarded the 2012 Prize for Innovation and Entrepreneurship by the German Informatics Society.

5.1. Outlook

In this thesis, several processing alternatives have been proposed to address specific challenges in separation research and ultimately, to improve separation quality. However, many questions still remain unanswered and many more have emerged during the completion of this thesis.

Probably one of the most relevant topics that needs to be addressed in sound separation research is perceptual quality evaluation of separation results. Even though great efforts have been placed into the development of objective perceptual quality measures specifically designed for the separation context, the weaknesses of current evaluation scores such as the ones proposed in the Perceptual Evaluation Methods for Audio Source Separation (PEASS) Toolkit have been long recognized by the separation community. Even when the measures have proven to be rough indicators of separation quality, they often fail to numerically evidence quality differences clearly audible and easily recognized by non-expert human listeners. This naturally represents a major weakness as a thorough and truthful evaluation of algorithm enhancements is often difficult without conducting a formal listening test. As part of the calculations returned by the PEASS Toolkit, audio signals with the distortion errors calculated by the toolkit are available. Each audio signal contains those distortions found in the estimated target source that are classified either as target, interference or artifacts distortions. Informal listening of these signals clearly show that the different types of distortions are not properly discerned, and resulting errors are very often a clear mix between different types of distortions. This evidences the shortcomings of the toolkit to properly categorize signal distortions.

Also relevant to quality evaluation of separation results is the understanding of the separation quality requirements posed by different applications. Ideally, separation algorithms would be powerful enough to obtain high quality scores for all the measures; however, given the current state-of-the-art, it is clear that separation quality is still far from maximal ratings and in most cases, separation approaches benefit certain measures over others. Additionally, taking into consideration that in many cases sound separation is not the final goal but an intermediate step for more advanced content analysis tasks—automatic music transcription, instrument recognition, re-mixing, tempo and beat extraction—, clearly understanding quality requirements of different applications is of critical importance.

Separation results have also shown that in many cases, separation strategies that are beneficial for certain target sources are not necessarily beneficial for separation of other sources in the audio mix. In these cases, it is often very difficult to come up with a separation method that can balance the quality of all the extracted sources under an unified processing approach. To date, most separation algorithms attempt to extract audio sources that exactly reconstruct the original audio mix; however, only very specific applications actually restrict separation algorithms to exactly reconstruct the audio mix from the separated tracks. It can be preferable in some cases to relax the Perfect Reconstruction constraint [O8] for the sake of improved perceptual quality of the separated tracks. Relaxing this constraint would allow different sources to be estimated with different algorithms that can better address their spectral characteristics. Additionally, introduced algorithm distortions would not propagate from one separation procedure to the consequent one and better separation quality can be achieved.

In the particular case of pitch-informed separation addressed in this thesis, it has become clear that having pitch detection and parameter estimation as independent processing entities does not allow the proper addressing of the estimation errors. The potential of combining separation techniques with other types of content retrieval methods has proven to improve robustness of results. In [126] for example, sound separation is used as a pre-processing step for beat tracking. In [O1] separation is used within an audio similarity context. In pitch-informed separation algorithms, a simple one-directional relationship has been the common approach used in the methods to date proposed. To improve results and make processing more robust under different signal characteristics, the use of a feedback loop between separation and pitch detection could bring beneficial results. The idea would be to perform an initial pitch-informed separation that would then return to pitch detection for refinement. Parameter estimation in separation approaches very often results in the extraction of valuable information that could be used for pitch detection refinement. In this thesis for example, a tone-based separation approach was proposed that estimates the harmonic structure of the solo instrument for every tone. It is an approach that brings Automatic Music Transcription (AMT) and sound separation a step closer to each other. Knowledge of musical instrument tones was also included in the estimation to model the spectral characteristics of the solo instrument. This information could be used for example, to discard erroneous tones detected by the pitch detection algorithm that do not fit the expected characteristics of an instrumental tone as defined by the proposed model. With this information, the pitch detection stage could be called again with the new constraints obtained at the separation stage. Several iterations between separation and pitch detection could be performed to further refine results.

A critical issue that also needs to be addressed in current separation research is the fact that most methods fail to include phase information directly in the estimation stage. Even when preliminary studies have been conducted and the importance of phase has been recognized, no robust solutions have been proposed so far that can take advantage of the valuable perceptual information contained in the phase of an audio signal. As opposed to spectral magnitude, phase information is extremely sensitive to modifications and minor changes can result in very audible distortions. This is most likely the reason why current approaches that include phase processing, are all very conservative in nature. However, results suggest that in many cases the performance boundary for current magnitude-based approaches is still far from reaching high quality ratings and the need of more complex estimation stages is critical if separation quality is to be improved.

A promising approach that could result in improved sound separation quality is the combination of sound separation techniques with sound synthesis technologies. These two fields of research have evolved independently over the years but the potential of combining them has scarcely been explored. One of the main difficulties faced by sound separation algorithms is the fact that sound sources inevitably overlap in the time-frequency domain. Different estimation techniques have been proposed in the attempt to resolve overlapping of spectral components and estimate the different sound sources. As a completely independent field of research, synthesis research has long worked on the refinement of instrument models that can more accurately capture musical instrument timbre and sound qualities. Synthesis models could be used to estimate sections of the signal whose estimation with standard separation methods is poor due to overlapping of spectral components. This would naturally mean that separation algorithms would have to be refined to specifically target a particular musical instrument. Nevertheless, the state-of-the-art in separation research has already recognized the need to more accurately model certain musical instruments such as percussion instruments or the voice, and separation methods to specifically address these instruments have been proposed in the literature. This concept would expand the idea of instrument-specific separation to any other musical instrument where a synthesis model can be used in conjunction with parameter extraction in the separation scheme.

Appendix A.

List of Music Education Applications (In alphabetical order)

A.1. Play Alongs and Instructional Videos

[A1] Alfred Music Publishing
http://www.alfred.com/Browse/Formats/DVD.aspx

[A2] Berklee Press
http://www.berkleepress.com/catalog/product-type-browse?product_type_id=10

[A3] Hal Leonard Corporation
http://www.halleonard.com/promo/promo.do?promotion=590001&subsiteid=1

[A4] Homespun
http://www.homespuntapes.com/home.html

[A5] Icons of Rock
http://www.livetojam.com/ltjstore/index.php5?app=ccp0&ns=splash

[A6] Jamey Aebersold
http://www.jazzbooks.com/jazz/category/AEBPLA

[A7] Music Minus One
http://www.musicminusone.com

A.2. Music Video Games

[A8] BandFuse: Rock Legends
`http://www.bandfuse.com`

BanFuse is scheduled for release for Xbox 360 and Sony PS3 in November 19, 2013. It offers the possibility to play real musical instruments —guitar, bass, and voice— and presents animated tablature and multi-player options. The game offers step-by-step guidance from famous rock legends.

[A9] Guitar Hero
`http://www.guitarhero.com`

Guitar Hero has been released for different video game consoles like Microsoft Xbox 360, Sony PS3, and also for Windows PCs.

[A10] Rock Band
`http://www.rockband.com`

Rock Band 3 has been released for Microsoft Xbox 360, Nintendo Wii, Sony PS3, and Nintendo DS. It supports up to three singers with a three-part harmony recognition feature. It was released with a set of 83 songs and has full compatibility with all Rock Band peripherals as well as most Guitar Hero instruments.

[A11] Rocksmith
`http://rocksmith.ubi.com/rocksmith/en-US/home/`

First commercial release in the video game community that allowed users to play with real guitars. It was released in the United States in September 2011 for Microsoft Xbox 360, Windows, and Sony PS3.

[A12] Singstar
`http://www.singstar.com`

SingStar was released for Sony PlayStation 2 & 3. It offers the possibility to sing along to the included songs with the lyrics shown synchronously.

A.3. Music Education Software & Online services

[A13] Garage Band
`http://www.apple.com/ilife/garageband/`

GarageBand is a software released by Apple for Mac and iPad. It provides the possibility to learn to play piano and guitar with specially designed content, performance feedback, and appealing user interfaces. Users can play directly to the computer microphone or through USB connection.

[A14] GuitarBots
`http://ovelin.com/guitarbots/`

Online service that allows to play with real guitars and offers different animated tablature with real-time performance feedback

[A15] Instinct
`http://instinct.com`

Web-based system to learn to play guitar. Users play to the computer microphone with real guitars. It offers an introductory course, and different lessons for chords, scales, rhythm, among others.

[A16] Music Delta
`http://www.musicdelta.com`

Music Delta is a web based system developed by Grieg Music Education comprising music curricula, content articles, and interactive tools.

[A17] Riffstation
`http://www.riffstation.com/`

Riffstation is a practice app for guitarists available for mac OSx an PC. If offers a tempo modification tool to slow down the audio track, pitch shifting to change the key of a song, and the possibility to mute the guitar in the mix.

[A18] Smart Music
`http://www.smartmusic.com`

SmartMusic is a Windows and Mac software developed by MakeMusic especially for bands, orchestras, and vocals. Users can play their instruments to the computer microphone and receive immediate feedback from the software. Teachers can assign tasks for the students to practice at home and track and rate their progress. Currently, there are around 2000 musical pieces available for the software.

[A19] Songle
`http://songle.jp/`

Web service for active music listening that uses Music Information Retrieval (MIR) technologies to extract melody, chord, beat, and structure information about the musical piece. The main idea behind this system is to allow users to have a deeper understanding of music and enrich their listening experience.

A.4. Mobile Apps

[A20] Cube Jam
`http://www.roland.com/amp/cubejam/`

Cube Jam is an iOS application developed by Roland as a companion for their series of guitar amplifiers. The app allows users to import songs from their iTunes library, play along to them, record their performances, tempo stretching, among others. It requires the use of a special cable to connect the mobile device to the i-Cube link jack on the amplifier.

[A21] Jamstar
`http://jamstar.co/`

Jamstar is an application to learn to play guitar available for iOs, Android, and as a browser application. It offers real-time feedback, a guitar tuner and a set of courses to address different topics from beginner to advanced level.

[A22] JoyTunes
`http://www.joytunes.com/`

JoyTunes offers different applications to learn piano and recorder. Their iPad application Piano Dust Buster allows the use of real pianos or keyboards as wells as touch input form the mobile device. It comes with a set of songs included and with a special game mode to learn to read sheet music.

[A23] Rock Prodigy
`http://www.rockprodigy.com/`

Rock Prodigy is a guitar playing app developed for the iPad, iPhone, and iPod Touch. Users can play their guitar directly to microphone and, based on a lesson plan and a rating system, receive performance feedback from the application.

[A24] Tonara
`http://tonara.com/`

Tonara is an interactive sheet music iPad app where users can download and view music directly on their iPad. The app records input directly from the microphone and automatically detects the user's position in the score.

[A25] Wild Chords
`http://www.wildchords.com/`

Wild Chords is a music game developed by Ovelin and designed to help beginners familiarize with guitar chords. It is available as an iPad app and uses appealing visuals and references to animals to help users identify the chords. The game records audio input directly from the microphone and no hardware controllers are needed.

A.5. Research Projects

[A26] i-maestro

http://www.i-maestro.org/

i-Maestro focused on the violin family and besides offering enhanced and collaborative practice tools, the project also addresses gesture and posture analysis based on audio visual systems and sensors attached to the performer's body.

[A27] IMUTUS

http://www.ilsp.gr/en/infoprojects/meta?view=project&task=show&id=120

IMUTUS focused on the recorder with the goal developing a practice environment where students could perform and get immediate feedback from their renditions.

[A28] KOPRA-M

http://www.idmt.fraunhofer.de/en/projects/Current_publicly_financed_research_projects/kopra_m.html

This is an ongoing project at the Fraunhofer Institute for Digital Media Technology that focuses on measurement of competencies in music. For this matter, a systematic methodology and a proprietary software solution to assign and control music tasks is developed.

[A29] Songs2See

http://www.songs2see.com

Songs2See dealt with the development of a music practice and learning tool that allow the use of real musical instruments and provided different options for content creation. It was conducted at the Fraunhofer Institute for Digital Media Technology.

[A30] VEMUS

http://www.tehne.ro/projects/vemus_virtual_music_school.html

VEMUS was proposed as a follow up project of IMUTUS and addressed the inclusion of further musical instruments and the development of tools for self-practicing, music teaching, and remote learning.

Appendix B.

Evaluation Datasets

Dataset	**Parts**	**Track N.**	**Name**	**Source**
SA_DS1	SA_DS1a Vocal	1	Dreams	Bass-dB
		2	Life as a disturbed infobeing	Bass-dB
		3	Mix tape	Bass-dB
		4	The ones we love	Bass-dB
		5	We weren't there	Bass-dB
		6	Wreck	Bass-dB
		7	bearlin-roads	SiSEC
		8	tamy-que_pena_tanto_faz	SiSEC
		9	another_dreamer	SiSEC
		10	ultimate_nz_tour	SiSEC
	SA_DS1a Instr.	11	Lussier	TRIOS
		12	Mozart	TRIOS
		13	Take Five	TRIOS
		14	Seed	CCMixter
		15	Free Music	CCMixter
		16	Mind Map1	CCMixter
		17	Mind Map2	CCMixter
	SA_DS1b	18	Saxophone Track 33	Commercial
		19	Saxophone Track 38	
		20	Saxophone Track 45	
		21	Saxophone Track 46	
		22	Saxophone Track 54	
		23	Saxophone Track 64	
		24	Saxophone Track 65	
		25	Saxophone Track 67	
		26	Saxophone Track 68	
		27	Saxophone Track 69	

Table B.1.: Description of the dataset SA_DS1.

Dataset	Num.	Length	Instrumentation	Copyright
SA_DS1	1	35	Male voice, guitar, bass, drums	CC 1.0
	2	57	Male voice, piano, guitar	CC 1.0
	3	46	Male voice, bass, drums, electric guitar, vocals	CC 1.0
	4	16	Male voice, guitar, bass, drums	CC 1.0
	5	32	Male voice, guitar1-2, bass, drums	CC 1.0
	6	19	Male voice, guitar, bass, drums	CC 2.0
	7	14	Male voice, piano, guitar, bass, drums	CC 3.0
	8	13	Guitar, female voice	CC 3.0
	9	25	Male voice, drums, guitar	CC 1.0
	10	18	Female voice, guitar, bass, drums	CC 3.0
	11	17	**Basson**, piano, trumpet	CC 2.0
	12	33	**Clarinet**, piano, viola	CC 2.0
	13	43	**Saxophone**, piano, drums	CC 2.0
	14	16	**Piano**, guitar	CC 3.0
	15	24	**Guitar**, rhythm guitar, organ, piano, bass, drums.	CC 3.0
	16	18	**Whistle**, ukulele	CC 3.0
	17	27	**Kazoo**, ukulele	CC 3.0
	19	16	**Saxophone**, Bass, electric guitar & drums	Schott Music GmBH ISBN 3795751594
	19	66		
	20	34		
	21	59		
	22	43		
	23	95		
	24	54		
	25	49		
	26	67		
	27	56		

Table B.2.: Description of the dataset SA_DS1. The short CC in the copyright column stands for Creative Commons License. For all the instrumental tracks, the solo instrument is shown in bold.

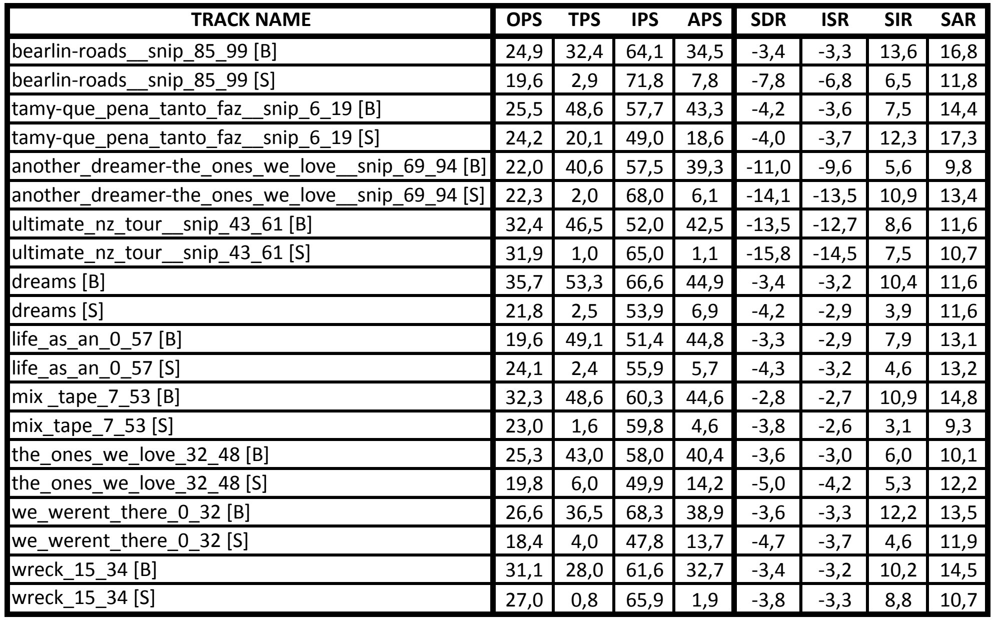
Dataset SA_DS2				
Tracks	**Genre**	**Instrumentation**	**Duration**	**Copyright**
test1	Jazz	Alto sax, drums, piano, bass	30 sec	Schott Music GmBH. ISBN 3795751594
test2	Pop Ballad	Male voice, piano, bass, vocals	30 sec	Mikestar
test3	Swing	Clarinet, piano, drums, bass	30 sec	Grieg Music Education

Table B.3.: Description of the dataset SA_DS2

Appendix C.

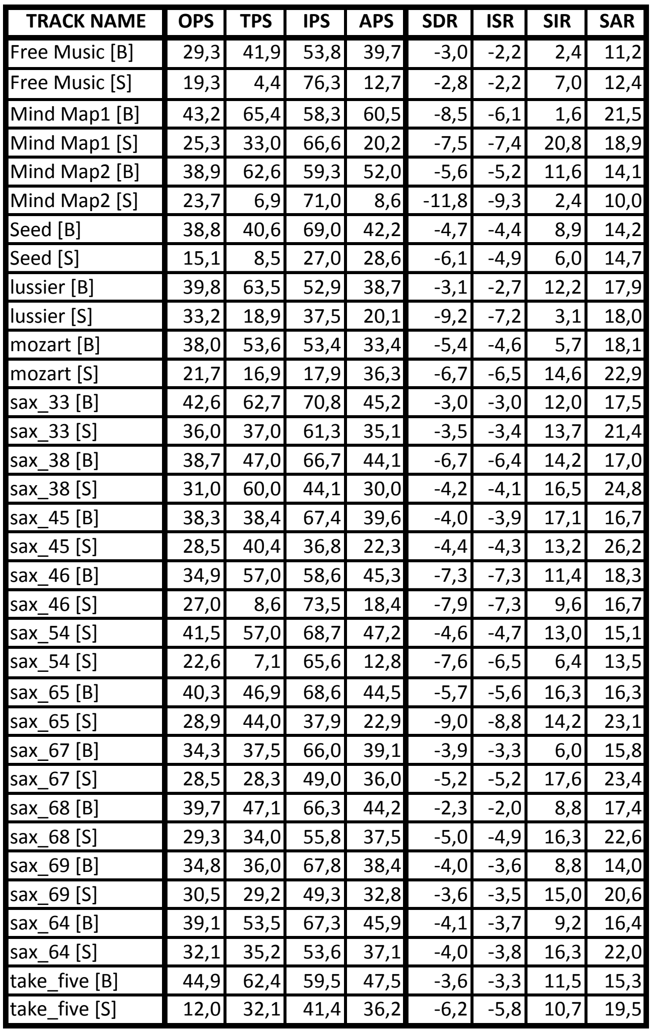
Complete Tables of Perceptual Quality Measures

In all the tables of results presented in this Appendix, both the PEASS measures—Overall Perceptual Score (OPS), Target-Related Perceptual Score (TPS), Interference-Related Perceptual Score (IPS), Artifact-Related Perceptual Score (APS)—, and the standard objective measures—Signal to Distortion Ratio (SDR), Source Image to Spatial Ration (ISR), Source to Artifact Ratio (SAR), and Source to Artifact Ratio (SAR)— are presented for reference.



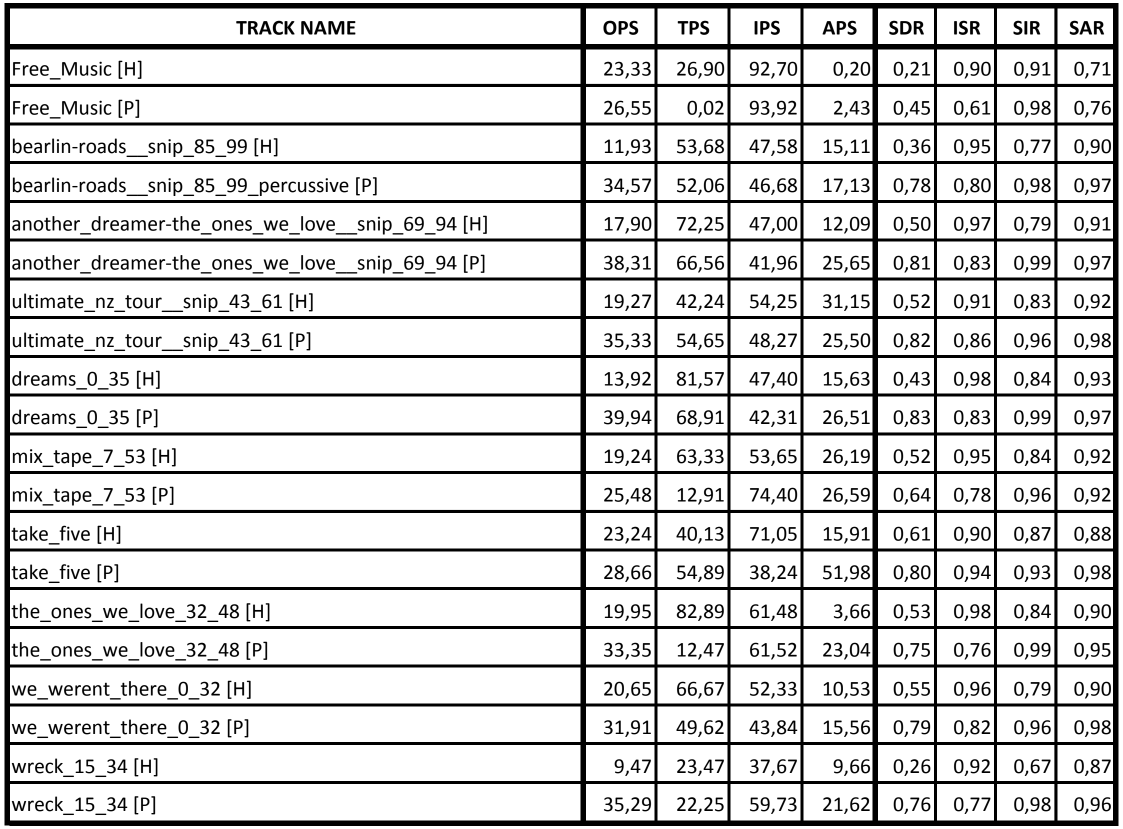
TRACK NAME	OPS	TPS	IPS	APS	SDR	ISR	SIR	SAR
bearlin-roads__snip_85_99 [B]	24,9	32,4	64,1	34,5	-3,4	-3,3	13,6	16,8
bearlin-roads__snip_85_99 [S]	19,6	2,9	71,8	7,8	-7,8	-6,8	6,5	11,8
tamy-que_pena_tanto_faz__snip_6_19 [B]	25,5	48,6	57,7	43,3	-4,2	-3,6	7,5	14,4
tamy-que_pena_tanto_faz__snip_6_19 [S]	24,2	20,1	49,0	18,6	-4,0	-3,7	12,3	17,3
another_dreamer-the_ones_we_love__snip_69_94 [B]	22,0	40,6	57,5	39,3	-11,0	-9,6	5,6	9,8
another_dreamer-the_ones_we_love__snip_69_94 [S]	22,3	2,0	68,0	6,1	-14,1	-13,5	10,9	13,4
ultimate_nz_tour__snip_43_61 [B]	32,4	46,5	52,0	42,5	-13,5	-12,7	8,6	11,6
ultimate_nz_tour__snip_43_61 [S]	31,9	1,0	65,0	1,1	-15,8	-14,5	7,5	10,7
dreams [B]	35,7	53,3	66,6	44,9	-3,4	-3,2	10,4	11,6
dreams [S]	21,8	2,5	53,9	6,9	-4,2	-2,9	3,9	11,6
life_as_an_0_57 [B]	19,6	49,1	51,4	44,8	-3,3	-2,9	7,9	13,1
life_as_an_0_57 [S]	24,1	2,4	55,9	5,7	-4,3	-3,2	4,6	13,2
mix _tape_7_53 [B]	32,3	48,6	60,3	44,6	-2,8	-2,7	10,9	14,8
mix_tape_7_53 [S]	23,0	1,6	59,8	4,6	-3,8	-2,6	3,1	9,3
the_ones_we_love_32_48 [B]	25,3	43,0	58,0	40,4	-3,6	-3,0	6,0	10,1
the_ones_we_love_32_48 [S]	19,8	6,0	49,9	14,2	-5,0	-4,2	5,3	12,2
we_werent_there_0_32 [B]	26,6	36,5	68,3	38,9	-3,6	-3,3	12,2	13,5
we_werent_there_0_32 [S]	18,4	4,0	47,8	13,7	-4,7	-3,7	4,6	11,9
wreck_15_34 [B]	31,1	28,0	61,6	32,7	-3,4	-3,2	10,2	14,5
wreck_15_34 [S]	27,0	0,8	65,9	1,9	-3,8	-3,3	8,8	10,7

Figure C.1.: Results from the frame-based separation algorithm for the vocal dataset



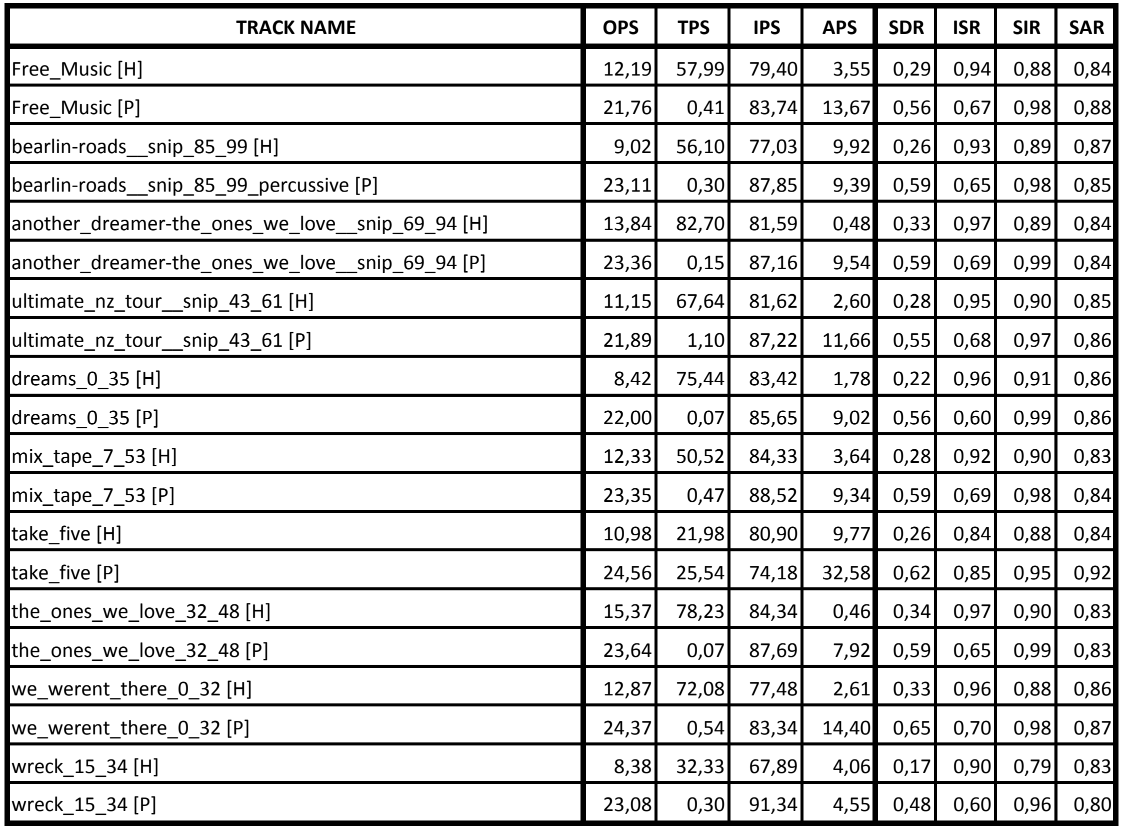
TRACK NAME	OPS	TPS	IPS	APS	SDR	ISR	SIR	SAR
Free Music [B]	29,3	41,9	53,8	39,7	-3,0	-2,2	2,4	11,2
Free Music [S]	19,3	4,4	76,3	12,7	-2,8	-2,2	7,0	12,4
Mind Map1 [B]	43,2	65,4	58,3	60,5	-8,5	-6,1	1,6	21,5
Mind Map1 [S]	25,3	33,0	66,6	20,2	-7,5	-7,4	20,8	18,9
Mind Map2 [B]	38,9	62,6	59,3	52,0	-5,6	-5,2	11,6	14,1
Mind Map2 [S]	23,7	6,9	71,0	8,6	-11,8	-9,3	2,4	10,0
Seed [B]	38,8	40,6	69,0	42,2	-4,7	-4,4	8,9	14,2
Seed [S]	15,1	8,5	27,0	28,6	-6,1	-4,9	6,0	14,7
lussier [B]	39,8	63,5	52,9	38,7	-3,1	-2,7	12,2	17,9
lussier [S]	33,2	18,9	37,5	20,1	-9,2	-7,2	3,1	18,0
mozart [B]	38,0	53,6	53,4	33,4	-5,4	-4,6	5,7	18,1
mozart [S]	21,7	16,9	17,9	36,3	-6,7	-6,5	14,6	22,9
sax_33 [B]	42,6	62,7	70,8	45,2	-3,0	-3,0	12,0	17,5
sax_33 [S]	36,0	37,0	61,3	35,1	-3,5	-3,4	13,7	21,4
sax_38 [B]	38,7	47,0	66,7	44,1	-6,7	-6,4	14,2	17,0
sax_38 [S]	31,0	60,0	44,1	30,0	-4,2	-4,1	16,5	24,8
sax_45 [B]	38,3	38,4	67,4	39,6	-4,0	-3,9	17,1	16,7
sax_45 [S]	28,5	40,4	36,8	22,3	-4,4	-4,3	13,2	26,2
sax_46 [B]	34,9	57,0	58,6	45,3	-7,3	-7,3	11,4	18,3
sax_46 [S]	27,0	8,6	73,5	18,4	-7,9	-7,3	9,6	16,7
sax_54 [S]	41,5	57,0	68,7	47,2	-4,6	-4,7	13,0	15,1
sax_54 [S]	22,6	7,1	65,6	12,8	-7,6	-6,5	6,4	13,5
sax_65 [B]	40,3	46,9	68,6	44,5	-5,7	-5,6	16,3	16,3
sax_65 [S]	28,9	44,0	37,9	22,9	-9,0	-8,8	14,2	23,1
sax_67 [B]	34,3	37,5	66,0	39,1	-3,9	-3,3	6,0	15,8
sax_67 [S]	28,5	28,3	49,0	36,0	-5,2	-5,2	17,6	23,4
sax_68 [B]	39,7	47,1	66,3	44,2	-2,3	-2,0	8,8	17,4
sax_68 [S]	29,3	34,0	55,8	37,5	-5,0	-4,9	16,3	22,6
sax_69 [B]	34,8	36,0	67,8	38,4	-4,0	-3,6	8,8	14,0
sax_69 [S]	30,5	29,2	49,3	32,8	-3,6	-3,5	15,0	20,6
sax_64 [B]	39,1	53,5	67,3	45,9	-4,1	-3,7	9,2	16,4
sax_64 [S]	32,1	35,2	53,6	37,1	-4,0	-3,8	16,3	22,0
take_five [B]	44,9	62,4	59,5	47,5	-3,6	-3,3	11,5	15,3
take_five [S]	12,0	32,1	41,4	36,2	-6,2	-5,8	10,7	19,5

Figure C.2.: Results from the frame-based separation algorithm for the instrumental dataset



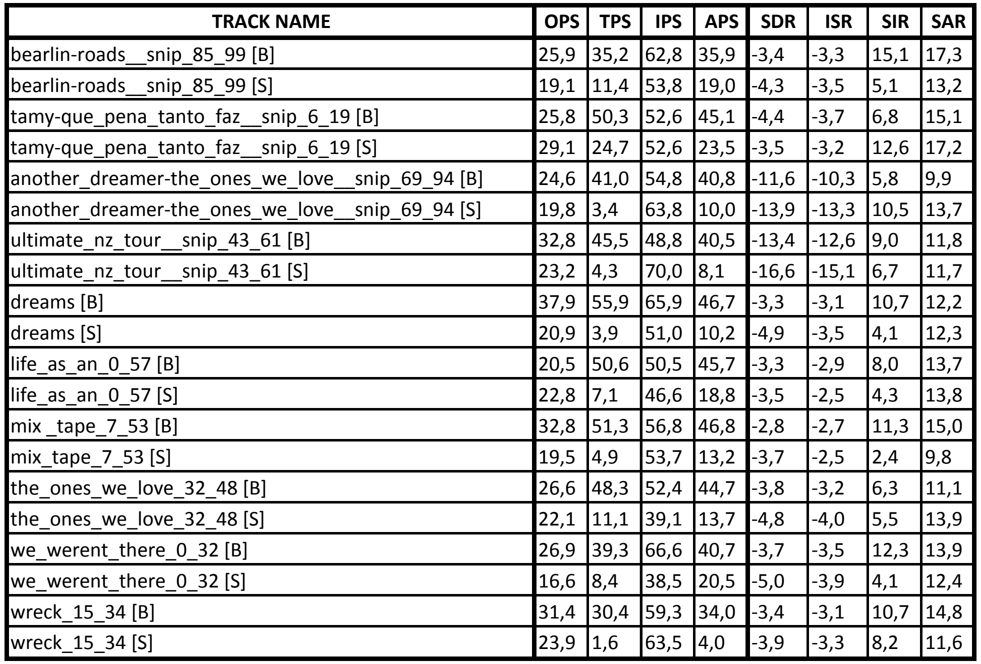
TRACK NAME	OPS	TPS	IPS	APS	SDR	ISR	SIR	SAR
Free_Music [H]	23,33	26,90	92,70	0,20	0,21	0,90	0,91	0,71
Free_Music [P]	26,55	0,02	93,92	2,43	0,45	0,61	0,98	0,76
bearlin-roads__snip_85_99 [H]	11,93	53,68	47,58	15,11	0,36	0,95	0,77	0,90
bearlin-roads__snip_85_99_percussive [P]	34,57	52,06	46,68	17,13	0,78	0,80	0,98	0,97
another_dreamer-the_ones_we_love__snip_69_94 [H]	17,90	72,25	47,00	12,09	0,50	0,97	0,79	0,91
another_dreamer-the_ones_we_love__snip_69_94 [P]	38,31	66,56	41,96	25,65	0,81	0,83	0,99	0,97
ultimate_nz_tour__snip_43_61 [H]	19,27	42,24	54,25	31,15	0,52	0,91	0,83	0,92
ultimate_nz_tour__snip_43_61 [P]	35,33	54,65	48,27	25,50	0,82	0,86	0,96	0,98
dreams_0_35 [H]	13,92	81,57	47,40	15,63	0,43	0,98	0,84	0,93
dreams_0_35 [P]	39,94	68,91	42,31	26,51	0,83	0,83	0,99	0,97
mix_tape_7_53 [H]	19,24	63,33	53,65	26,19	0,52	0,95	0,84	0,92
mix_tape_7_53 [P]	25,48	12,91	74,40	26,59	0,64	0,78	0,96	0,92
take_five [H]	23,24	40,13	71,05	15,91	0,61	0,90	0,87	0,88
take_five [P]	28,66	54,89	38,24	51,98	0,80	0,94	0,93	0,98
the_ones_we_love_32_48 [H]	19,95	82,89	61,48	3,66	0,53	0,98	0,84	0,90
the_ones_we_love_32_48 [P]	33,35	12,47	61,52	23,04	0,75	0,76	0,99	0,95
we_werent_there_0_32 [H]	20,65	66,67	52,33	10,53	0,55	0,96	0,79	0,90
we_werent_there_0_32 [P]	31,91	49,62	43,84	15,56	0,79	0,82	0,96	0,98
wreck_15_34 [H]	9,47	23,47	37,67	9,66	0,26	0,92	0,67	0,87
wreck_15_34 [P]	35,29	22,25	59,73	21,62	0,76	0,77	0,98	0,96

Figure C.3.: Results from the proposed harmonic/percussive separation algorithm



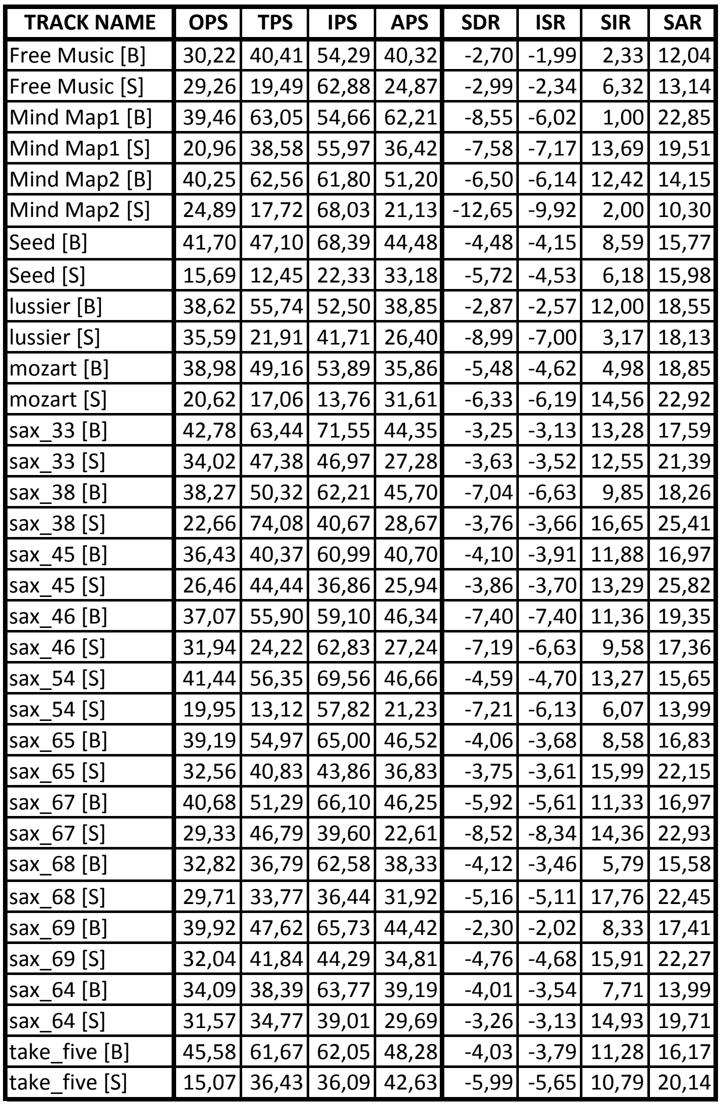
TRACK NAME	OPS	TPS	IPS	APS	SDR	ISR	SIR	SAR
Free_Music [H]	12,19	57,99	79,40	3,55	0,29	0,94	0,88	0,84
Free_Music [P]	21,76	0,41	83,74	13,67	0,56	0,67	0,98	0,88
bearlin-roads__snip_85_99 [H]	9,02	56,10	77,03	9,92	0,26	0,93	0,89	0,87
bearlin-roads__snip_85_99_percussive [P]	23,11	0,30	87,85	9,39	0,59	0,65	0,98	0,85
another_dreamer-the_ones_we_love__snip_69_94 [H]	13,84	82,70	81,59	0,48	0,33	0,97	0,89	0,84
another_dreamer-the_ones_we_love__snip_69_94 [P]	23,36	0,15	87,16	9,54	0,59	0,69	0,99	0,84
ultimate_nz_tour__snip_43_61 [H]	11,15	67,64	81,62	2,60	0,28	0,95	0,90	0,85
ultimate_nz_tour__snip_43_61 [P]	21,89	1,10	87,22	11,66	0,55	0,68	0,97	0,86
dreams_0_35 [H]	8,42	75,44	83,42	1,78	0,22	0,96	0,91	0,86
dreams_0_35 [P]	22,00	0,07	85,65	9,02	0,56	0,60	0,99	0,86
mix_tape_7_53 [H]	12,33	50,52	84,33	3,64	0,28	0,92	0,90	0,83
mix_tape_7_53 [P]	23,35	0,47	88,52	9,34	0,59	0,69	0,98	0,84
take_five [H]	10,98	21,98	80,90	9,77	0,26	0,84	0,88	0,84
take_five [P]	24,56	25,54	74,18	32,58	0,62	0,85	0,95	0,92
the_ones_we_love_32_48 [H]	15,37	78,23	84,34	0,46	0,34	0,97	0,90	0,83
the_ones_we_love_32_48 [P]	23,64	0,07	87,69	7,92	0,59	0,65	0,99	0,83
we_werent_there_0_32 [H]	12,87	72,08	77,48	2,61	0,33	0,96	0,88	0,86
we_werent_there_0_32 [P]	24,37	0,54	83,34	14,40	0,65	0,70	0,98	0,87
wreck_15_34 [H]	8,38	32,33	67,89	4,06	0,17	0,90	0,79	0,83
wreck_15_34 [P]	23,08	0,30	91,34	4,55	0,48	0,60	0,96	0,80

Figure C.4.: Results from the reference harmonic/percussive separation algorithm proposed in [115]



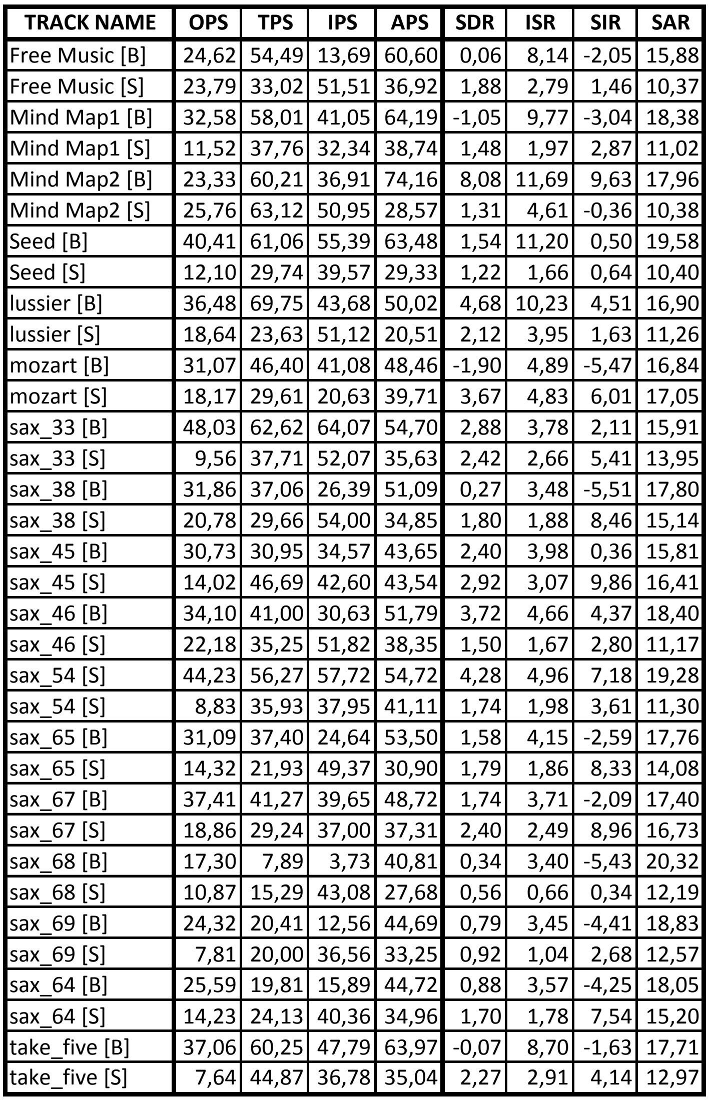
TRACK NAME	OPS	TPS	IPS	APS	SDR	ISR	SIR	SAR
bearlin-roads__snip_85_99 [B]	25,9	35,2	62,8	35,9	-3,4	-3,3	15,1	17,3
bearlin-roads__snip_85_99 [S]	19,1	11,4	53,8	19,0	-4,3	-3,5	5,1	13,2
tamy-que_pena_tanto_faz__snip_6_19 [B]	25,8	50,3	52,6	45,1	-4,4	-3,7	6,8	15,1
tamy-que_pena_tanto_faz__snip_6_19 [S]	29,1	24,7	52,6	23,5	-3,5	-3,2	12,6	17,2
another_dreamer-the_ones_we_love__snip_69_94 [B]	24,6	41,0	54,8	40,8	-11,6	-10,3	5,8	9,9
another_dreamer-the_ones_we_love__snip_69_94 [S]	19,8	3,4	63,8	10,0	-13,9	-13,3	10,5	13,7
ultimate_nz_tour__snip_43_61 [B]	32,8	45,5	48,8	40,5	-13,4	-12,6	9,0	11,8
ultimate_nz_tour__snip_43_61 [S]	23,2	4,3	70,0	8,1	-16,6	-15,1	6,7	11,7
dreams [B]	37,9	55,9	65,9	46,7	-3,3	-3,1	10,7	12,2
dreams [S]	20,9	3,9	51,0	10,2	-4,9	-3,5	4,1	12,3
life_as_an_0_57 [B]	20,5	50,6	50,5	45,7	-3,3	-2,9	8,0	13,7
life_as_an_0_57 [S]	22,8	7,1	46,6	18,8	-3,5	-2,5	4,3	13,8
mix _tape_7_53 [B]	32,8	51,3	56,8	46,8	-2,8	-2,7	11,3	15,0
mix_tape_7_53 [S]	19,5	4,9	53,7	13,2	-3,7	-2,5	2,4	9,8
the_ones_we_love_32_48 [B]	26,6	48,3	52,4	44,7	-3,8	-3,2	6,3	11,1
the_ones_we_love_32_48 [S]	22,1	11,1	39,1	13,7	-4,8	-4,0	5,5	13,9
we_werent_there_0_32 [B]	26,9	39,3	66,6	40,7	-3,7	-3,5	12,3	13,9
we_werent_there_0_32 [S]	16,6	8,4	38,5	20,5	-5,0	-3,9	4,1	12,4
wreck_15_34 [B]	31,4	30,4	59,3	34,0	-3,4	-3,1	10,7	14,8
wreck_15_34 [S]	23,9	1,6	63,5	4,0	-3,9	-3,3	8,2	11,6

Figure C.5.: Results from the tone-based separation algorithm for the vocal dataset



TRACK NAME	OPS	TPS	IPS	APS	SDR	ISR	SIR	SAR
Free Music [B]	30,22	40,41	54,29	40,32	-2,70	-1,99	2,33	12,04
Free Music [S]	29,26	19,49	62,88	24,87	-2,99	-2,34	6,32	13,14
Mind Map1 [B]	39,46	63,05	54,66	62,21	-8,55	-6,02	1,00	22,85
Mind Map1 [S]	20,96	38,58	55,97	36,42	-7,58	-7,17	13,69	19,51
Mind Map2 [B]	40,25	62,56	61,80	51,20	-6,50	-6,14	12,42	14,15
Mind Map2 [S]	24,89	17,72	68,03	21,13	-12,65	-9,92	2,00	10,30
Seed [B]	41,70	47,10	68,39	44,48	-4,48	-4,15	8,59	15,77
Seed [S]	15,69	12,45	22,33	33,18	-5,72	-4,53	6,18	15,98
lussier [B]	38,62	55,74	52,50	38,85	-2,87	-2,57	12,00	18,55
lussier [S]	35,59	21,91	41,71	26,40	-8,99	-7,00	3,17	18,13
mozart [B]	38,98	49,16	53,89	35,86	-5,48	-4,62	4,98	18,85
mozart [S]	20,62	17,06	13,76	31,61	-6,33	-6,19	14,56	22,92
sax_33 [B]	42,78	63,44	71,55	44,35	-3,25	-3,13	13,28	17,59
sax_33 [S]	34,02	47,38	46,97	27,28	-3,63	-3,52	12,55	21,39
sax_38 [B]	38,27	50,32	62,21	45,70	-7,04	-6,63	9,85	18,26
sax_38 [S]	22,66	74,08	40,67	28,67	-3,76	-3,66	16,65	25,41
sax_45 [B]	36,43	40,37	60,99	40,70	-4,10	-3,91	11,88	16,97
sax_45 [S]	26,46	44,44	36,86	25,94	-3,86	-3,70	13,29	25,82
sax_46 [B]	37,07	55,90	59,10	46,34	-7,40	-7,40	11,36	19,35
sax_46 [S]	31,94	24,22	62,83	27,24	-7,19	-6,63	9,58	17,36
sax_54 [S]	41,44	56,35	69,56	46,66	-4,59	-4,70	13,27	15,65
sax_54 [S]	19,95	13,12	57,82	21,23	-7,21	-6,13	6,07	13,99
sax_65 [B]	39,19	54,97	65,00	46,52	-4,06	-3,68	8,58	16,83
sax_65 [S]	32,56	40,83	43,86	36,83	-3,75	-3,61	15,99	22,15
sax_67 [B]	40,68	51,29	66,10	46,25	-5,92	-5,61	11,33	16,97
sax_67 [S]	29,33	46,79	39,60	22,61	-8,52	-8,34	14,36	22,93
sax_68 [B]	32,82	36,79	62,58	38,33	-4,12	-3,46	5,79	15,58
sax_68 [S]	29,71	33,77	36,44	31,92	-5,16	-5,11	17,76	22,45
sax_69 [B]	39,92	47,62	65,73	44,42	-2,30	-2,02	8,33	17,41
sax_69 [S]	32,04	41,84	44,29	34,81	-4,76	-4,68	15,91	22,27
sax_64 [B]	34,09	38,39	63,77	39,19	-4,01	-3,54	7,71	13,99
sax_64 [S]	31,57	34,77	39,01	29,69	-3,26	-3,13	14,93	19,71
take_five [B]	45,58	61,67	62,05	48,28	-4,03	-3,79	11,28	16,17
take_five [S]	15,07	36,43	36,09	42,63	-5,99	-5,65	10,79	20,14

Figure C.6.: Results from the tone-based separation algorithm for the instrumental dataset



TRACK NAME	OPS	TPS	IPS	APS	SDR	ISR	SIR	SAR
Free Music [B]	24,62	54,49	13,69	60,60	0,06	8,14	-2,05	15,88
Free Music [S]	23,79	33,02	51,51	36,92	1,88	2,79	1,46	10,37
Mind Map1 [B]	32,58	58,01	41,05	64,19	-1,05	9,77	-3,04	18,38
Mind Map1 [S]	11,52	37,76	32,34	38,74	1,48	1,97	2,87	11,02
Mind Map2 [B]	23,33	60,21	36,91	74,16	8,08	11,69	9,63	17,96
Mind Map2 [S]	25,76	63,12	50,95	28,57	1,31	4,61	-0,36	10,38
Seed [B]	40,41	61,06	55,39	63,48	1,54	11,20	0,50	19,58
Seed [S]	12,10	29,74	39,57	29,33	1,22	1,66	0,64	10,40
lussier [B]	36,48	69,75	43,68	50,02	4,68	10,23	4,51	16,90
lussier [S]	18,64	23,63	51,12	20,51	2,12	3,95	1,63	11,26
mozart [B]	31,07	46,40	41,08	48,46	-1,90	4,89	-5,47	16,84
mozart [S]	18,17	29,61	20,63	39,71	3,67	4,83	6,01	17,05
sax_33 [B]	48,03	62,62	64,07	54,70	2,88	3,78	2,11	15,91
sax_33 [S]	9,56	37,71	52,07	35,63	2,42	2,66	5,41	13,95
sax_38 [B]	31,86	37,06	26,39	51,09	0,27	3,48	-5,51	17,80
sax_38 [S]	20,78	29,66	54,00	34,85	1,80	1,88	8,46	15,14
sax_45 [B]	30,73	30,95	34,57	43,65	2,40	3,98	0,36	15,81
sax_45 [S]	14,02	46,69	42,60	43,54	2,92	3,07	9,86	16,41
sax_46 [B]	34,10	41,00	30,63	51,79	3,72	4,66	4,37	18,40
sax_46 [S]	22,18	35,25	51,82	38,35	1,50	1,67	2,80	11,17
sax_54 [S]	44,23	56,27	57,72	54,72	4,28	4,96	7,18	19,28
sax_54 [S]	8,83	35,93	37,95	41,11	1,74	1,98	3,61	11,30
sax_65 [B]	31,09	37,40	24,64	53,50	1,58	4,15	-2,59	17,76
sax_65 [S]	14,32	21,93	49,37	30,90	1,79	1,86	8,33	14,08
sax_67 [B]	37,41	41,27	39,65	48,72	1,74	3,71	-2,09	17,40
sax_67 [S]	18,86	29,24	37,00	37,31	2,40	2,49	8,96	16,73
sax_68 [B]	17,30	7,89	3,73	40,81	0,34	3,40	-5,43	20,32
sax_68 [S]	10,87	15,29	43,08	27,68	0,56	0,66	0,34	12,19
sax_69 [B]	24,32	20,41	12,56	44,69	0,79	3,45	-4,41	18,83
sax_69 [S]	7,81	20,00	36,56	33,25	0,92	1,04	2,68	12,57
sax_64 [B]	25,59	19,81	15,89	44,72	0,88	3,57	-4,25	18,05
sax_64 [S]	14,23	24,13	40,36	34,96	1,70	1,78	7,54	15,20
take_five [B]	37,06	60,25	47,79	63,97	-0,07	8,70	-1,63	17,71
take_five [S]	7,64	44,87	36,78	35,04	2,27	2,91	4,14	12,97

Figure C.7.: Results from the REPET separation algorithm for the instrumental dataset

List of Figures

List of Tables

Glossary of Acronyms

Acronyms

AMT	Automatic Music Transcription
APS	Artifacts-related Perceptual Score
BSS	Blind Source Separation
CAM	Common Amplitude Modulation
CISS	Coding-based Informed Source Separation
CQT	Constant-Q Transform
DTW	Dynamic Time Warping
FFT	Fast Fourier Transform
GCTF	Generalized Coupled Tensor Factorization
IBM	Ideal Binary Masks
IF	Instantaneous Frequency
IFD	Instantaneous Frequency Distribution
IG	Inverse Gamma
IPS	Interference-related Perceptual Score
ISS	Informed Source Separation
IS	Itakura-Saito
ISTFT	Inverse Short Time Fourier Transform
KL	Kullback-Leibler
MIR	Music Information Retrieval
MDCT	Modified Discrete Cosine Transform
MFCC	Mel Frequency Cepstral Coefficients
MIREX	Music Information Retrieval Evaluation eXchange
MMSE	Minimum Mean-Square Error
MP	Matching Pursuits
MSTFT	Modified Short Time Fourier Transform
NMF	Non-negative Matrix Factorization
NTF	Non-negative Tensor Factorization
OMP	Orthogonal Matching Pursuits
OPS	Overall Perceptual Score
OSC	Octave-based Spectral Contrast
PEASS	Perceptual Evaluation Methods for Audio Source Separation
PLCA	Probabilistic Latent Component Analysis
PLCA	Probabilistic Latent Component Analysis
SAOC	Spatial Audio Object Coding
SiSEC	Signal Separation Evaluation Campaign
SNR	Signal to Noise Ratio
STFT	Short Time Fourier Transform
SVM	Support Vector Machine
TPS	Target-related Perceptual Score

Glossary

blue is the color of glossary terms. 6

BPM Beats per minute is a unit typically used as a measure of tempo in music. 93

brass A brass instrument is a musical instrument that produces sound by sympathetic vibration of air in a tubular resonator in sympathy with the vibration of the player's lips. The trumpet, trombone, and tuba are example of brass instruments. 95

cent is a logarithmic unit of measure for musical intervals. An equally tempered semitone spans 100 cents. An octave spans of 12 semitones and thus, 1200 cents. 58

double reeds The term double reed comes from the fact that some musical instruments produce sound by the vibration against each other of two pieces of cane. The bassoon, oboe, and English horn are examples of double reed instruments. 95

equal temperament is a system of tuning in which every pair of adjacent notes has an identical frequency ratio. 97

F-Measure can be interpreted as a weighted average of the precision and recall.. 74

major a major chord is one whose third degree is a major third above the tonic or root note. 97

MIDI is a technical standard that describes a protocol, digital interface and connectors and allows a wide variety of electronic musical instruments, computers and other related devices to connect and communicate with one another. MIDI carries event messages that specify notation, pitch and velocity, control signals for parameters such as volume, vibrato, audio panning, cues, and clock signals that set and synchronize tempo between multiple devices. 24

minor a minor chord is one whose third degree is a minor third above the tonic or root note. 97

mixture is the resulting signal after combining multiple recorded sounds into one or more audio channels. 56

monaural single-channel signal. 56

Precision in information retrieval, Precision is the fraction of relevant instances that are retrieved. 74

real-time In digital signal processing, real-time means that the mean processing time per sample is no greater than the sampling period. 68

Recall in information retrieval, Recall is the fraction of retrieved instances that are relevant. 74

Bibliography

Publications as Co- or First Author

[O1] Krasser, J., Abeßer, J., Grossmann, H., Dittmar, C. & Cano, E. Improved Music Similarity Computation based on Tone Objects Categories. In *Audio Mostly Conference*, 47–54 (Corfu, Greece, 2012).

[O2] Cano, E., Dittmar, C. & Grollmisch, S. Songs2See: Learn to Play by Playing. In *12th International Society for Music Information Retrieval Conference (ISMIR 2011)* (Miami, USA, 2011).

[O3] Cano, E., Schuller, G. & Dittmar, C. Exploring Phase Information in Sound Source Separation Applications. In *13th International Conference on Digital Audio Effects (DAFx-10)*, 1–8 (Graz, Austria, 2010).

[O4] Cano, E., Dittmar, C. & Schuller, G. Influence of Phase, Magnitude, and Location of Harmonic Components in the Perceived Quality of Extracted Solo Signals. In *AES 42nd International Conference on Semantic Audio*, 1–6 (Ilmenau, Germany, 2011).

[O5] Cano, E., Dittmar, C. & Schuller, G. Efficient Implementation of a System for Solo and Accompaniment Separation in Polyphonic Music. In *20th European Signal Processing Conference (EUSIPCO 2012)*, 285–289 (Bucharest, Romania, 2012).

[O6] Cano, E., Grollmisch, S. & Dittmar, C. Songs2See : Towards a New Generation of Music Performance Games. In *9th International Symposium on Computer Music Modeling and Retrieval CMMR*, 19–22 (London, UK, 2012).

[O7] Dittmar, C., Cano, Estefanía, Abeßer, J. & Grollmisch, S. Music Information Retrieval Meets Music Education. In Müller, M., Goto, M. & Schedl, M. (eds.) *Multimodal Music Processing*, 95–120 (Dagstuhl Publishing, 2012).

[O8] Cano, E., Dittmar, C. & Schuller, G. Re-thinking Sound Separation: Prior Information and Additivity Constraint in Separation Algorithms. In *16th International Conference on Digital Audio Effects (DAFx-13)*, 1–7 (Maynooth, Ireland, 2013).

References by Other Authors

[9] Salamon, J., Gómez, E., Ellis, D. P. W. & Richard, G. Melody Extraction from Polyphonic Music Signals : Approaches, Applications and Challenges. *IEEE Signal Processing Magazine* (2013).

[10] FitzGerald, D. Upmixing from mono; a source separation approach. In *17th International Conference on Digital Signal Processing* (Corfu, Greece, 2011).

[11] Liutkus, A., Ozerov, A., Badeau, R. & Richard, G. Spatial Coding-based Informed Source Separation. In *20th European Signal Processing Conference (EUSIPCO 2012)*, 2407–2411 (Bucharest, Romania, 2012).

[12] Salamon, J. & Gómez, E. Melody Extraction from Polyphonic Music Signals using Pitch Contour Characteristics. *IEEE Transactions on Audio, Speech and Language Processing* **20**, 1759–1770 (2012).

[13] Liutkus, A., Durrieu, J.-L., Daudet, L. & Richard, G. An Overview of Informed Audio Source Separation. In *14th International Workshop on Image and Audio Analaysis for Multimedia Interactive Services*, 3–6 (Paris, France, 2013).

[14] Department of Geography. University of Oregon. Color Schemes Appropriate for Scientific Data. URL `http://geography.uoregon.edu/datagraphics/color_scales.htm`.

[15] Fraunhofer IDMT. Songs2See: Learn to Play by Playing. URL `http://www.idmt.fraunhofer.de/en/Service_Offerings/technologies/q_t/songs2see.html`.

[16] Le Roux, J. *et al.* Consistent Wiener Filtering : Generalized Time-Frequency Masking Respecting Spectrogram Consistency. In *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)* (St. Malo, France, 2010).

[17] Duong, N. Q. K., Vincent, E. & Gribonval, R. Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model. *IEEE Transactions on Audio, Speech and Language Processing* **18**, 1830–1840 (2010).

[18] Benaroya, L., Donagh, L. M., Bimbot, F. & Gribonval, R. Non-Negative Sparse Representation for Wiener based Source Separation with Single Sensor. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, 613–616 (2003).

[19] Li, Y. & Wang, D. On the Optimality of Ideal Binary Time - Frequency Masks. *Speech Communication* **51**, 230–239 (2009).

[20] FitzGerald, D. & Jaiswal, R. On the Use of Masking Filters in Sound Source Separation. In *15th International Conference on Digital Audio Effects (DAFx-12)*, 1–7 (York, UK, 2012).

[21] Röbel, A. Between Physics and Percepction: Signal Models for High Level Audio Processing. In *13th International Conference on Digital Audio Effects (DAFx-10)* (Graz, Austria, 2010).

[22] Burred, J. J. *From Sparse Models to Timbre Learning : New Methods for Musical Source Separation.* Ph.D. thesis, Technischen Universität Berlin (2009).

[23] Serra, X. & Smith, J. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal* 12–24 (1990).

[24] Fant, G. The Source Filter Concept in Voice Production. *STL-QPSR* **22**, 21–37 (1981).

[25] Durrieu, J.-L., David, B. & Richard, G. A Musically Motivated Mid-Level Representation for Pitch Estimation and Musical Audio Source Separation. *IEEE Journal of Selected Topics in Signal Processing* **5**, 1180–1191 (2011).

[26] Plumbley, M. D., Blumensath, T., Daudet, L., Gribonval, R. & Davies, M. Sparse Representations in Audio and Music: From Coding to Source Separation. *Proceedings of the IEEE* **98**, 995–1005 (2010).

[27] Rath, G. & Sahoo, A. A Comparative Study of some Greedy Pursuit Algorithms for Sparse Approximations. In *17th European Signal Processing Conference (EUSIPCO 2009)*, Eusipco, 398–402 (Glasgow, Scotland, 2009).

[28] Huang, P.-S., Chen, S. D., Smaragdis, P. & Hasegawa-Johnson, M. Singing-voice Separation from Monaural Recordings using Robust Principal Component Analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, 57–60 (Kyoto, Japan, 2012).

[29] Rafii, Z. & Pardo, B. REpeating Pattern Extraction Technique (REPET): A Simple Method for Music / Voice Separation. *IEEE Transactions on Audio, Speech and Language Processing* **21**, 73–84 (2013).

[30] Lee, D. D., Laboratories, B., Hill, M. & Ý, H. S. S. Algorithms for Non-negative Matrix Factorization. In *Advances in Neural Information Processing Systems 13*, 1, 556–562 (MIT Press, Lee2001, 2001).

[31] Févotte, C., Bertin, N. & Durrieu, J.-L. Nonnegative Matrix Factorization with the Itakura-Saito Divergence: with Application to Music Analysis. *Neural computation* **21**, 793–830 (2009).

[32] FitzGerald, D., Cranitch, M. & Coyle, E. On the use of the beta divergence for musical source separation. In *Signals and Systems Conference (ISSC 2009), IET Irish*, 1–6 (2009).

[33] Wang, Y.-X. & Zhang, Y.-J. Nonnegative Matrix Factorization: A Comprehensive Review. *IEEE Transactions on Knowledge and Data Engineering* **25**, 1336–1353 (2013).

[34] Smaragdis, P. Non-Negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs. In *5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2004)*, 494–499 (Granada, Spain, 2004).

[35] Virtanen, T. Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria. *IEEE Transactions on Audio, Speech, and Language Processing* **15**, 1066–1074 (2011).

[36] Virtanen, T. O. Monaural Sound Source Separation by Perceptually Weighted Non-Negative Matrix Factorization. Tech. Rep., Tampere University of Technology (2007).

[37] Kirbiz, S. & Gunsel, B. Perceptually Weighted Non-Negative Matrix Factorization for Blind Single-Channel Music Source Separation. In *21st International Conference on Pattern Recognition (ICPR2012)*, Icpr, 226–229 (Tsukuba, Japan, 2012).

[38] Smaragdis, P., Raj, B. & Shashanka, M. Supervised and Semi-Supervised Separation of Sounds from Single-Channel Mixtures. In *7th International Conference on Independent Component Analysis and Signal Separation (ICA 2007)* (London, UK, 2007).

[39] Ganseman, J., Scheunders, P., Mysore, G. J. & Abel, J. S. Evaluation of a Score-Informed Source Separation System. In *11th International Society for Music Information Retrieval Conference (ISMIR 2010)* (Utrecht, Netherlands, 2010).

[40] Guo, Y. & Zhu, M. Audio Source Separation by Basis Function Adaptation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 20011)*, 2192–2195 (Prague, Czech Republic, 2011).

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

[42] Fitzgerald, D., Cranitch, M. & Coyle, E. Using Tensor Factorisation Models to Separate Drums from Polyphonic Music. In *12th International Conference on Digital Audio Effects (DAFx-09)*, 1–5 (Como, Italy, 2009).

[43] Simsekli, U. & Cemgil, A. T. Score Guided Musical Source Separation using Generalized Coupled Tensor Factorization. In *20th European Signal Processing Conference (EUSIPCO 2012)*, 2639–2643 (Bucharest, Romania, 2012).

[44] Ozerov, A., Liutkus, A., Badeau, R. & Richard, G. Coding-Based Informed Source Separation : Nonnegative Tensor Factorization Approach. Tech. Rep., Institut Mines-Télécom; TÉLÉCOM PARISTECH (2012).

[45] Dressler, K. Pitch Estimation by the Pair-Wise Evaluation of Spectral Peaks. In *AES 42nd International Conference on Semantic Audio*, 1–10 (Ilmenau, Germany, 2011).

[46] Li, Y., Woodruff, J. & Wang, D. Monaural Musical Sound Separation based on Pitch and Common Amplitude Modulation. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **17**, 1361–1371 (2009).

[47] Klapuri, A. P. Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness. *IEEE Transactions on Speech and Audio Processing* **11**, 804–816 (2003).

[48] Marxer, R., Janer, J. & Bonada, J. Low-Latency Instrument Separation in Polyphonic Audio using Timbre Models. *Latent Variable Analysis and Signal Separation* **7191**, 314–321 (2012).

[49] Ewert, S., Müller, M. & Grosche, P. High Resolution Audio Synchronization using Chroma Onset Features. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009)*, 1869–1872 (Taipei, Taiwan, 2009).

[50] Müller, M. *Information Retrieval for Music and Motion* (Springer Verlag, 2007).

[51] Bosch, J. J., Kondo, K., Marxer, R. & Janer, J. Score-Informed and Timbre Independent Lead Instrument Separation in Real-World Scenarios. In *20th European Signal Processing Conference (EUSIPCO 2012)*, 2417–2421 (Bucharest, Romania, 2012).

[52] Joder, C. & Schuller, B. Score-Informed Leading Voice Separation from Monaural Audio. In *13th International Society for Music Information Retrieval Conference (ISMIR)*, 277–282 (Porto, Portugal, 2012).

[53] Fritsch, J. & Plumbley, M. D. Score Informed Audio Source Separation using Constrained Nonnegative Matrix Factorization and Score Synthesis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, 888–891 (Vancouver, Canada, 2013).

[54] Duan, Z. & Pardo, B. Soundprism : An Online System for Score-Informed Source Separation of Music Audio. *IEEE Journal of Selected Topics in Signal Processing* **5**, 1205–1215 (2011).

[55] Ewert, S. & Müller, M. Score-Informed Source Separation for Music Signals. In Müller, M., Goto, M. & Schedl, M. (eds.) *Multimodal Music Processing*, vol. 3, 73–94 (Dagstuhl Publishing, 2012).

[56] Hsu, C.-L. & Jang, J.-S. R. On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset. *IEEE Transactions on Audio, Speech, and Language Processing* **18**, 310–319 (2009).

[57] Vembu, S. & Baumann, S. Separation of Vocals from Polyphonic Audio Recordings. In *6th International Society for Music Information Retrieval Conference (ISMIR 2005)* (2005).

[58] Coïc, M. & Burred, J. J. Bayesian Non-Negative Matrix Factorization with Learned Temporal Smoothness Priors. In *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 280–287 (Tel-Aviv, Israel, 2012).

[59] Ozerov, A., Liutkus, A., Badeau, R. & Richard, G. Informed Source Separation: Source Coding Meets Source Separation. In *20011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 8–11 (New Paltz, NY, 2011).

[60] Parvaix, M., Girin, L. & Brossier, J.-M. A Watermarking-Based Method for Informed Source Separation of Audio Signals with a Single Sensor. *IEEE Transactions on Audio, Speech, and Language Processing* **18**, 1464–1475 (2010).

[61] Smaragdis, P. & Mysore, G. J. Separation by Humming: User-Guided Sound Extraction from Monophonic Mixtures. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (New Paltz, NY, 2009).

[62] FitzGerald, D. User Assisted Separation using Tensor Factorisations. In *20th European Signal Processing Conference (EUSIPCO 2012)*, 2412–2416 (Bucharest, Romania, 2012).

[63] Gerber, T., Dutasta, M., Girin, L. & Févotte, C. Professionally-Produced Music Separation Guided by Covers. In *13th International Society for Music Information Retrieval Conference (ISMIR)*, 85–90 (Porto, Portugal, 2012).

[64] Leveau, P., Maller, S., Burred, J. J. & Juareguiberry, X. Convolutive Common Audio Signal Extraction. In *20011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 165–168 (New Paltz, NY, 2011).

[65] Durrieu, J.-L. & Thiran Jean-Philippe. Musical Audio Source Separation based on User-Selected F0 Track. *Latent Variable Analysis and Signal Separation* 1–8 (2012).

[66] Fuentes, B., Badeau, R. & Richard, G. Blind Harmonic Adaptive Decomposition Applied to Supervised Source Separation. In *20th European Signal Processing Conference (EUSIPCO 2012)*, 2654–2658 (Bucharest, Romania, 2012).

[67] Vincent, E., Jafari, M. G. & Plumbley, M. D. Preliminary Guidelines for Subjective Evaluation of Audio Source Separation Algorithms. In *ICA Research Network International Workshop*, 93–96 (Liverpool, UK, 2006).

[68] ITU. RECOMMENDATION ITU-R BS.1284-1 General Methods for the Subjective Assessment of Sound Quality. Tech. Rep. (2003).

[69] ITU. RECOMMENDATION ITU-R BS.1116-1 Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems. Tech. Rep. (1997).

[70] ITU. RECOMMENDATION ITU-R BS . 1534-1 Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems. Tech. Rep. (2003).

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

[72] Févotte, C., Gribonval, R. & Vincent, E. BSS_Eval Toolbox User Guide Revision 2.0. Tech. Rep., IRISA, Rennes, Bretagne Atlantique (2011).

[73] Vincent, E., Sawada, H., Bofill, P., Makino, S. & Rosca, J. P. First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results. In *7th International Conference on Independent Component Analysis and Signal Separation (ICA 2007)*, 552–559 (London, UK, 2007).

[74] Hörtech GmbH. PEMO-Q. URL `http://www.hoertech.de/web_en/produkte/downloads.shtml`.

[75] INRIA. The PEASS Software. URL `http://bass-db.gforge.inria.fr/peass/PEASS-Software.html`.

[76] Vincent, E. *et al.* The Signal Separation Evaluation Campaign (2007 - 2010): Achievements and Remaining Challenges. Tech. Rep., INRIA, Rennes, Bretagne Atlantique (2011).

[77] Araki, S., Nesta, F. & Vincent, E. The 2011 Signal Separation Evaluation Campaign (SiSEC2011): Audio Source Separation. In *Latent Variable Analysis and Signal Separation*, 414–422 (2012).

[78] INRIA. BASS-dB: Multitrack Recordings (2003). URL `http://bass-db.gforge.inria.fr/BASS-dB/?show=browse&id=mtracks`.

[79] IRISA. SiSEC 2008 (2008). URL `http://sisec2008.wiki.irisa.fr/tiki-index.php`.

[80] IRISA. SiSEC 2010. URL `http://sisec2010.wiki.irisa.fr/tiki-index.php?page=Professionally+produced+music+recordings`.

[81] Fritsch, J. & Queen Mary University of London. The TRIOS Score-aligned Multitrack Recordings Dataset (2012). URL `http://c4dm.eecs.qmul.ac.uk/rdr/handle/123456789/27`.

[82] IRISA. SiSEC 2013 (2013). URL `http://sisec.wiki.irisa.fr/tiki-index.php?page=Professionally+produced+music+recordings`.

[83] Cano, E. & Fraunhofer IDMT. Solo and Accompaniment Separation: Towards its use in Music Education Applications (2013). URL http://www.idmt.fraunhofer.de/en/Departments_and_Groups/smt/solo_and_accompaniment_separation.html.

[84] INRIA. SiSEC. URL http://sisec.wiki.irisa.fr/tiki-index.php.

[85] Bay, M. & Beauchamp, J. W. Harmonic Source Separation Using Prestored Spectra. In *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 561–568 (Charlestone, USA, 2006).

[86] Wolfe, J. What is Sound Spectrum? URL http://www.phys.unsw.edu.au/jw/sound.spectrum.html.

[87] Every, M. R. & Szymanski, J. E. Separation of Synchronous Pitched Notes by Spectral Filtering of Harmonics. *IEEE Transactions on Audio, Speech and Language Processing* **14**, 1845–1856 (2006).

[88] Wolfe, J., Garnier, M. & Smith, J. Singing Voice Acoustics: An Introduction. URL http://www.phys.unsw.edu.au/jw/voice.html#resonances.

[89] Gunawan, D. & Sen, D. Separation of Harmonic Musical Instrument Notes Using Spectro-Temporal Modeling of Harmonic Magnitudes and Spectrogram Inversion with Phase Optimization. *Journal of the Audio Engineering Society (AES)* **60** (2012).

[90] Viste, H. & Evangelista, G. A Method for Separation of Overlapping Partials based on Similarity of Temporal Envelopes in Multichannel Mixtures. *IEEE Transactions on Audio, Speech and Language Processing* **14**, 1051–1061 (2006).

[91] Fletcher, N. H. Mode Locking in Nonlinearly Excited Inharmonic Musical Oscillators. *Journal of the Acoustical Society of America* **64**, 1566–1569 (1978).

[92] Brown, J. C. Frequency Ratios of Spectral Components of Musical Sounds. *Journal of the Acoustical Society of America* **99**, 1210–1218 (1996).

[93] Ando, S. & Yamaguchi, K. Statistical Study of Spectral Parameters in Musical Instrument Tones. *Journal of the Acoustical Society of America* **94**, 37–45 (1993).

[94] Fletcher, N. H. & Rossing, T. D. *The Physics of Musical Intruments* (Springer, New York, 1998), 2nd edn.

[95] Fastl, H. & Völk, F. Inharmonicity of Sounds from Electric Guitars: Physical Flaw or Musical Asset? In *International Conference on Music Perception and Cognition* (Sapporo, Japan, 1998).

[96] Järveläinen, H., Välimäki, V. & Karjalainen, M. Audibility of Inharmonicity in String Instrument Sounds, and Implications to Digital Sound Synthesis. In *International Computer Music Conference (ICMC 1999)*, 359–362 (Beijing, China, 1999).

[97] Godsill, S. & Davy, M. Bayesian Computational Models for Inharmonicity in Musical Instruments. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 283–286 (New Paltz, NY, 2005).

[98] Dixon, S., Mauch, M. & Tidhar, D. Estimation of Harpsichord Inharmonicity and Temperament from Musical Recordings. *The Journal of the Acoustical Society of America* **131**, 878–87 (2012).

[99] Rigaud, F., David, B. & Daudet, L. Piano Sound Analysis Using Non-Negative Matrix Factorization with Inharmonicty Constraint. In *20th European Signal Processing Conference (EUSIPCO 2012)*, 1–5 (Bucharest, Romania, 2012).

[100] Kob, M. *et al.* Analysing and Understanding the Singing Voice: Recent Progress and Open Questions. *Current Bioinformatics* **6**, 362–374 (2011).

[101] Rad, A. B. & Virtanen, T. Phase Spectrum Prediction of Audio Signals. In *5th International Symposium on Communications, Control, and Signal Processing (ISCCP)* (Rome, Italy, 2012).

[102] Alsteris, L. D. & Paliwal, K. K. Short-time Phase Spectrum in Speech Processing: A Review and Some Experimental Results. *Digital Signal Processing* **17**, 578–616 (2007).

[103] Wang, D. L. & Lim, J. S. The Unimportance of Phase in Speech Enhancement. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **30**, 679–681 (1982).

[104] Paliwal, K. K. & Alsteris, L. D. On the Usefulness of STFT Phase Spectrum in Human Listening Tests. *Speech Communication* **45**, 153–170 (2005).

[105] Shi, G., Shanechi, M. M. & Aarabi, P. On the Importance of Phase in Human Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing* **14**, 1867–1874 (2006).

[106] Ewert, S. *Signal Processing Methods for Music Synchronization, Audio Matching, and Source Separation.* Ph.D. thesis, Rheinischen Friedrich-Wilhelms-Universität Bonn (2012).

[107] Griffin, D. W. & Lim, J. S. Signal Estimation from Modified Short-Time Fourier Transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **32**, 236–243 (1984).

[108] Woodruff, J., Li, Y. & Wang, D. Resolving Overlapping Harmonics for Monaural Musical Sound Separation using Pitch and Common Amplitude Modulation. In *9th International Society for Music Information Retrieval Conference (ISMIR 2008)*, 538–543 (Philadelphia, USA, 2008).

[109] of IOWA. Electronic Music Studios, U. Musical Instrument Samples (2013). URL `http://theremin.music.uiowa.edu/MIS.html`.

[110] Abe, T., Takao, K. & Imai, S. The IF Spectrogram: The New Spectral Representation. In *International Symposium on Simulation, Visualization and Auralization for Acoustic Research and Education* (Tokyo, Japan, 1997).

[111] Bregman, A. S. *Auditory Scene Analysis* (The MIT Press, Cambridge, MA, 1990).

[112] Dubnov, S. & Rodet, X. Investigation of Phase Coupling Phenomena in Sustained Portion of Musical Instruments Sound. *The Journal of the Acoustical Society of America* **113**, 348 (2003).

[113] Opticom. Perceptual Evaluation of Audio Quality (PEAQ). URL `http://www.peaq.org/`.

[114] Ono, N., Miyamoto, K. & Roux, J. L. Separation of a Monaural Audio Signal into Harmonic/Percussive Components by Complementary Diffusion on Spectrogram. In *16th European Signal Processing Conference (EUSIPCO 2008)*, 1–4 (Lausanne, Switzerland, 2008).

[115] Fitzgerald, D. Harmonic/Percussive Separation Using Median Filtering. In *13th International Conference on Digital Audio Effects (DAFx-10)*, 10–13 (Graz, Austria, 2010).

[116] Luce, D. & Clark, M. J. Durations of Attack Transients of Nonpercussive Orchestral Instruments. *Journal of the Audio Engineering Society (AES)* **13**, 194–200 (1965).

[117] MIREX. Multiple Fundamental Frequency Estimation and Tracking Results. URL http://www.music-ir.org/mirex/wiki/2013:Multiple_Fundamental_Frequency_Estimation_%26_Tracking_Results.

[118] Cannam, C., Landone, C. & Sandler, M. Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files. In *Proceedings of the ACM Multimedia 2010 International Conference*, 1467—-1468 (Firenze, Italy, 2010).

[119] Sonic Visualiser. URL http://www.sonicvisualiser.org/.

[120] IRISA. Results SiSEC 2011 (2011). URL http://www.irisa.fr/metiss/SiSEC11/professional/test_eval2011.htm.

[121] Janer, J. & Marxer, R. Separation of Unvoiced Fricatives in Singing Voice Mixtures with Music Semi-Supervised NMF. In *16th International Conference on Digital Audio Effects (DAFx-13)*, 1–4 (Maynooth, Ireland, 2013).

[122] IRISA. Results SiSEC 2013 (2013). URL http://www.onn.nii.ac.jp/sisec13/evaluation_result/MUS/testMUS2013.htm.

[123] Rafii, Zafar and Pardo, Bryan. REpeating Pattern Extraction Technique (REPET). URL http://music.cs.northwestern.edu/research.php?project=repet#download.

[124] Stowell, D. & Dixon, S. MIR in School? Lessons from Ethnographic Observation of Secondary School Music Classes. In *12th International Society for Music Information Retrieval Conference (ISMIR 2011)* (Miami, USA, 2011).

[125] Grollmisch, S., Dittmar, C. & Gatzsche, G. Concept, Implementation and Evaluation of an Improvisation based Music Video Game (2009).

[126] Zapata, J. R. & Gómez, E. Improving Beat Tracking in the Presence of Highly Predominant Vocals Using Source Separation Techniques: Preliminary Study. In *9th International Symposium on Computer Music Modeling and Retrieval (CMMR 2012)* (London, UK).

Index