An iterative, residual-based approach to unsupervised musical source separation in single-channel mixtures

Georgios Siamantas

A thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy to

University of York
Department of Electronics

December 2009

Abstract

This thesis concentrates on a major problem within audio signal processing, the separation of source signals from musical mixtures when only a single mixture channel is available. Source separation is the process by which signals that correspond to distinct sources are identified in a signal mixture and extracted from it. Producing multiple entities from a single one is an extremely underdetermined task, so additional prior information can assist in setting appropriate constraints on the solution set. The approach proposed uses prior information such that: (1) it can potentially be applied successfully to a large variety of musical mixtures, and (2) it requires minimal user intervention and no prior learning/training procedures (*i.e.*, it is an unsupervised process). This system can be useful for applications such as remixing, creative effects, restoration and for archiving musical material for internet delivery, amongst others.

Here, specific priors include that the signal contains detectable musical events, with characteristic partial structures, often assumed to be harmonic. The harmonicity cue is incorporated by employing an adapted and extended frame-based multiF0 estimator for identifying the sources. This acts as a front-end to a source estimation and extraction stage. Further, an iterative procedure is introduced between the two stages, enabling improved performance via increased adaptivity to signal content: this novel approach becomes possible by exploiting a residual signal.

Experimental results show that the proposed residual-based method achieves better average performance compared to alternative methods in terms of source separation and multiF0 estimation on a range of mixtures of varying complexity. Unmodelled content of the separated mixture will appear in the residual, which can be exploited further. In particular, a novel onset detection technique is proposed that works entirely with the residual. Considering its simplicity, the technique shows promising results compared to two existing methods that do not use the residual.

Contents

List of Figures

List of Tables

Acknowledgements

I want to express my warmest gratitude to my supervisor, Dr. John Szymanski. His immense support, patience and enthusiasm were absolutely invaluable to me in order to go forward with this. I feel lucky to have learnt so much from him about the nature of research and probably so many other things I have not yet realised. Also, I want to thank Mark Every for providing me with the code of his system and being a positive example of what it is to be a young and passionate researcher; his influence during the first year of my research was crucial. Many thanks go to Tony Tew for his genuine interest in my studies, his always careful and valuable advice and comments on my thesis. Additional thanks go to Prof. Mark Plumbley; his constructive comments have been greatly appreciated and have been crucial for the improvement of my thesis. To Brenda Christmas, the Graduate and Research Support Manager I thank her for doing her best to take a load of worries off me. A big thank you also goes to Anssi Klapuri for offering the code for his multiF0 estimator, as well as to John Woodruff, Zhiyao Duan, Chunghsin Yeh and Sebastian Böck for running their systems on my data.

To the other (ex- and current) PhD students in the audio lab: Rachel Van Besouw, Alastair Moore, Alex Southern, Sandra Pauletto, Jude Brereton, Simon Shelley, Jonathan Thorpe, Helena Daffern, Sorrel Hoare, Aglaia Foteinou, Matt Speed, Andrew Chadwick, Christin Kirchhübel, Anocha Rugchatjaroen; as well as the rest of the teaching staff: Jez Wells, Damian Murphy, David Howard and Andy Hunt, you have given me enormous support and you have contributed to a warm and stimulating research environment, thank you.

Furthermore, I acknowledge the important financial support by a PhD studentship from Greek State Scholarship Foundation (IKY).

My wholehearted gratitude goes to my family: my parents and my brother for their calmness and 'quiet faith' in me, as well as their unconditional support during this process on my needs (or the needs of the PhD – difficult to tell one from the other at times), to the point where they would tell me "are you sure you want to come home for Christmas?" Somehow even if they didn't *really* understand what a PhD is like, I still think they did to a very large degree. This thesis is dedicated to them.

I am so grateful for the continuous, unconditional support of my long-time friends from Greece: Giorgos P., Evi, Fotini, Giorgos K., Takis, Panagiotis, Froso, Chrysi, Giorgos V., Christina and Lampros. Further, to everyone in the *Burn the Jukebox* collective, and especially Hugh, Jim, Bobby, James, Chris and Kathryn: thank you. You have given me so much, confirmed my faith in music and inspired new possibilities of transformative cultural interventions. Special thanks goes to Rasmus, my first ever housemate in York, and also Lizzie, Graeme, Rita and Louise. Angie has given me an incredible support during my studies – she has been a treasure and I thank her wholeheartedly for that. Also, I thank my housemates Georg, Ilaria, Lucas, Chloe, Priscille, James, Miranda, Achim and Bridget – I feel so lucky to have lived with them. I would especially like to thank Ilaria: she has probably been the person who witnessed most strongly the dramatic but positive psychological shift in me; her role was instrumental in that.

I also gratefully thank Cirihn, for sharing with me a large, transforming part of this journey, and for helping me to open up immensely. Finally, for the last few months, a warm thank you to Katerina.

In my view, engineering, as with any other endeavour of the human spirit, will not reach its true potential without a commitment to socio-economic justice and ecological sustainability. In a world where almost half of the population lives on less than $2.50 a day, this cannot be stressed enough. I am hoping that this thesis might be a very small step towards this aim.

Author's declaration

Except where references are made to other sources, the work presented in this thesis is the sole contribution of the author. Parts of the review presented in Ch. 4 previously appeared in:

G. Siamantas, M. R. Every, and J. E. Szymanski, "Separating sources from single-channel musical material: A review and future directions," in *Proc. DMRN Doctoral Research Conference 2006*, London, UK, 22-23 July 2006.

CHAPTER 1

Introduction

With the advent of the digital age and computers in the last decade of the 20th century the way we create, acquire and listen to music has changed rapidly. For example, the very idea of what is music has been put into question. Nowadays, if an artist intends to include any audible sound as a part of a piece [144], then this sound can automatically be considered as musical. This creative freedom leads naturally to the need for processing tools that will enhance expression through digital musical sound. These tools will, ideally, have to provide flexible ways of conveying, representing and manipulating information from musical signals.

Moreover, the link between music and technology has been greatly reinforced by the internet. It is now extremely easy for the internet user to listen to and/or download music or, generally, audio recordings. Ever-expanding databases of speech and audio recordings are being created, offering a wide variety of choices for every need. Also, a large number of today's music recordings are produced employing only digital means, and digital music players (such as the MP3 player) are now becoming as common as the mobile phone. For this reason, there is a growing demand for techniques that will enhance usability, fidelity and the ways that audio and music content are being archived, classified and delivered.

In this framework, physical science, psychology, music theory, computing and engineering have all contributed to the exploration of the possible ways that we can listen to and create music by:

- Increasing our understanding of the physical properties of sound itself and its production from physical musical instruments.
- Providing mathematical tools for representing and modelling sound in such a way that its musical properties are evident and that they provide many possibilities for content manipulation.
- Exploring the way that humans perceive and understand sounds.

Audio and, in particular, musical source separation is one of these multi-disciplinary areas of research that makes use of these findings in order to provide methods for enhancing the above applications. Musical recordings are in most cases polyphonic, *i.e.* containing audio coming simultaneously from different musical instruments or, more generally, *sources*[1]). Source separation techniques aim to estimate and, most often, extract and isolate from the recording the signals corresponding to each one of these sources.

1.1 Potential applications

If the audio source signals can be made available through separation algorithms, a number of advanced processing operations can be performed by operating on these individual signals, rather than the whole mixture. If the separation is carried out with sufficiently good results, the access to the extracted source signals provides more 'freedom' for manipulation and analysis. For example, it has been reported [5] that certain music and speech feature extraction methods for Music Information Retrieval (MIR) applications have reached a "glass ceiling" in terms of performance level. When these methods are applied to polyphonic segments, they most often use the original signal mixtures, which are typically a lot more

[1] For a discussion regarding polyphony and the concept of 'sources', see §4.1.4.

complex and challenging than their constituent source signals. Source separation could, thus, be a way forward from there, and a number of important applications (particularly related to music) could become possible or, rather, significantly enhanced.

Vincent *et al.* [162] have proposed a distinction between source separation applications according to whether the estimated and/or separated signals are supposed to be listened to or not: these are the *Audio Quality Oriented* (AQO) applications and the *Significance Oriented* (SO) applications, respectively. Examples of AQO applications are:

- **Upmixing and remixing.** This includes a variety of situations where (a) one or all of the sources are extracted from the mixture, or (b) one or more sources are suppressed beforehand. The upmixing is usually carried out in order for the sources to be remixed afterwards, after performing some kind of 'creative' manipulation: addition of audio effects, change of instrumentation or arrangement, phase and gain adjustment for creating artificial spatial source images (*e.g.*, mono to stereo [101], stereo to stereo [180] or stereo to multichannel [31]). Karaoke applications, where the suppression of main vocal melody in the mixture is required, can fall under the category of creative remixing (*e.g.*, [143]).
- **Restoration and denoising of mixtures.** At least during the last 20 years (and with the rise of the digital format), the need for higher quality audio has been increasing. At the same time, there is still a large amount of recorded material which, for nostalgic or historical/preservation purposes, needs to be restored appropriately before being transferred into digital format [65, 57, 62]. Denoising also includes enhancing the quality of hearing aids (*e.g.*, [126]) and noise reduction in communication devices.

In general, the AQO applications are expected to require extracted signals of a reasonably high quality, although this can be debatable to some degree. For example, in the case of remixing, the previously extracted signals are expected to be masked by others to some degree, obscuring some of the separation artefacts.

Representative examples of SO applications include:

- ***Automatic Music Transcription* (AMT).** The automatic extraction of the musical score corresponding to a piece of musical audio is often a highly desirable feature for an audio content processing system [96], a key part of which is often a multiple fundamental frequency (or *multiF0*) estimation stage [40]. Source separation can potentially simplify the complexity of this problem by allowing the AMT algorithms to be applied to the separated sources rather than on the complete mixture at once.
- **MIR.** How to automatically annotate a huge amount of highly varied musical material for the creation of internet-based or private databases and devise methods for effective retrieval of this information is a crucial challenge. Applications that fall under this category are, *e.g.*, query-by-humming, music recommendation, musical instrument recognition and genre classification.[2]
- **Audio compression.** Compression approaches that fall under the term *Structured Audio Coding* [160] and can potentially achieve much lower bit rates than the widely used MP3 (for example) with comparable perceptual quality, have started to emerge recently. If high-level parameters can be extracted from the audio (such as, *e.g.*, spectral envelopes, amplitudes of sinusoids), they can be used to represent the signal as a collection of "objects" (hence the related term, *Object-based Audio Coding*). What is encoded and transmitted are the parameters of those objects, which allows resynthesis at the receiver end (see, *e.g.*, [166, 36]). If it was possible for complex audio mixtures to be replaced with their constituent parts, it would be expected that the effectiveness of these coding strategies would be enhanced.

[2] The International Society for Music Information Retrieval (ISMIR) website [84] is a useful resource for MIR-related topics.

The next section describes the main challenges that the design of a separation system poses, outlines the context of the work in this thesis, and presents its basic contributions.

1.2 Context and thesis contributions

The fact that human beings are capable of distinguishing sounds, and are able to make sense of them despite the presence of others, has long been considered as an engineering problem in the form of the *cocktail party problem* [29, 72]. Thus, a very important and constant inspiration in the course of development of the area of source separation has been the exploration of the *Human Auditory System* (HAS) and, in particular, the field of *Auditory Scene Analysis* (ASA) – the way humans infer meaning from the auditory environment through perceptual organisation mechanisms. The book by Bregman of the same name [16], was seminal also by helping to lay the foundations for computational attempts to imitate the way humans 'separate' distinct sounds; in other words to help the field of *Computational Auditory Scene Analysis* (CASA) emerge.

Around the same time as the first CASA systems were proposed (the early 1990's), a parallel line of work saw the problem of source separation from a more mathematically rigourous point of view. No matter what the nature of the source signals was (*i.e.*, they did not particularly concentrate on audio), as long as the sources satisfied a number of statistical assumptions, techniques such as *Independent Component Analysis* (ICA) could be applied to them. This group of work usually falls under the term *Blind Source Separation* (BSS). The word 'blind' is used to signify that the prior information used is minimal.

The prior information is one of the most important factors that has to be taken into account when designing a source separation system. Even if we disregard any additional factors, single-channel mixtures pose considerable challenges for source separation approaches just because of the sheer lack of initial information – in a perfectly blind approach only a single version of the signal mixture is available. Unless some kind of appropriate information is employed *a priori*, this is a problem with an infinite number of solutions. Prior information can take the form of application-specific assumptions and models for the source signals and the mixing process. These assumptions and models can be incorporated directly into the processing algorithm and/or inserted within it beforehand by the user or through learning/training procedures. In contrast to the blind approaches, others that employ some additional, sufficiently generic information (such as, instrument models or psychoacoustic cues) can be called *semi-blind*, while the ones involving user-inserted information (such as, *e.g.*, the MIDI score of the mixture), *non-blind.*

1.2.1 Between 'understanding' and 'separation'

The degree of blindness in separation methods is inextricably tied to their respective applications. For example, the outputs of some SO applications (such as the estimates of F0 contours, or the note onset timings) can be used as prior information for a separation system because they provide assistance with the identification of source structures. A further insight in the relationship between prior information, separation and intended application can be acquired if the separation methods are described with the use of the combination of two terms: 'understanding' and 'separation'.

Scheirer [145] introduced the term *Understanding Without Separation* (UWS), to emphasise that his work on constructing music listening systems did not require the separation of sources beforehand. The word 'understanding' is used to describe a number of SO applications of source separation, such as AMT and other MIR-related ones. As mentioned, these applications do not usually require explicit source separation; however, as is argued here and by others, separation could lead to their improvement, if employed appropriately. All the separation methods which can potentially aid SO applications can be described as *Separation For Understanding* (SFU) methods. This different kind of taxonomy is completed by referring to the *Understanding For Separation* (UFS) methods (these include, ba-

Primary processing goal	No connection	Monodirectional connection	Bidirectional connection
Separation	SWU	UFS	SAU (UFSFUFS…)
Understanding	UWS	SFU	

Table 1.1: A taxonomy of audio processing systems, where the 'understanding' and 'separation' elements and the ways that are connected to each other are highlighted. The method proposed in this thesis belongs to the 3rd column, where UFSFUFS… is an alternative to the SAU term, which gives emphasis on the iterative process between SFU and UFS. (See text for an explanation of the acronyms.)

sically, the semi-blind and non-blind ones) and *Separation Without Understanding* (SWU) methods (these are the blind methods).[3]

The approach presented here attempts to close the circle between understanding and separation by making their connection operate in both ways (*i.e.*, bidirectionally), so that the one feeds the other. This can be realised through an iterative process between UFS and SFU. It can, thus, be characterised as a *Separation And Understanding* (SAU) method (see Table 1.1[4]).

More specifically, the work concentrates on the design of a system that would be able to extract the source signals from a musical mixture without necessarily trying to emulate human-inspired music scene analysis. In other words, a number of characteristics of the HAS capabilities are taken into account (those concerned with the nature of the musical sounds and mixtures), while others which could be limiting for SO applications (such as the fact that human listeners show difficulty recognising the correct number of sources if there are more than three sources [81][5]) are not.

To do this, a shift is made, first, from a non-blind to a semi-blind system. Since semi-blindness implies the use of advanced generic models – in the sense that a blind system would not, in theory, use any explicit signal model, while a non-blind system would use very restricting models – it provides the flexibility required for

[3] This taxonomy was first introduced by Burred [23].

[4] An extended classification framework (also involving the separation/understanding relationship) particularly for separation systems is presented in §4.3.2.

[5] Huron uses the term 'voice' to refer to what is essentially called a 'source' in this thesis. (See also §4.1.4.)

an SAU system. The harmonicity psychoacoustic cue, along with the assumption that the source signals contain music events with well localised onsets and offsets are considered to satisfy the need for generality. Harmonicity is applied by using an automatic multiF0 estimator. This is combined with a particular source extraction method that can deliver (apart from the sources) a residual signal which will contain all the unmodelled content. It is then shown that the use of the residual channel in an iterative framework can improve the robustness of both the multiF0 estimation and the separation.

Specifically, the key contributions of the work presented in this thesis are summarised below:

- Removal of the need for significant user input (shift from a non-blind to a semi-blind system) by replacing the score-informed front-end of an existing separation system with a more automatic, while still accurate, alternative that is based on a multiF0 estimation algorithm.
- Proposal of a method that carries out multiF0 error correction as part of a F0 track disentangling stage.
- Establishment of the residual channel as a central concept in the proposed iterative framework for single-channel separation and multiF0 estimation. Confirmation that the iterative use of the residual can improve the robustness of both processes.
- Comparative analysis and discussion on the effectiveness of two widely used separation performance evaluation measures, using both a theoretical and a practical framework.
- Use of the proposed system to improve the performance of specific applications, such as note onset detection. To be specific, a novel note onset detection algorithm that operates on the residual channel is introduced.

It is worth mentioning that, alongside the above contributions, there are also a number of subsidiary ones. These have to do with a number of modifications that led to improvement in the harmonic parameter estimation stage of the existing separation system, and the extension of the frame-based multiF0 estimation process to a note-based estimation one through the F0 track disentangling process.

1.3 Thesis overview

In Ch. 2 an overview is given of the basic concepts related to the analysis of musical audio signals. The nature of musical audio is discussed and then some important ideas from music theory, auditory perception and cognition are introduced, together with a brief introduction on sound spectra.

Ch. 3 discusses a number of different transformations and models that are frequently used for representing musical audio signals. In particular, the principles of additive modelling, along with parametric and nonparametric methods, are briefly discussed.

Ch. 4 gives an introduction to the problem of single-channel source separation. The factors that define the complexity and the challenges of the problem are discussed, and a thorough review of the various ways it has been approached so far with regards to musical audio is carried out. The chapter continues by referring to a few of the measures most usually employed for analysing the performance of source separation systems.

Furthermore, the proposed approach for an unsupervised system for single-channel source separation is presented in Ch. 5. After giving an overview of the complete system and going through a few additional definitions, the chapter continues by explaining the different stages of its basic one-way infrastructure. The differences, modifications and improvements compared to an existing non-blind system, are shown. This includes a detailed account on the choice and adaptation of the multiF0 estimation algorithm, along with a description of the supplementary stage of F0 track disentangling which provides a method for an initial multiF0 error correction. Next, a performance comparison is carried out between the one-pass proposed approach and its non-blind version for a variety of mixtures.

Additionally, in order to choose the appropriate means for analysing separation performance, the same section discusses the effectiveness of two widely used measures by comparing them using a theoretical and a practical framework.

The idea of the residual channel is introduced next in Ch. 6, along with an exploration of the multiple functions that it can fulfil. This is followed by the proposal for the iterative framework that uses the residual. Evaluation results are presented that show improvement of both separation and multiF0 estimation when the feedback loop is incorporated in the system and, in addition, a performance that is better on average in comparison with other alternative methods. Lastly, a novel, residual-based note onset detection algorithm is introduced, and a brief proposal for an extension to stereo mixtures is made. In particular, promising results are shown through performance comparisons of the onset detection algorithm with two other methods, one of which could be considered as current state-of-the-art. The audio results of the source separation experiments can be listened to on the web at [150].

Finally, Ch. 7 gives a summary of all the material presented in the thesis, and a structured outline for future work is proposed.

CHAPTER 2

Analysis of musical audio signals

The separation approach that is the central part of this thesis is designed to be applied to musical sounds. This chapter reviews some important concepts around the analysis of musical sounds, which will help in clarifying the decisions behind the design of the particular system. Firstly, a number of definitions are made, paying particular attention to discussing the musical character of sounds. This is followed by an overview of musical theoretic concepts, the physical attributes of musical sounds and how these are perceived, understood and mentally organised by humans.

2.1 Defining the character of musical audio

We begin this section by emphasising first two statements that will be explained below:

→ Not all musical signals are audio signals.

→ Not all audio signals are musical signals.

The first statement can be explained by the following definition, proposed here, of a musical signal: *The representation of a varying quantity or any other medium*

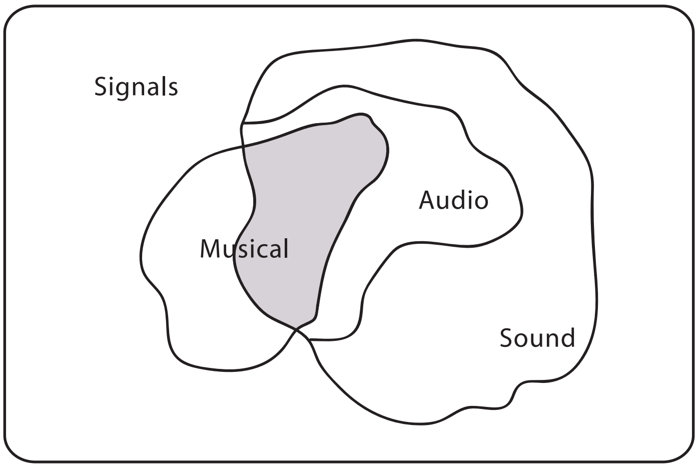


Figure 2.1: The relations between sound, musical and audio signals.

*that can carry information (*any *kind of* [1]*), when this information is a function of time and it has a* musical nature *(it can be interpreted using musical terms).* According to this definition, a series of musical notes appearing on the staff or even the sequence of letters G, E, and D written on paper (representing a melody made of the succession of musical chords G, E and D, respectively) could well be considered as musical signals.

Now, a particular case of musical signals are musical *audio* signals. These signals represent vibrations of physical media (*i.e. sounds* [140]) that are within the range of the *audible* frequencies, and hence they can be perceived as sound waves by the human listener. In Fig. 2.1 the relations of sound, audio and musical signals are depicted graphically with the help of a simple Venn diagram.[2] It is worth noting, also, that the above definition helps to distinguish between the kind of analysis that will be dealt with here, from traditional *musical analysis* (that is, primarily the analysis of non-audio musical signals such as the score of musical notes and timings).

[1] Although the term 'information' can have different meanings depending on the context, it is safe to attempt a general definition for the context of this research: assuming the existence of a general type of a communication system (transmitter-channel-receiver), we regard information as *any sort of* knowledge *that possesses a specific meaning and importance, because it acts by* adding *to the knowledge of the receiver.*

[2] Hereafter and for the sake of simplicity, the words *sound* and *audio* will be used interchangeably. The same will hold for the terms *musical signals* and *musical audio signals.*

The second statement refers to the fact that arguments for distinguishing musical from non-musical sound could easily end up trying to distinguish music and non-music. Clearly, this is not a way to tackle this, since after the birth of electronic and computer music *any* sort of audible vibration can exist (or be *allowed* to exist) in a musical piece. So, although defining music is still nowadays a difficult but interesting problem, there exists a relatively accepted idea (at least among researchers) about which sounds are considered musical and which are not, something that we will try to show through this thesis. This is an interesting paradox within contemporary music: in its context, any kind of sound can exist, no matter if it can be considered to be musical or not.

One obvious and important remark that we could make is that *music is the sound coming from one or more musical acoustic instruments (including the human voice).* This is definitely true but it does not constitute a complete definition of musical sound. For example, what about sounds generated by other sources, or artificially made ones? There are certainly sounds of this sort that do not resemble common instrument sounds, although they still retain a musical character. Hence, rather than searching for a clear definition of the 'musicality' of a piece of sound, it is more appropriate to *understand* well known musical sounds/signals.

Towards understanding musical signals, researchers have devised suitable *representations* or *models* by using a variety of analytical tools. These models help to describe the sound wave by employing terms that relate to perceptual properties such as pitch, loudness and timbre, even though some of these measures are really only intended to describe the properties of an isolated instrument, and may be of limited value when applied to a typical mixture. Furthermore, other models relate music theoretic concepts such as harmony or melody to mental structures in the brain. So, research has been progressing along two main parallel paths that, quite frequently, happen to merge together:

- Investigating the physical aspects of traditional instrument sounds and what is involved during their *generation.* This is what the areas of physics, mathematics, computing and engineering have been contributing to.

- Finding out how humans *perceive* sounds and their musical features, and infer meaning from higher-level musical structures, and finding correspondences of these perceptual attributes with the physical properties of the sound. Music and its evolution, in general, has relied on the capabilities and limitations of the human auditory system (HAS) and if the HAS cannot tell the difference in pitch or timbre between two sounds, it does not matter musically.[3] This what the areas of music theory, psychology, cognitive psychology and neuroscience have been contributing to.

The following sections will examine briefly the ways in which these areas have contributed to the understanding and analysis of musical signals. In particular, we will focus here more on those concepts which are most relevant to the discussions later in this report. Unfortunately, due to the multidisciplinary nature of these areas, it is difficult to avoid using some terms before explaining them fully.

Fig. 2.2 outlines the relationships between the main concepts that are discussed in this chapter. The conceptual formulation carried out by Scheirer in his thesis [145] (as well as some of the associated terminology) was employed here as a useful starting point. Vibrating sources (the *auditory objects*, or *sound objects*) produce sounds that can be characterised as *auditory events*, which all together are presented to the listener's ear as a single *auditory stimulus.* In order to reach the listener, though, the sound often goes through a number of 'communication' processes: the audible part is transduced by microphones into electrical signals (audio signals) and then it may be digitised, compressed, coded, *etc.* before it is converted back to sound vibrations. From the listener's side, this complex stimulus is believed to be disentangled into *auditory images* [145]. An auditory image is the perception of a sound as coming from a single source. The rest of the figure represents the highest-level mental classification of these auditory images into

[3] Here we refer, mainly, to the limitations of the HAS in perceiving melodies (see §2.2.1) of isolated sounds and distinguishing differences between these sounds, based on their timbral features: there are certain limits beyond which we cannot observe pitch fluctuations in time when we listen to a single sound; the same holds for observing differences between two sounds. It is important to note, though, that if two sounds with the same (or very similar) pitch and timbral properties are played *simultaneously*, there will be harmonic reinforcement and amplitude changes, which will still be perceived (corresponding, for example, to the sense of harmony).

musical and non-musical, including some other concepts related to the perception of sound that will be mentioned throughout this thesis.

2.2 Musical properties and perception

2.2.1 Music theoretic concepts

This research work, in line with the majority of research in music signal analysis and understanding, will concentrate on musical pieces based on Western musical theory. Although many other systems of music exist, the vast majority of popular and classical music content is based on the Western musical tradition. According to this tradition the musical works are structured around elementary elements, which are *notes*. These are symbols for representing a number of the salient characteristics of the sound: primarily its pitch and duration, and to a lesser degree, its intensity, timbre and tempo [121]. *Pitch* is a perceptual attribute which allows us to order sounds from high to low on a frequency scale. A more exact definition would be '*the frequency of a sine wave that is matched to the target sound by human listeners*' [71]. For ideally harmonic or near-harmonic sounds the pitch is definite and is normally equivalent to the *fundamental frequency* (F0) of these sounds. This is assumed when we talk about a note's pitch, or the *musical pitch*. The notes are often arranged in the *equal-tempered* tuning scale that divides each octave into 12 logarithmically spaced semitones. Each note's pitch can be calculated as $440 \times 2^{n/12}$ Hz where n varies from -48 to 39 on a standard piano keyboard [94]. A *melody* is the sequence of notes in time and can itself be thought of as a single entity. On the other hand, a *chord* is a set of notes sounding simultaneously. Chords can be *consonant* or *dissonant*. These last terms are perceptual attributes of chords related to the field of music theory called *harmony*. A common explanation of consonance (originally proposed by L.M.F. Helmholtz [74]) is the fact that the combined notes have *shared harmonics* (see §2.2.3). This leads to a fusion effect that makes consonant chords sound pleasing to the listener, with the dissonant ones having the opposite effect. As a

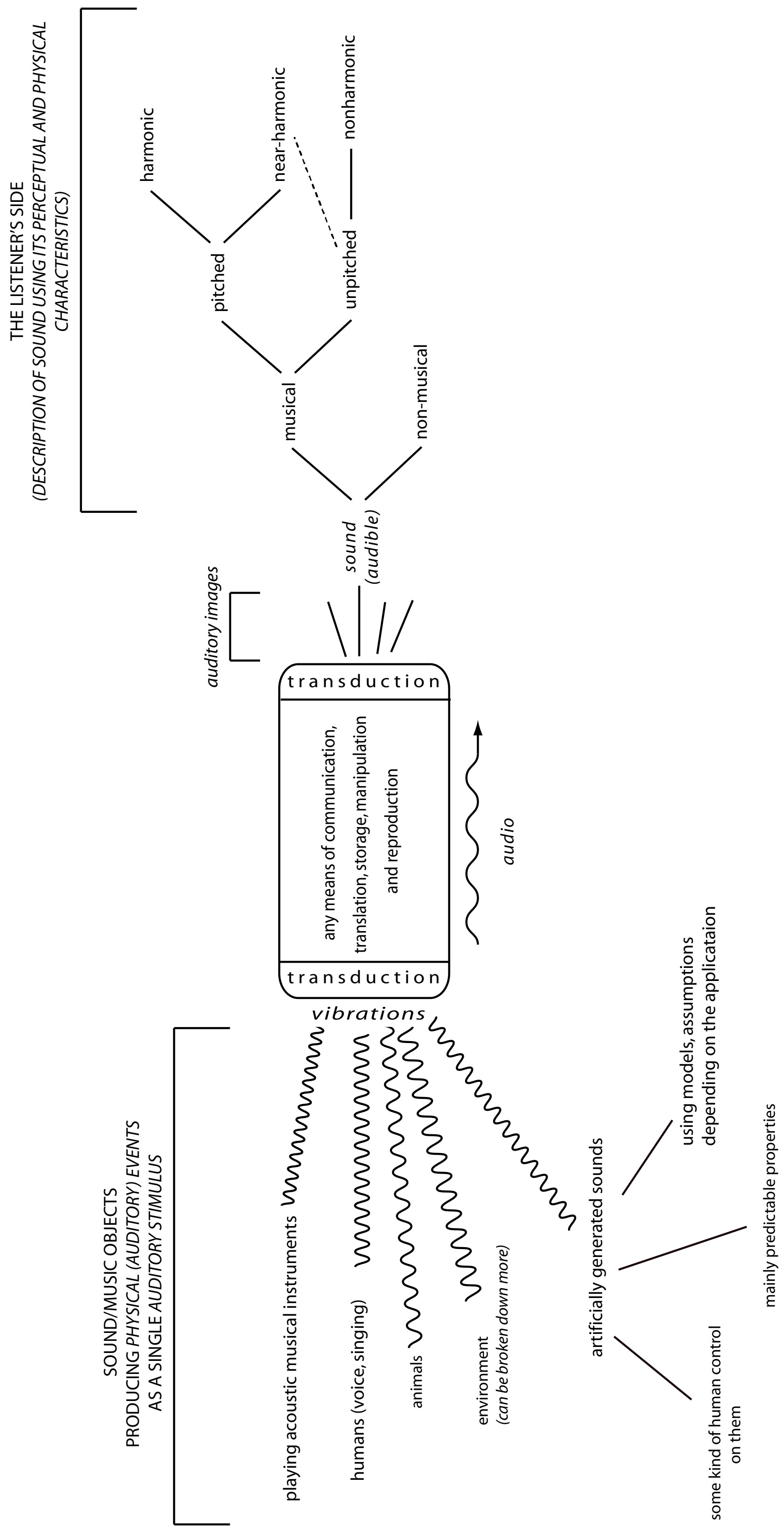

THE LISTENER'S SIDE
(DESCRIPTION OF SOUND USING ITS PERCEPTUAL AND PHYSICAL CHARACTERISTICS)
harmonic
near-harmonic
nonharmonic
pitched
unpitched
musical
non-musical
sound
(audible)
auditory images
transduction
any means of communication, translation, storage, manipulation and reproduction
audio
transduction
vibrations
SOUND/MUSIC OBJECTS
PRODUCING PHYSICAL (AUDITORY) EVENTS
AS A SINGLE AUDITORY STIMULUS
playing acoustic musical instruments
humans (voice, singing)
animals
environment
(can be broken down more)
artificially generated sounds
using models, assumptions depending on the applicataion
mainly predictable properties
some kind of human control on them

Figure 2.2: From physical auditory events (vibrations) to a classification of sounds in terms of their physical properties and their perception by humans.

result, consonant intervals are favoured, as opposed to dissonant ones. We will see later on that this very common characteristic of musical sounds, that they are 'fused' together when they have shared harmonics and played simultaneously, is an important one when we want to separate them.

2.2.2 Limitations of the auditory system

The HAS is generally sensitive to sounds within the range from 20 Hz to around 20 kHz. However, frequencies up to 10 kHz are more significant than the rest of the audible spectrum. Due to the particular way the sound is transduced into neural impulses in the *basilar membrane*, it is hard to distinguish two different notes when played simultaneously and their frequencies are separated by less than a certain frequency range called the *critical band*. The range of this band, called the *critical bandwidth*, has a constant value for centre frequencies up until 1 kHz. After 1 kHz the critical bandwidth increases *logarithmically* with the frequency. So, for example, at frequencies 100 Hz and 200 Hz the critical bandwidth is 90 Hz, while for a frequency of 5 kHz it increases to 700 Hz [140]. Another important parameter (which appears to be the result of the same mechanism in the ear that is responsible for the critical band) is the Just-Noticeable Difference (JND) with regards to pitch: this is the smallest frequency change that has to take place in a tone in order for it to become noticeable by the average ear. The JND depends on the frequency, sound level, duration and suddenness of the frequency change, and corresponds roughly to 1/30 of a critical band [140]. This means that while it is possible to detect very small differences in the frequency of a single tone, a much larger frequency difference is needed to discern between two tones when they are played simultaneously.

Some additional important limitations of the HAS are related to time. Experiments have shown, for example, that in order to tell reliably the order of the onsets of two 0.5 s tones (a duration appropriate for a short musical sound segment), they have to be at least 20 ms apart [129]. These properties of the auditory system are exploited for creating reliable signal representations (as we will see in Ch. 3) as well as for effective compression and coding algorithms.

2.2.3 Sound spectra

The HAS (as well as the auditory systems of all mammalian animals) in effect performs a kind of analysis of the sound similar to that of 'Fourier analysis'. Musical sounds, as with all sounds, can be represented as the sum of a number of different sine waves (also called *tones*, or *sinusoids*). The amplitudes of these sinusoids characterise the sound's *spectrum*, which is, generally, changing over time. This idea was inspired by J.B.J. Fourier who first showed that periodic functions can be decomposed into a sum of sinusoidal components (the *Fourier series*). Periodic sounds, hence, are represented by a number of discrete sinusoidal components whose frequency and amplitude remain constant over time (theoretically, for ever). These sinusoidal components are the *harmonics* (integer multiples) or *harmonic partials* of a F0. Thus, if m is the partial index, the harmonic frequencies will be:

$$f_m = m\,\mathrm{F0}, \qquad m = 1, 2, \ldots. \tag{2.1}$$

The components that are placed in non-integer multiples of the F0 are called *inharmonic partials*, and these partials dominate in *aperiodic* or *nonharmonic* sounds.

Of course, sounds coming from real acoustic instruments are not strictly periodic and they don't have infinite duration. As a consequence, they cannot be decomposed using the Fourier series. However, for time-limited relatively stationary portions of the sound, that can be characterised as *quasi-periodic*, other tools exist for analysing its frequency content (see Ch. 3). The spectral analysis of these sounds shows many more frequency components than just the expected harmonics, although the harmonics still remain the dominant components. This is what we usually call a *harmonic* sound.[4] In general, most Western non-percussive mu-

[4] In real musical sounds the harmonics may not appear in exact integer multiples, as will be shown below. For this case, where the sinusoidal components are placed in *nearly* integer multiples of the F0, the term 'harmonic' will still be used for characterising the sound.

sical instruments produce harmonic sounds (at least during the sustained part of the sound). We have to make two notes, though:

- Their spectra are not always perfectly harmonic.

This is evident for plucked and struck string instruments (a common example is the piano) for which the partials deviate to a certain degree from perfect harmonicity. In fact, they are placed according to the approximate formula[5]

$$f_m \;=\; m\,\mathrm{F0}\sqrt{1+B\,(m^2-1)}, \tag{2.2}$$

where $m = 1, 2, \ldots$ is the partial index and B the *inharmonicity coefficient* [63].

- Most of the harmonic instruments produce sounds also with both *transient* and *noise* content. This is due both to the way an instrument is played (*e.g.* repeated plucking of strings) and nonlinearities introduced by the instrument's body.

The transients occur during note *attacks* (the beginning of a note) and *decays* (the end of a note), in other words, during the non-steady part of the note. They have a very rich spectral content, almost noise-like[6] at the attacks, in contrast to the steady-state portion of the note (*sustain*) where the harmonics dominate. However, transients are not entirely random signals in that their sound *does* reveal a sense of structure, although this is difficult to model effectively. Finally, by 'noise content' we mean any other residual content in the sound that does not reveal any sort of clear spectral structure (an example is the 'breathiness' of a flute). To conclude, these features which are responsible for the nonharmonic content of the sound play an important role in characterising its 'naturalness'.

[5] It is important to note that, in theory, *all* the partials deviate from the positions predicted by perfect harmonicity, even the fundamental. So, in this formula F0 is really the *measured* value of the fundamental frequency, rather than the predicted one. Practically though, the difference between the two is often not perceptually significant. This is why the majority of authors do not make this distinction, assuming no deviation of the fundamental from its predicted value.

[6] A 'noisy' signal's spectrum is a distribution of frequency components that resembles the white or coloured noise spectrum.

The rest of the chapter will highlight how the physical properties of sounds are believed to be perceived and understood by a human being, and how they are used for building high-level cognitive structures.

2.2.4 Basic perceptual attributes of sound

The perception of sounds by humans is thought to be carried out by employing a number of basic *perceptual attributes*. These attributes are part of a *subjective* experience, as opposed to physical phenomena, which are objective and can be described in terms of quantifiable parameters [140]. The most commonly-studied perceived attributes are pitch, loudness, timbre and duration. These attributes depend in some ways on a set of well known physical parameters such as frequency, pressure, spectrum, envelope and duration. The parameters could be seen as dimensions defining a space on which the perceptual attributes could be projected. Table 2.1 (taken from [140]) shows the 'dimensionality' of each of the attributes. What we can see from this table is that all the attributes depend to a certain degree on all the physical parameters. Apart from timbre, though, the rest of them can be broadly characterised as 'one-dimensional' (*e.g.* pitch depending mostly on frequency and loudness mostly on pressure). Timbre, on the other hand, is a multi-dimensional attribute, and thus difficult to describe and quantify. We will continue, here, by focusing only on the attributes of pitch and timbre, which are considered important to this research. For a discussion on the perceptual attributes of loudness and duration, the reader is referred to key texts such as [46] and [140].

Pitch

As mentioned in the previous section, there are sounds for which a person can make a definite decision about how high or low in a frequency scale they are, compared to other sounds. This decision is based on the sense of a perceptual attribute called *pitch*. Quantifying this attribute and relating it to measurable properties of the sound, though, is not necessarily straightforward when we are

	Subjective Quality			
Physical Parameter	Loudness	Pitch	Timbre	Duration
Pressure	+++	+	+	+
Frequency	+	+++	++	+
Spectrum	+	+	+++	+
Duration	+	+	+	+++
Envelope	+	+	++	+

Table 2.1: The dependence of subjective qualities of sound on physical parameters (from [140]). The quantity of '+'s indicates the degree of dependence, from weak dependence ('+') to strong dependence ('+++').

dealing with very complex signals. The brain is capable of assigning pitch even to sounds composed just of inharmonic partials or some types of wideband noise [140, Ch. 7]. As shown in Fig. 2.2 there is no clear boundary between unpitched and near-harmonic (pitched) sounds. Although a tremendous amount of work has been published representing attempts at finding a reliable pitch model, none of them has been able to deal with all the different cases in matching humans' decisions regarding complex sounds (for a review see [71]). We are not going into deep analysis here, instead we will present some results that have been extensively verified and can hence be applied with confidence to musical instrument sounds.

If the target sound is another sinusoid or a periodic sound, the pitch would almost definitely match the frequency of that sinusoid or the F0 of the sound, respectively, as experiments show. However, the F0 does not necessarily have to be present. There are cases, for example, where the F0 or the first few partials are absent from a signal, and yet the sense of pitch evoked to the listener corresponds to the (missing) F0 [146]. Examples of such instrumental sounds include that of the bassoon or the organ, when playing very low notes. This has become known as *residue pitch* or *virtual pitch* (a term coined by Terhardt [157]). What is known is that the pitch is determined by the most prominent harmonics. The prominence of the harmonics depends on the frequency range: for high pitches it is the lower harmonics which have greater effect, while for low pitches the higher harmonics are more important. Generally, it can safely be said that the perception of pitch depends on the position and amplitude of the lowest six harmonics in some fashion [129]. Finally, another important observation concerns sounds in which the odd harmonics are dominant: the sense of pitch in this case may not correspond to the F0 [128].

Timbre

Various definitions have been presented in the literature for the perceptual attribute of timbre. Following Scheirer [145] a definition with a broad sense will be adopted: “[The timbre of a sound is] the quality or set of qualities that allows a listener to identify the physical source of a sound.” As Plomp [130] notes though, this is a ‘negative’ description, *i.e.* it states that timbre is neither pitch nor loudness, but does not give any more information about it. Which are these qualities, then, that make a sound distinguishable from one other sound? Table 2.1 shows clearly that these qualities depend considerably on more than one physical property, such as the spectrum, the envelope and the frequency. *How* exactly these dependencies are formed is the difficult part.

Firstly, it is generally true that the ear is largely insensitive to phase alterations. By ‘phase’ we refer here to phase relationships between harmonic partials of periodic tones [137]. This can be observed, for example, from the fact that reverberant environments (generally causing large phase changes to the sound) do not appear to alter the sound’s perceived timbre. Thus, the waveform shape is not the only factor affecting timbre.

However, it does appear that the timbre is greatly affected by how the sound energy is distributed among the partials: in other words, the spectral shape. One interesting point, however, is that we are able to distinguish the difference between the words ‘we’ and ‘you’ when we hear them, although it can be shown that they have approximately the same spectrum [129][7]. Thus, the importance of *transitions* (time domain variations) also has to be acknowledged. Moreover, by filtering we can change the ‘colour’ of a sound (make it brighter, or more dull), but

[7] This could be also shown (with regards to the magnitude spectrum) by listening to a piano note played backwards in time: though the magnitude spectrum would be the same, it may not be recognised as a piano sound.

we can still easily recognise it. These facts raise questions about the significance of purely spectral information in timbre discrimination.

Among other approaches towards understanding timbre perception, some have employed sound analysis and synthesis [137]. According to this process, a suitable model or representation (see Ch. 3) is chosen for the sound in question, and its parameters are calculated. These parameters are used to create a synthetic version of this sound. By elaborate auditory comparisons between the original and the synthesized sound, useful insight can be gained into the relevance of a sound's physical parameters to the resulting timbre.

Finally, it is worth noting that instrument recognition and classification [76, 75] is a field of research closely related to timbre discrimination. Work on this field tries to find sets of quantitative features that can best describe timbre and then use them to build algorithms that will automatically recognise or classify musical sounds into different classes of instruments. Examples of features may include harmonic irregularity, vibrato, Amplitude Modulation (AM) frequency, spectral centroid, or the zero-crossing rate of the waveform [127].

2.2.5 Grouping and segmentation mechanisms

The human brain has the ability to recognise or build structures out of highly complex musical sounds. A usual case in music is when we are 'hearing out' a melody played from a certain instrument in a mixture. This is caused by some kind of mechanism that deconstructs the sound into *auditory streams* (or *perceptual units*). The tendency to form perceptual organisations from sound is innate, and of course this does not hold only for music. Humans try to infer meaning from their overall sound environment; in other words, they perform an *auditory scene analysis* [16].

The basic models of grouping mechanisms were proposed by the *Gestalt* psychologists [16, p. 18]. According to them, we can group elements together by employing a number of simple rules, or *cues*. Examples of these cues are proximity, common fate, similarity and continuity. Especially for musical sounds, two of the most important cues are harmonicity and common onset. Grouping by harmonicity can be seen in pitch perception: what we perceive as the pitch of a sound, is due in part to the grouping of those sinusoids with harmonic or near-harmonic relationships.[8] According to the common onset cue, when a set of frequency components start simultaneously, it is likely to have originated from the same source. Some other cues used in musical signals are common Frequency Modulation (FM) or AM. AM is not a very popular cue, though, since an instrument's harmonics often evolve in a different way and decay in different times [45].

If we consider grouping in terms of notes (and not sinusoids), the same Gestalt principles can be applied. In this case, two main dimensions of grouping are generally encountered: *horizontal grouping* (or *sequential integration*) and *vertical grouping* (or *simultaneous integration*). Horizontal grouping is responsible for the perception of melody, while vertical grouping is responsible for the perception of harmony.

All these cues that have been found to be good models for the brain's grouping mechanisms appear to be very useful in musical signal processing. By including computational versions of them within the analysis framework of a musical signal the process of separation of musical structures can be enhanced.

2.3 Summary

This chapter presented the basic principles of musical audio analysis and introduced a number of definitions, outlining the broad framework of this research work. In order to analyse musical audio it is important to understand the physical properties of sound, its perception by the HAS and how sound is employed to make music, according to the Western musical tradition. So firstly, a number of basic music theoretic concepts are mentioned. Next, some important limitations of the HAS (related to its time and frequency resolution) are discussed, since

[8] As it was mentioned in §2.2.4, perfect harmonicity is not required for the perception of pitch.

these limitations can help to define useful musical audio characteristics. This is followed by briefly describing how the spectrum of the sound provides important information that can be used for further analysis. Depending on their spectral properties, musical sounds contain harmonic or near-harmonic frequency components and nonharmonic (transient or noise) content. Although this classification is not always straightforward, identifying these different types of content in a musical signal is valuable: analysis methods can be more effective when they are designed specifically for a certain type of content.

Furthermore, we discussed research findings related to basic perceptual attributes, which are thought to be employed by the HAS for perceiving sounds. In particular, pitch and timbre are mentioned here. Quantifying these attributes is not something straightforward, since they depend on more than one physical parameter. Finally, a number of auditory cues were presented, which are believed to be employed by the brain in order to recognise structures out of highly complex sounds. For musical signals consisting of notes, common structures recognised by the brain are melodies (largely due to sequential integration of frequency components) or chords (largely due to simultaneous integration of frequency components). The next chapter will discuss the ways in which musical signal structures can be represented so that they can be processed by computers.

CHAPTER 3

Representations of musical audio signals

This chapter describes some of the important contributions of signal processing techniques for representing musical audio signals. The audio waveform is considered to be a *low level* source of information on a scale of abstraction, as opposed to the high-level structures of information encountered in music theory (*e.g.* notes, melodies, chords, motifs). This hierarchy, of course, is an artificial one, corresponding to a human point of view,[1] because for computers there is nothing abstract (in the sense of a high-level representation) about a series of numbers (*i.e.* a digital signal). On the other hand, a sequence of 44100 numbers[2] does not mean anything to humans, unless they can relate it to some kind of 'real-world' information. This situation (which is depicted in Fig. 3.1) is directly related to one of the purposes of musical signal processing: we want to make machines understand musical information and structures the way we do, so that with their help we can solve problems like source separation. But this has to be done through abstract representations of the waveform which best describe the underlying properties of these signals in the way that corresponds to *our* perception of these properties.

All the information that we need is hidden in the time signal (waveform amplitude vs. time). As we have seen so far, though, our auditory system performs an

[1] or, more correctly, point of *hearing*!

[2] This many numbers can represent one second of digital audio signal, sampled at a rate of 44.1 kHz.

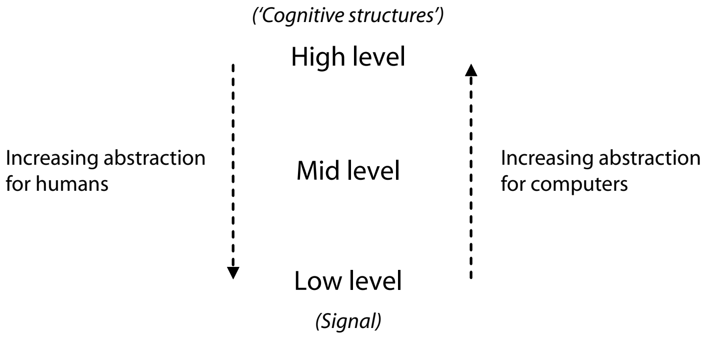

('Cognitive structures')
High level
Increasing abstraction for humans
Mid level
Increasing abstraction for computers
Low level
(Signal)

Figure 3.1: Levels of representation for humans and computers. The human world and its constructed ideas is abstract to the computer, and vice versa.

approximate Fourier analysis and then makes an effective use of the spectral information in order to build mental representations of the sound environment. Hence, it makes sense to try to find representations that will 'translate' the time signal into its corresponding spectrum. Moreover, since our goal is to separate musical structures, it makes sense to try to *project* them into a domain where these structures are evident, *i.e.*, the frequency domain.[3] These representations are called *time-frequency (TF) representations* or *distributions*, and they fall into the category of *mid-level representations.* This last concept, originally related to computer vision [111], appears to be equally useful for devising computational models of auditory perception [16, 56]. They form a representational area that can be 'located' between the low level (basically the waveform before it reaches the cochlea) and the high level (cognitive processes in the brain, related to the recognition of events or objects) in human auditory perception (see Fig. 3.1) and they are usually grouped into *parametric* and *non-parametric methods.* Although the division between these two classes can often be ill-defined, the general situation is as follows:

Non-parametric methods These approaches are usually based on signal transformations between the time and frequency domains and they do not require any assumptions or prior information about the signal. In other words, they do not offer an *interpretation* of the representation [113].

[3] Of course, source separation approaches that operate in the time domain do exist (*e.g.*, [77, 86, 14]). However, emphasis is given here on the ability of mid-level representations to highlight signal structures in ways that bear similarity to the way the HAS operates.

Parametric methods These approaches construct *models* by parameterising physically meaningful features of the signal. This means that some kind of prior knowledge is implied about the signal under question and thus, they implicitly offer some kind of interpretation.

Before starting with the description of the methods, an introduction to signal expansions, a concept that encompasses both non-parametric and parametric methods, will be given.

3.1 Signal expansion

A signal expansion is basically an additive model that represents a signal as the linear weighted sum of basic components. In its most general form, an observed discrete-time signal $x(n)$ can be expressed as:

$$x(n) \;=\; \sum_{k=1}^{K} g_k \, b_k(n), \quad \forall n \in \mathbb{Z}, \tag{3.1}$$

where $\{b_k(n)\}_{k=1}^{K}$ is the set of basic components called *expansion functions* which are summed using the set of linear weighting coefficients $\{g_k\}_{k=1}^{K}$. The family of the selected expansion functions is called a *dictionary.* If T is the length of x in samples, Eq. 3.1 can be written in vector notation as

$$\mathbf{x} \;=\; \sum_{k=1}^{K} g_k \, \mathbf{b}_k, \tag{3.2}$$

where $\mathbf{x} = [x(0)\, x(1) \,\ldots\, x(T-1)]^{\mathsf{T}}$ and $\mathbf{b} = [b(0)\, b(1) \,\ldots\, b(T-1)]^{\mathsf{T}}$. If we now organise the expansion functions in a $T \times K$ matrix $\mathbf{B} = [\mathbf{b}_1 \; \mathbf{b}_2 \; \ldots \; \mathbf{b}_K]$ and the coefficients in a vector $\mathbf{g} = [g_1 \; g_2 \; \ldots \; g_K]^{\mathsf{T}}$, we end up with:

$$\mathbf{x} \;=\; \mathbf{B}\,\mathbf{g}. \tag{3.3}$$

Eq. 3.3 represents a linear system of equations, where the expansion coefficient vector $\mathbf{g}$ is the unknown. When $T = K$ and $\{\mathbf{b}_k\}_{k=1}^{K}$ are linearly independent, the expansion functions are called *basis functions* and the transformation (or, in other words, the representation) is said to be *complete* [67].[4] A complete representation is invertible, so the expansion coefficients can be calculated using:

$$\mathbf{g} \;=\; \mathbf{B}^{-1}\,\mathbf{x}. \tag{3.4}$$

In this particular context, Eq. 3.4 is called the *analysis equation*, while Eq. 3.3 is the *synthesis equation*. A common restriction on a set of basis functions that is particularly useful is to use *orthogonal functions*. If the orthogonality constraint $\langle \mathbf{g}_i, \mathbf{g}_j \rangle \;=\; \delta_{ij}$ is satisfied, it follows that $\mathbf{B}^{-1} \;=\; \mathbf{B}^{\mathsf{H}}$. This means that the calculation of the coefficients is simplified as there is no need for matrix inversion; projecting $\mathbf{x}$ onto each of the basis functions can give us the desired result:

$$g_k \;=\; \mathbf{b}_k^{\mathsf{H}}\,\mathbf{x} \;=\; \frac{\langle \mathbf{b}_k, \mathbf{x} \rangle}{\langle \mathbf{b}_k, \mathbf{b}_k \rangle}, \quad k = 1, 2, \ldots, K. \tag{3.5}$$

If the norm $\langle \mathbf{b}_k, \mathbf{b}_k \rangle$ is equal to unity, the basis functions can be said to be *orthonormal* to each other. This is often the case in signal modelling applications, which is why the terms ‘orthogonal’ and ‘orthonormal’ (incorrectly) tend to be used interchangeably in the literature. Well-known and used examples of complete representations include, for example, the Discrete Fourier Transform (DFT), the Short-time Fourier Transform (STFT) and the Discrete Wavelet Transform (DWT). In fact, it has to be noted that the last two, since they are linear TF representations, are generalisations of the additive model of Eq. 3.1, that have the form:

$$x(n) \;=\; \sum_{r=1}^{R}\sum_{k=1}^{K} g_{kr}\, b_{kr}(n), \quad \forall\, n \in \mathbb{Z}. \tag{3.6}$$

where $r = 1, 2, \ldots, R$ is the time-frame number. In this case, the functions $\{b_{kr}\}$ are localised both in the time and frequency domain and they are called TF *atoms*. Henceforth, the TF or frequency-domain representation of a time-domain signal will be denoted by its corresponding capital calligraphic symbol. For example, $\mathcal{X}$

[4] The dictionary associated with the representation is also said to be complete.

corresponds to x, where

$$\mathcal{X}(k,r) \;:=\; g_{kr}, \qquad \begin{array}{l} k = 1, 2, \ldots, K, \\ r = 1, 2, \ldots, R, \end{array} \tag{3.7}$$

is the TF representation of $x(n)$.

3.2 Non-parametric methods

Non-parametric methods generally involve some sort of reversible transformation, using a fixed set of basis functions. Because of this, they are broadly applicable to any kind of signal, and their calculation is usually highly optimised. For the case of musical signals (which have an evident structure in the frequency domain) complex exponentials are a common type of basis function, because of their ability to explicitly describe time and frequency information.

3.2.1 Fourier-related methods

The standard non-parametric method for decomposing a time-limited signal into a distribution of sinusoidal components (*i.e.* using complex exponentials as the basis functions) is the Fourier Transform (FT). For a continuous-time signal $x(t)$ the continuous FT (*i.e.*, the analysis equation CFT) and its inverse (*i.e.*, the synthesis equation ICFT) are defined as follows:

$$\mathsf{CFT}_x(\omega): \qquad \mathcal{X}(\omega) \;=\; \int_{-\infty}^{\infty} x(t)\,\mathrm{e}^{-\mathrm{j}\omega t}\;\mathrm{d}t \tag{3.8}$$

$$\mathsf{ICFT}_{\mathcal{X}}(t): \qquad x(t) \;=\; \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathcal{X}(\omega)\,\mathrm{e}^{\mathrm{j}\omega t}\;\mathrm{d}\omega. \tag{3.9}$$

For a discrete-time real signal $x(n)$ of length N, it can be shown [133] that the DFT and its inverse are

$$\begin{aligned}
\mathsf{DFT}_x(k): \qquad \mathcal{X}(k) &= \sum_{n=0}^{N-1} x(n)\,\mathrm{e}^{-\mathrm{j}2\pi kn/N}, \qquad k = 0, 1, \ldots, N-1 \quad (3.10)\\
\mathsf{IDFT}_{\mathcal{X}}(n): \qquad x(n) &= \frac{1}{N}\sum_{k=0}^{N-1} \mathcal{X}(k)\,\mathrm{e}^{\mathrm{j}2\pi kn/N}, \quad n = 0, 1, \ldots, N-1. (3.11)
\end{aligned}$$

$\mathcal{X}(k)$ is the (complex) value of the k-th *frequency coefficient.* The frequency axis is sampled in N (called the *length of the transform*) uniformly spaced frequencies centred at $f_k = (k/N)F_s$, where F_s is the sampling frequency. Thus, the frequency resolution is proportional to N (assuming that F_s remains constant). What we can see from Eq. 3.11 is that the Inverse DFT (IDFT) decomposes the signal $x(n)$ into a weighted sum of complex exponentials $e^{\mathrm{j}2\pi kn/N}$ which, for this case, are the basis functions. The complex exponentials are also orthonormal: if $s_k \equiv e^{\mathrm{j}2\pi kn/N}$, then

$$\begin{aligned}
\langle s_k, s_m\rangle &= \sum_{n=0}^{N-1} s_k(n)\,\overline{s_m(n)} = \sum_{n=0}^{N-1} \mathrm{e}^{\mathrm{j}2\pi kn/N}\mathrm{e}^{-\mathrm{j}2\pi mn/N}\\
&= \sum_{n=0}^{N-1} \mathrm{e}^{\mathrm{j}2\pi(k-m)n/N} = \begin{cases} 1, & \text{for } k = m\\ \frac{1-\mathrm{e}^{\mathrm{j}2\pi(k-m)}}{1-\mathrm{e}^{\mathrm{j}2\pi(k-m)/N}} = 0, & \text{for } k \neq m.\end{cases}
\end{aligned}$$

The fraction is the closed form expression of a geometric series, and the end result is equivalent to $\langle s_k, s_m\rangle = \delta_{km}$, *i.e.* the complex sinusoids are orthonormal. This leads to a general reduction of calculation complexity, as explained above.

One of the advantages of using this transformation is that it is *linear* (as are the majority of the non-parametric methods). The advantage of the linearity can be shown with an example: suppose we have 3 audio signals summed into a mixture signal that has been transformed through the DFT into the frequency domain and we subtract from the spectrum the frequency components corresponding to one of the signals. If we then inverse-transform (reconstruct) the residual onto the time domain, we get exactly what we had before, minus the extracted signal. In other words, removal of content via the frequency domain does not introduce additional artefacts, *i.e.*, components that were not there before the transformation. This is something that is highly important for any realistic approach towards source separation.[5] Nevertheless, the DFT is not totally suitable for signals such as those encountered in music. Musical signals must generally be considered to be *nonstationary* signals: assuming that they contain sinusoidal components and some stochastic content, the number and amplitude of the sinusoids is potentially changing quickly, as are the statistical properties of the stochastic content.[6] The result is a *time-varying spectrum.* For a situation like this a spectrum analysis method using the DFT would not be adequate, since this transform tells us which frequencies existed for the *total* duration of the signal, and not the frequencies that existed at any particular time. Hence, a method that would give a description of the energy density of the signal simultaneously in time and frequency (resulting in a TF *distribution*) would be more appropriate. A common method used for this purpose is the STFT. This method segments the signal into a number of short-duration frames and performs a DFT separately for each of these frames. Mathematically, it is a joint function of time and frequency and for the discrete-time case is

$$\mathsf{STFT}_x^h(k,m): \quad \mathcal{X}(k,m) \;=\; \sum_{n=-\infty}^{\infty} x(n)\,h(n-m)\,\mathrm{e}^{-\mathrm{j}2\pi kn/N}. \tag{3.12}$$

$m \in \mathbb{N}$ is the time instant defining the starting point of the frame over which the transform is calculated. $h(n)$ is a *window* function that is applied before the calculation of the transform. Various kinds of window functions have been designed so far for this purpose. Many of them have a shape that approaches zero at its boundaries. This is in order to prevent spurious spectral components (spectral 'leakage') arising from discontinuities in the signal amplitude between opposite window boundaries. The discontinuities are due to the fact that the STFT effectively assumes that it is creating a Fourier series expansion of a periodically extended version of the analysis frame. Common examples are the Hamming, Hanning and Blackman-Harris windows [70, 123]. The main considerations af-

[5] §4.2.3 and §4.12.2 include further discussions regarding the relation between linearity of the mixture model and the representation in the context of examining existing methods.

[6] By contrast, we can define a signal as *stationary* during a specified amount of time when this signal *is comprised of sinusoidal components whose statistical properties do not change over that specified duration.*

fecting its selection are its main spectral lobe width and the energy of its side lobes.

The window is non-zero only within the interval $[0, M-1]$, where $M \leq N$ is the size of the window. This changes the limits of the above sum accordingly. Also, in order for the calculation of the STFT to be done efficiently, we introduce one more term in Eq. 3.12, the *hop size* L. In this way, instead of doing the calculation after each step m, it will be done after every L samples of $x(n)$. Let us define, now, a windowed segment of $x(n)$ as:

$$x_r(n) \equiv x(n)\, h(n - rL) \tag{3.13}$$

and the fixed-time-origin sequence:

$$\check{x}_r(n) \equiv x(n + rL)\, h(n), \qquad \text{for } n = 0, 1, \ldots, M-1 \tag{3.14}$$

where $r = 0, 1, \ldots, R-1$ is the index of the calculation frame. The STFT can be written now as:

$$\begin{aligned}
\mathsf{STFT}_x^h(k,r): \quad \mathcal{X}(k,r) &= \sum_{n=rL}^{rL+M-1} x(n)\, h(n-rL)\, \mathrm{e}^{-\mathrm{j}2\pi kn/N} && (3.15)\\
&= \sum_{n=0}^{M-1} x(n+rL)\, h(n)\, \mathrm{e}^{-\mathrm{j}2\pi k(n+rL)/N} && (3.16)\\
&= \mathsf{DFT}_{\check{x}_r}(k)\, \mathrm{e}^{-\mathrm{j}2\pi krL/N}, && (3.17)
\end{aligned}$$

Eq. 3.17 shows that the STFT can be directly implemented by calculating the DFT of each of the blocks of M samples $\check{x}_r(n)$. Regarding the choice of the length L, it has to be done so that it is reasonable to expect no significant variation of the parameters within that interval. When $L < M$ (which is the most common practice), the frames are overlapping. Also, the transform length N is often taken to be larger than M, in order to increase the apparent spectral resolution. The additional samples for $M \leq |n| \leq N-1$ have zero value, thus this method is called *zero-padding.* It is worth noting that while zero-padding can, indeed, increase the frequency resolution, it does not provide any extra spectral information (such as revealing hidden partials). Rather, it provides an interpolated viewpoint for improving the localisation of all the spectral information, including the maximum on a partial's main lobe.

The visualisation of the STFT analysis is a 2-D image, called the *spectrogram*. It displays the *magnitude* of the coefficients, displayed on a logarithmic scale. For a particular TF point (k, r) the spectrogram is defined as:

$$\mathsf{spectrogram}_x^h(k, r) \ \equiv \ 20 \log_{10} |\mathcal{X}(k, r)|. \tag{3.18}$$

So far, we have seen how the TF analysis is performed using the STFT. After this step and any processing operations on the TF content, the signal will have to be reconstructed in the time domain. This is the *synthesis* step and it is performed in a similar manner to the analysis. If $\mathcal{X}'(k, r) \equiv \mathcal{X}'$ is the modified frame spectrum, then the transformed signal can be obtained by a *weighted overlap-add* procedure [37, 132]

$$\begin{aligned} x'(n) &= \sum_{r=0}^{R-1} v(n - rL) \cdot \mathsf{IDFT}_{\mathcal{X}'}(n) \\ &= \sum_{r=0}^{R-1} v(n - rL) \, \breve{x}'_r(n - rL) \\ &= \sum_{r=0}^{R-1} v(n - rL) \, x'_r(n), \end{aligned}$$

where $v(n)$ is a synthesis window and $x'_r(n)$ a potentially modified version of $x_r(n)$. If $x'(n) = x(n)$ (*i.e.*, no modifications have been performed on the original signal) perfect reconstruction is achieved when the windows $v(n)$ and $h(n)$ satisfy the constraint

$$\sum_{r=0}^{R-1} v(n - rL) \, h(n - rL) \ = \ 1, \qquad \forall \, n. \tag{3.19}$$

This can be achieved when *perfect reconstruction windows* (*i.e.*, the windows the shifted copies of which overlap and add to 1) are employed. The window functions mentioned above fulfil this requirement. In practice, the choice of $v(n)$ depends on whether the STFT has been modified or not before reconstruction. If it has, then possibly erroneous phase estimations or inadequate parameter interpolation before synthesis could lead to frame-edge discontinuities that could be audible. A suitable synthesis window would be one which, while still achieving perfect reconstruction, it would try to minimise those discontinuities. The use of the triangular window is one sensible choice for this purpose [67, p. 74]. In fact, it is used as a part of a hybrid synthesis window: if $t(n)$ is the triangular window, then the synthesis window will be

$$v(n) \;=\; \frac{t(n)}{h(n)}. \tag{3.20}$$

The use of a hybrid-type increases flexibility, because now $h(n)$ does not necessarily have to fulfil the perfect reconstruction constraint as long as $t(n)$ does. However, $h(n)$ still needs to be nonzero within the width of $t(n)$, otherwise unwanted discontinuities arise at the edges of the frame.

Finally, applying a DFT with N chosen to be a power of 2 or 4 enables us to use the Fast Fourier Transform (FFT), a well-established algorithm which is known for its high computational efficiency [17].

Time-Frequency localisation

The negative effect of using a window function is that the spectral peaks appear broadened. This is due to the fact that a multiplication operation between a window $h(n)$ with $x(n)$ in the time domain is a convolution of $\mathsf{DFT}_h(k)$ with $\mathcal{X}(k)$ in the frequency domain. Hence, although a spectral peak corresponding to a single frequency value should have the shape of a delta function[7] placed at that frequency value, it ends up having the shape of the Fourier transform of the window function (the observed bandwidth is the bandwidth of the window). Increasing the length of the window in the time domain could result in a better localisation

[7] The *Dirac delta* is often described as the 'function' or, more correctly, the distribution that satisfies $\delta(t) = \begin{cases} \infty, & \text{for } t = 0 \\ 0, & \text{otherwise,} \end{cases}$ and $\int_{-\infty}^{\infty} \delta(t)\,\mathrm{d}t = 1$. This leads to the property $\int_{-\infty}^{\infty} f(t)\delta(t - t_0)\,\mathrm{d}t = f(t_0)$, for any t_0. It is the continuous equivalent of the Kronecker delta (see p. 200).

in the frequency domain. However, as well as the stationarity issues discussed above, this comes with an additional price: the localisation in time worsens, and so any information about rapid temporal signal changes is averaged. This is a general problem of TF representations, known as the *uncertainty principle*, which gives a lower bound on the product of the time duration Δ_t and the frequency bandwidth Δ_ω of the window function:[8]

$$\Delta_t \, \Delta_\omega \;\geq\; \frac{1}{2}. \tag{3.21}$$

The localisation of TF components can be depicted by *tiles* in a TF plane. A tile is rectangular with dimensions Δ_t and Δ_ω centred at a point (t_0, ω_0). Furthermore, the actual lower bound in Eq. 3.21 is achieved by using Gaussian window functions. Eq. 3.21 clearly shows that there is a *trade-off* between time and frequency resolution. In musical signal processing the decision about the window length is often difficult to make: slowly time-varying partials are ideally represented by long windows, while quickly time-varying segments (like note attacks) require short windows. These two extremes appear very often in the same mixture. Practically, this can cause problems when it comes to calculating the parameters of frequency components using the STFT representation. Also, when two or more frequencies are very close together their spectral peaks will overlap. Often, this makes it hard to tell (by just observing the spectrogram) whether a peak corresponds to a single frequency component or several. As has already been said, the case of overlapping partials (or partials that are very close together) is something common in musical signals. For source separation purposes this is a crucial issue that has to be dealt with, as we will see in the following chapters.

It is important to stress that the present discussion about TF localisation is carried out under the assumption that the signal remains stationary for the duration of the window. However, as mentioned on p. 33, musical signals are realistically expected

[8] Δ_t and Δ_ω are in fact the *time domain standard deviation* and the *frequency domain standard deviation* respectively, satisfying the equations $\Delta_t^2 = \frac{1}{E}\int_{-\infty}^{\infty}(t - \langle t\rangle)^2 |s(t)|^2 \,\mathrm{d}t$ and $\Delta_\omega^2 = \frac{1}{2\pi E}\int_{-\infty}^{\infty}(\omega - \langle\omega\rangle)^2 |S(\omega)|^2 \,\mathrm{d}\omega$, where: $s(t)$ is the *analytic* form of the continuous-time signal $x(t)$; E is the energy of $s(t)$; $\langle t\rangle$ and $\langle\omega\rangle$ are the mean values of time and angular frequency, respectively [134].

to exhibit nonstationary behaviour. This behaviour often causes the shape of the spectral peaks to be different, and more importantly, *wider* than the shape of the Fourier transform of the window function. In other words, the blurring of frequency content in STFT is a result of three factors: the non-conformity of the signal to the bin frequencies, the windowing process and any nonstationarity of the signal.

3.2.2 Multiresolution approaches

In order to address the TF trade-off, various *multiresolution* approaches have been investigated. In these representations Δ_t and Δ_ω vary with time and/or frequency, corresponding to different tilings in the TF plain. One common multiresolution representation is provided using wavelet basis functions [110]. This can be done, for example, by using the DWT or an optimised version of it, the Wavelet Packet Transform (WPT). The tilings produced by the DWT constitute a dyadic sampling grid in such a way that good frequency localisation is achieved at low frequencies and good time localisation at high frequencies. In other words, the window length is inversely proportional to frequency. Kisilev *et al.* reported an increase in sparsity[9] through the use of wavelets and, as a result, in separation performance in the context of Blind Source Separation (BSS) on a few simple audio mixtures [93].

A similar multiresolution transform is the Constant-Q (CQ) transform ($Q = f/\Delta f$, where Δf is the resolution bandwidth and f the centre frequency in a TF tile) [22]. These CQ representations are similar to the way the HAS distributes frequency components into critical bands. So, it could be argued that they would be well suited for musical signals and mixtures.

Finally, frequency-warped representations such as the Bark scale and the Mel scale [24], as well as auditory representations such as the cochleagram [108] and the correlogram [152] also belong in this category.

[9] The sparsity of a representation is defined in §3.2.3.

3.2.3 Brief introduction on sparsity and W-disjoint orthogonality

All the above representations have the effect of increasing the *sparsity* of the signal, compared to its time-domain representation. This is generally defined as

$$\zeta = \left(\sum_{k=1}^{K} |g_k|^p \right)^{1/p} \tag{3.22}$$

where $p \in [0, 1]$. If 'sparsified' (*i.e.*, transformed to a mid-level TF domain where they appear sparser than in the time domain) signals are mixed together, it would be expected that they would not overlap significantly in that particular TF domain. The condition that is satisfied when two signals are non-overlapping at each TF point is called W-Disjoint Orthogonality (WDO). Since sparsity and WDO are employed as assumptions by source separation approaches, their implications and consequences on the separation performance will be examined in more detail in the context of reviewing related work (see §4.12).

3.3 Sinusoidal modelling

In contrast to the above methods, the sinusoidal model is a parametric method. It is one of the most common analysis models in speech and music. It was first introduced by McAulay and Quatieri [114] and it was originally applied to speech signals. The deterministic part of the signal is expressed as a time-varying sum of sinusoids $M \equiv M(t)$ with instantaneous amplitude $a_m(t)$ and phase $\phi_m(t)$:

$$x(t) = \sum_{m=1}^{M} a_m(t) \cos(\phi_m(t)). \tag{3.23}$$

The sinusoidal representation of a sound in terms of its instantaneous amplitude and phase is obtained in the following way:

1. Compute the TF representation (usually the STFT) of the sound,

2. detect the spectral peaks and calculate their parameters (magnitude, frequency and phase), and
3. track the sinusoidal parameters from frame to frame, identifying in this way the time-varying sinusoidal tracks.

Although the basic sinusoidal model has proven to be very useful for performing certain basic musical effects (*e.g.* time-stretching or pitch-shifting) and to a good fidelity, it still cannot represent transient and noise content, which is usually an important part of musical sounds. Several popular models have tried to incorporate a model of the *residual* part of the sound, along with the sinusoidal model. One well-known method is Spectral Modeling Synthesis (SMS), otherwise known as *deterministic plus stochastic decomposition* [147, 148]. The residual part is assumed to be the stochastic component, modelled as filtered white noise (where the filter is time-varying), while the deterministic part is the sum of time-varying sinusoids. The sinusoidal plus residual model can be seen as a generalisation of the basic sinusoidal model. Various modifications and extensions have been proposed for this model. For example, a noise model based on perceptual properties [66], and the use of hidden Markov models along with the Viterbi algorithm for harmonic tracking [138].

3.4 Summary

The task of representing musical signals in a way that their underlying mid-level structure becomes evident is very important. This is because they make it possible to perform advanced and flexible manipulations on their content (such as decomposing a mixture into its source signals). The methods for representing signals are often categorised into non-parametric and parametric, and the selection of the right method depends strongly on the way the resulting representation will be exploited.

Non-parametric methods do not generally imply prior knowledge of the signal, making them applicable to a large range of cases. They usually result in a projection of the signal onto the TF plane, and a widely applied method is the STFT. The STFT is a good solution for displaying simultaneously the time and frequency structure of the signal. Although what is displayed is a crude approximation of the 'real' situation (due to the inevitable TF trade-off), this representation has the advantage of allowing for flexible analysis/resynthesis operations on music/speech signals and a variety of content-altering transformations. Its time and frequency resolution is constant though, and is directly dependent on the choice of the window length. Multiresolution approaches, such as the DWT and the CQ transform, overcome this limitation by offering frequency-dependent resolution. In this way they provide both a more realistic display of rapidly changing signal energy and slowly-moving partials. However, it is not always straightforward to extract partial content from these representations.

Parametric methods make use of prior knowledge for the signals, offering in this way an interpretation of the representation. Because of that, they can be applied to musical signals for performing advanced operations in an efficient way. For example, by using the sinusoidal plus residual model and its extensions we can create musical effects and other manipulations, with a good fidelity.

Both parametric and non-parametric methods are used heavily for identifying and extracting source structures from mixtures in the context of source separation, as will be seen in the next chapter.

CHAPTER 4

Single-channel source separation

So far, we have discussed the meaning of the 'musicality' of sounds or signals and its relation to human perception and understanding. We have also reviewed some of the important ways for representing those signals as a mid-level stage for designing systems capable of 'understanding' them. This chapter considers the specific problem of single-channel source separation. First, the problem is introduced, underlining the challenges associated with it. A focus on musical sources is made, followed by a detailed description of existing methods. This is carried out with the help of a classification framework, intended to assist the reader in placing the proposed method within its current context. The chapter ends with an introduction to the available performance analysis approaches for separation systems.

4.1 The audio source separation problem

4.1.1 Polyphonic mixtures

It is usual for audio signals originating from different sources to coexist simultaneously in the form of a mixture. For the majority of Western musical pieces this mixture consists of interweaving melodies[1] coming from musical instruments. Roughly speaking, the pitched (harmonic or near-harmonic) instruments are usually responsible for the melodies, while the pitched nonharmonic ones are mostly responsible for the sense of rhythm. We can call this type of mixture signal a *polyphonic* signal.

The term "polyphonic" is originally taken from music theory, where it is used in a more limited sense than it is here. In music theory, polyphonic describes a piece of music that "combines several distinct melodic lines simultaneously" [135]. This term contrasts with *monophonic*: a piece of music consisting of a single melodic line, and *homophonic*: a piece of music where several melodic lines are combined, although instead of being distinct (as in polyphonic music) they move in the same rhythm, creating a clear succession of chords. One thing to note is that it is often hard to draw clear distinctions between these terms when we want to characterise a musical piece [135]. In fact, the area between the formal definitions of polyphony and homophony is more of a continuum, and the position of a musical piece in this continuum depends on the degree of independence between the melodic lines: the higher the inter-melodic independence, the closer a musical piece is to being classified as polyphonic; the lower the independence, the closer it is to being classified as homophonic.[2] Furthermore, the majority of Western musical pieces (especially popular music) very often involve both polyphonic and homophonic characteristics. A musical source separation system should ideally be able to deal with all those cases, regardless of where they are placed on the polyphony/homophony continuum, as long as they are mixtures comprised of different musical source signals. This defines a polyphonic signal for the purposes of this thesis. A corollary of this is that the source signals will be monophonic signals (*i.e.*, no chords are assumed for this thesis).

[1] We note here that the reference to the melody takes us immediately to the mid and high levels of human perception, mentioned in Ch. 3. We stress again that it is *according to humans* that most of the musical signals consist of melodies (and other relevant structures).

[2] A strict definition or use of independence (an example of which would be to talk about *statistical independence* on the melody level) is not needed here, as the word is only used for making the point of unclear distinctions between polyphony and homophony.

In addition to the polyphony-related definitions, a few further clarifications are needed to set up the context for the description of the source separation problem. As will be seen further below, many of the existing separation methods do not produce a residual, or if they do, they treat it as undesirable content. The work presented here treats the residual much differently, since it deliberately expects a part of the sources (the one not belonging to the current model) to be found in the residual. The following definitions reflect the particular way the involved signals are considered in this work.

4.1.2 Signal categorisation in mixing and separation scenarios

Two main sets of entities are involved in any kind of signal mixing or separation process: the *original source signals* and the *mixture channels*. In a mixing scenario we refer to the original source signals as the signals that are the input to the mixing process, which produces the mixture; they correspond to the sound produced by a particular source, usually at the point where this sound was produced.[3] The set of J original source signals can be denoted by $\{s_j(n)\}_{j=1}^{J}$ (see Fig. 4.1a). In a separation scenario we also have the *estimated* or *extracted source signals*. These are the output signals of a separation system, and by definition they correspond to and are expected to match to a certain degree the original source signals. The set of J extracted source signals can be denoted by $\{\hat{s}_j(n)\}_{j=1}^{J}$ (see Fig. 4.1b).

As an aside, it is worth making a few additional remarks regarding definitions related to separation. Estimation and extraction are distinguished in this thesis as two different processes appearing in series in a separation system, as will be seen in §4.3. However, for the purpose of differentiating between the two 'kinds' of source signals (original and estimated/extracted), we can refer to either the estimation and extraction processes, since they fall into the same category. Furthermore, it is worth highlighting the contextual difference between the words 'extraction' and 'separation': the word 'extraction' focuses on the idea that a particular source

[3] This point may not necessarily correspond to a real point (or area) in the recording space; instead, it can be a simulated one. The mixtures that are created using simulated recording conditions can be called *artificially-created* mixtures, in contrast to the *naturally-created* mixtures.

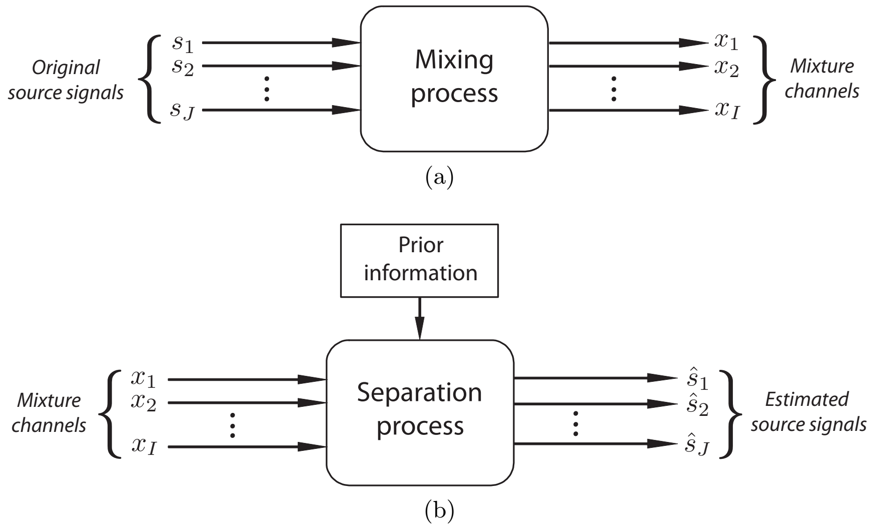


Figure 4.1: General view of the ASS problem (see §4.1.3 for a definition), for any combination of I and J. (a) shows how the original source signals and the mixture channels are related through the mixing process, while (b) shows how the mixture channels and the extracted signals are related through the separation process.

needs to be isolated from a composite signal, which may not contain other intended sources (see §4.1.4). On the other hand, the word 'separation' implies that there are more than one source signals that constitute a mixture, and that they all need to be isolated from *each other*, rather than a background signal. Finally, with regards to the word 'separation', there are other authors who make use of different words to describe the same process. Common examples are the words 'unmixing' and 'demixing', 'segregation' and 'decomposition' (although these last two can cause confusion, since they are used differently in other signal processing contexts).

The last set of signals to consider is the *mixture channels*. Each of them corresponds to a different version of the mixture; they constitute the outputs of a mixing process and the inputs of a separation process. The difference in the mixture channels lies in the point where the mixing process took place. This mixing point can be one placed in a real acoustic space, or it can be a virtual mixing point. When the mixing points are placed in the real acoustic space, they correspond to the sensors (or microphones, the transducers in Fig. 2.2) placed in different positions in space, with which the sound coming from the sources is observed and translated into signals (the mixture channels). Virtual mixing points are the ones established by an artificial mixing process *i.e.*, a process that simulates the acoustics of a certain space and the arrangement of the acoustic sources in that space and it is often carried out in the studio. The set of I mixture channels can be denoted by $\{x_i(n)\}_{i=1}^{I}$.

4.1.3 Definition of the problem

The procedure of extracting, or estimating, the audio signals corresponding to each source, given the mixture channels, is called the *Audio Source Separation* (ASS) problem. This problem can be classified, according to the respective values of I and J, into three cases:

- The *determined* case: $I = J$
- The *overdetermined* case: $I > J$
- The *underdetermined* case: $I < J$

These terms are borrowed from linear algebra, since the mixing process can be approximated as a system of linear equations, as will be seen below in Eq. 4.5. Considering the above cases from the point of view of source separation difficulty, the underdetermined case is generally the hardest one. Indeed, while blind methods (such as ICA) that do not belong in this category tend to give good results [83, 26], underdetermined situations (such as the one dealt with in this thesis) can pose greater challenges, as will be shown below. This is simply because of the fact that there is less information available for inferring the sources.

In order to construct signal processing techniques for solving the source separation problem, the relationship between the mixture and the source signals has to be described mathematically. In other words, a model of the mixing process has to be defined.[4] At this point it is deemed important to give an emphasis on the interconnections between the various models used in this thesis (some of which will be introduced below), because it will help understand better the process of

[4] We are starting from the low-level representation going up, see Fig. 3.1.

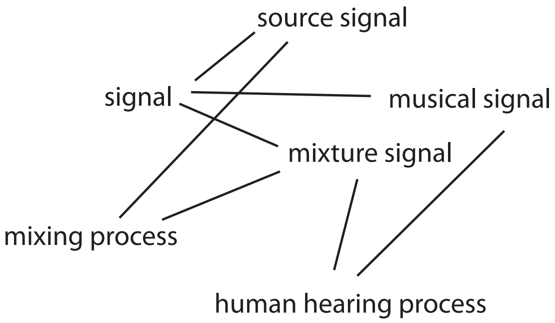


Figure 4.2: Various models used in this thesis, along with their interconnections with regards to how these models are defined.

designing a musical source separation system. Fig. 4.2 shows the interconnections of some of the mostly used models made in this thesis.

Before we continue with the model of the mixing process, though, the idea of what we mean by ‘source’ in source separation has to be clarified.

4.1.4 What is a ‘source’?

First of all, it is worth making clear that by referring to “source separation”, what authors most often mean is “source *signal* separation”: the word ‘signal’ is omitted for the sake of simplicity, when used in this particular phrase. This convention is used here, as well.

The need for a definition of source or source signal comes from the fact that in source separation we have to deal already with two different versions of these entities: the original sources and the extracted sources. Naturally, a successful separation algorithm would process the mixture in such a way that the extracted sources match the original ones to a sufficiently high degree. The problem is that, in the situation where only the mixture is available (and not the original sources), defining the extracted source signals conceptually is not straightforward. Below, it will be shown why this is, and how a non-problematic definition can be made.

The extracted source signals have to satisfy two certain characteristics:

- They have to correspond to an one-to-one relationship with the original source signals.
- Since the mixtures are assumed to be musical, the extracted signals have to conform to the definition of what kind of signal is considered musical.

Since musicality is a subjective quality related to the HAS, it appears appropriate to define the sources comprising a mixture using terminology from the field of auditory psychology and scene analysis. In §2.1 the concepts of *auditory objects*, *auditory stimulus* and *auditory images* were employed for describing the perception and mental organisation of sounds by humans. Thus, by attempting a simple parallelism, the original sources can be said to correspond to the auditory objects which contribute to the overall auditory stimulus (corresponding to the signal mixture), and the estimated sources to the auditory images which appear in the mind of the listener. However, this parallelism is inadequate for describing the sources, since the auditory object/auditory image relationship is not always one-to-one (*i.e.*, the first desired characteristic for the sources is not satisfied). Indeed, there is often a nonlinear relationship between how the stimulus is created as the 'sum of auditory objects' and how the stimulus is perceived as the 'sum of auditory images'. For example, a mixture of seven violins playing simultaneously would probably be perceived by most listeners as a mixture of less than seven violins, because of the limitations of the HAS. Also, the notion of the auditory object can change with the context, depending on the particular focus or requirements of the listener. An example of this would be the difference between considering the drums as one object (*i.e.*, when referring to a mixture of "bass, guitar and drums"), or several (*i.e.*, when referring to a mixture of "bass, guitar, snare drum, kick drum and hi-hat"). Because of this, the conditions under which the relationship auditory object/auditory image is always one-to-one have to be identified.

A sensible way to go about identifying these conditions is to start by considering the auditory image. This is because the auditory image is more stable as a descriptor than the auditory object, meaning that the first refers always to a

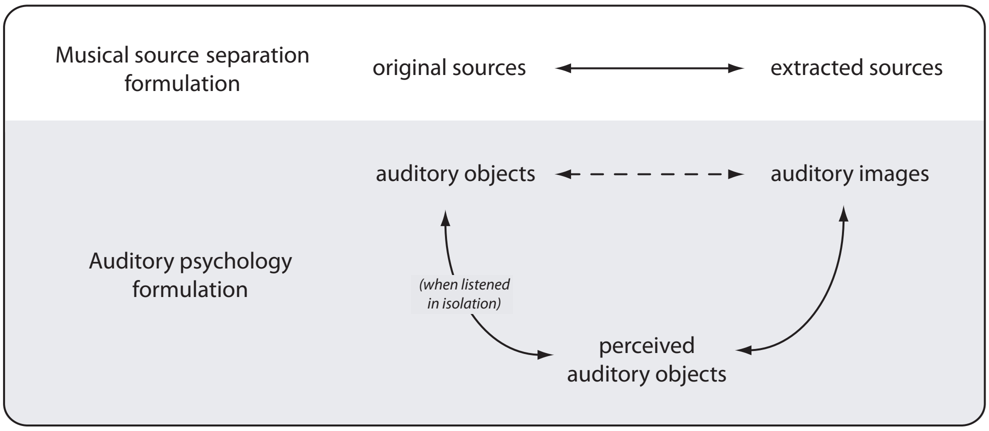

Musical source separation formulation
original sources
extracted sources
Auditory psychology formulation
auditory objects
auditory images
(when listened in isolation)
perceived auditory objects

Figure 4.3: A definition of original sources and extracted sources as conceptual entities and how they are related to each other in the context of musical source separation by drawing parallels from the area of auditory psychology. Straight arrows signify one-to-one relationship while the dashed arrow signifies a relationship which is not always one-to-one.

distinct entity. This entity can be called the *perceived auditory object*. For example, when the listener perceives an auditory stimulus as containing 5 saxophone melodies, the melodies correspond to 5 auditory images and these images are associated with 5 perceived auditory objects (the saxophones). The latter may not, in general, correspond to the auditory objects responsible for the overall auditory stimulus – the number of perceived auditory objects may often be less than the number of auditory objects [81]. The only time when a perceived auditory object is guaranteed to refer exclusively to a distinct auditory object is when the sound coming from the auditory object is perceived in isolation. The relationship between these perceptual entities is illustrated in Fig. 4.3. As it can be seen, the one-to-one relationship required for formulating a complete parallelism is now conditionally satisfied. This lets us introduce the following definition of the source in source separation: it is *the auditory object (or the group of auditory objects) which produces auditory stimuli that are expected to be perceived by humans as a single auditory image when presented in isolation (*i.e.*, not in a mixture).*

The above definition enables us to make sense of the mixture from the point of view of the *creator* of the mixture.[5] The sources are defined according to how the creator of the mixture perceives those sources as a listener (the objective view).

[5] The word ‘creator’ refers, here, to the person who is responsible for the mixing process.

Obviously, the estimated source signals have to make sense as being parts of the mixture for the listener, but not only that; they also have to satisfy the intention of the creator of the mixture. However, the prior information available for the separation is limited to the degree that it cannot be known with certainty what the intention of the creator was. This relative uncertainty is highlighted by the use of the word 'expected' in the above definition.

4.2 Models of the mixing process

4.2.1 Multi-channel models

The most general expression of the mixing process between the source signals $\{s_j(n)\}_{j=1}^{J}$ and the mixture channels $\{x_i(n)\}_{i=1}^{I}$ can be formulated mathematically as follows:

$$x_i(n) \;=\; \sum_{j=1}^{J} \sum_{\nu=-\infty}^{\infty} a_{ij}(n-\nu,\nu)\, s_j(n-\nu) \;+\; v_i(n), \qquad i = 1, 2, \ldots, I. \tag{4.1}$$

$a_{ij}(n,\nu)$ represents a time-varying filtering process between the j-th source and the i-th channel. ν denotes delay in samples and $v_i(n)$ any background noise introduced at the i-th channel. This is a model of a mixing process where the sources and/or the microphones are moving in time inside a reverberant space. If we remove the time variance of the mixing filters and the noise element we end up with the following expression:

$$x_i(n) \;=\; \sum_{j=1}^{J} \sum_{\nu=-\infty}^{\infty} a_{ij}(\nu)\, s_j(n-\nu), \qquad i = 1, 2, \ldots, I. \tag{4.2}$$

where the impulse response of the filters is:[6]

$$\xi_{ij}(n) = \sum_{\nu=-\infty}^{\infty} a_{ij}(\nu)\, \delta(n-\nu). \tag{4.3}$$

If the mixture is assumed to have been recorded anechoically (*i.e.*, in a non-reverberant space) the above expression can be simplified further:

$$x_i(n) \;=\; \sum_{j=1}^{J} a_{ij}\, s_j(n-\nu_{ij}), \qquad i = 1, 2, \ldots, I, \tag{4.4}$$

where ν_{ij} is the delay of the signal going from the j-th source to the i-th microphone. This is called an *anechoic* mixture, as opposed to the mixtures expressed in Eq. 4.1 and 4.2, which are *convolutive* ones. If we ignore the delays, Eq. 4.4 becomes:

$$x_i(n) \;=\; \sum_{j=1}^{J} a_{ij}\, s_j(n), \qquad i = 1, 2, \ldots, I. \tag{4.5}$$

This equation describes the simplest of all the mixture models, which is the *instantaneous* mixture. The form of Eq. 4.5, *i.e.*, the fact that it describes a system of linear equations shows the main characteristic of instantaneous mixtures: they are the result of a *linear* process. Linear processes or systems have a particular significance when constructing models, and this is because of their special properties. The main property that a linear system satisfies is the *superposition* property, which is described as follows: if there are two causes c_1 and c_2 the effects of which on a system are e_1 and e_2 respectively:

$$c_1 \;\longrightarrow\; e_1 \quad \text{and} \quad c_2 \;\longrightarrow\; e_2, \tag{4.6}$$

then the system is linear if:

$$k_1\, c_1 \,+\, k_2\, c_2 \;\longrightarrow\; k_1\, e_1 \,+\, k_2\, e_2, \tag{4.7}$$

[6] Eq. 4.3 makes use of a general theoretical expression for the filtering process. In practice, a finite number of filter taps is used (the filters are assumed to be Finite Impulse Response (FIR) ones). In this case, if the length of the impulse response of the filter between the j-th source and i-th channel is Z_{ij}, the impulse response of the filter will be $\xi_{ij}(n) = \sum_{\zeta=1}^{Z_{ij}} a_{ij}(\zeta)\delta(n-\nu_{ij\zeta})$.

for any real or complex constants k_1 and k_2 [103]. The superposition property is a combination of the *additivity* and *homogeneity* properties, which become apparent through the addition and multiplication operations in the above equation.

Linearity plays a crucial role in the design of separation strategies, as will also be discussed in §4.2.3. In particular, the assumption of mixture linearity helps to simplify the problem and reduce its indeterminacy. Although real musical recordings are often far from linear mixtures, as long as they do not involve extremely reverberant spaces, the instantaneous mixing model would be expected to provide a viable solution (or at least a good starting point) for cases of SO and AQO applications that do not require particularly high separation fidelity. For this reason, the separation approach proposed in Ch. 5 will make use of the instantaneous mixing model.

4.2.2 Single-channel instantaneous mixtures

The case of extreme indeterminacy of the ASS problem is when $I = 1$, *i.e.*, when there is just one mixture channel available. In this case, the mixing model of Eq. 4.5 becomes:

$$x(n) \;=\; \sum_{j=1}^{J} a_j \, s_j(n), \tag{4.8}$$

This can be written in vector notation as:

$$x(n) \;=\; \mathbf{s}^{\mathsf{T}} \, \mathbf{a}, \tag{4.9}$$

where $\mathbf{a} = [a_1 \, a_2 \, \ldots \, a_J]^{\mathsf{T}}$ and $\mathbf{s} = [s_1(n) \, s_2(n) \, \ldots \, s_J(n)]^{\mathsf{T}}$. If the length of the mixture signal in samples is T (*i.e.*, $n \in [1, T]$) Eq. 4.9 can be written more compactly as:

$$\mathbf{x} \;=\; \mathbf{S}^{\mathsf{T}} \, \mathbf{a}, \tag{4.10}$$

where $\mathbf{x} = [x(1) \, x(2) \, \ldots \, x(T)]^{\mathsf{T}}$ and $\mathbf{S} = [\mathbf{s}(1) \, \mathbf{s}(2) \, \ldots \, \mathbf{s}(T)]$ a matrix of size $J \times T$. The problem of single-channel ASS in instantaneous mixtures is, thus, the problem of estimating $\mathbf{S}$ and $\mathbf{a}$ when the only known quantity is the mixture vector $\mathbf{x}$. Strictly speaking, this can be characterised as a *Blind Audio Source Separation* (BASS) problem – the term 'blind' being used to signify minimal or (as it is the case here) no *a priori* information (see §1.2).

In fact, it has to be noted that, when $I = 1$, $\mathbf{a}$ is usually already known in practice, or at least assumed to be equal to $[1 \ldots 1]^\mathsf{T}$. An exception to this would be if a stereo source separation scenario was considered as 'dual-mono' separation: carrying out separation in each channel independently using a method suited for single-channel mixtures (see §6.5). In that case it would be generally expected that $\mathbf{a} \neq [1 \ldots 1]^\mathsf{T}$.

4.2.3 Representation of the model in the time-frequency domain

As discussed in Ch. 3, transforming the time-domain audio signal to an equivalent mid-level representation is a crucial step towards effective analysis and manipulation procedures of that signal. This is because the new representation – assuming that it has been appropriately chosen – will hopefully highlight the signal features that matter for the processing goal. Within the wider context of this thesis, this goal is the identification and extraction of source signals from one or several mixtures. The majority of audio source separation methods do, indeed, operate largely on a mid-level representation domain (or a combination/succession of multiple domains), thus the use of the chosen representation has to be incorporated in the models of the mixing process. Although – as seen in Ch. 3 – TF non-parametric representations are just one category of mid-level representations, source separation methods (and, indeed, the approach presented in this thesis) make wide use of them, if only as the first stage for deriving subsequent parametric representations. The use of a TF representation implies that the model of the mixing process includes an explicit consideration of this different representation. The reason that this consideration has to be explicit in the mixing model is that in many cases the move from the time domain to a TF representation alters some properties of the time-domain model. A common property that may not hold in the chosen TF domain (while it does in the time domain) is the assumption of linear summation of the sources.

As will be discussed in more detail in §4.12.2, discarding the phase information from a TF representation has often proved to be a reasonable simplification practice for many methods. By choosing not to deal with the phases, the focus is moved to the absolute values of the expansion coefficients (*i.e.*, the magnitude values). This results in a loss of the linearity property since, while linearity is preserved in the mixing model when including the phases of the TF coefficients:[7]

$$x(n) = s_1(n) + s_2(n),\ \forall n \quad \Rightarrow \quad \mathcal{X}(r,k) = \mathcal{S}_1(r,k) + \mathcal{S}_2(r,k),\ \forall r,k. \tag{4.11}$$

the same does not generally hold when the magnitudes are used:

$$x(n) = s_1(n) + s_2(n),\ \forall n \quad \not\Rightarrow \quad |\mathcal{X}(r,k)| = |\mathcal{S}_1(r,k)| + |\mathcal{S}_2(r,k)|,\ \forall r,k, \tag{4.12}$$

where $r \in [1, R]$ and $k \in [1, K]$ are the time and spectral axis indexes, respectively. In fact, the possibility for nonlinearity appears only at those values of r and k in the representation where energy for more than one source exists.

If linearity is not preserved, this will have to be taken into account in the estimation and extraction processes: since overlapping energy regions introduce a certain degree of ambiguity regarding parameter estimation (see §4.12), a judgement has to be made as to what extent this ambiguity affects the desired result.

The single-channel instantaneous model of Eq. 4.10 is now presented in its equivalent form, where the mixture x and source signals $\{s_j\}_{j=1}^J$ have been transformed to a TF domain:

$$\boldsymbol{\mathcal{X}} \ = \ [\boldsymbol{\mathcal{S}}_1\,\boldsymbol{\mathcal{S}}_2\,\ldots\,\boldsymbol{\mathcal{S}}_J]^{\mathsf{T}}\,\mathbf{a}, \tag{4.13}$$

where $\boldsymbol{\mathcal{X}}$ is the TF representation of the mixture $\mathbf{x}$ (such as its STFT), while $\{\boldsymbol{\mathcal{S}}_j\}_{j=1}^J$ are the TF representations of the corresponding original source signals

[7] The first part of Eqs. 4.11 and 4.12 corresponds to the time-domain instantaneous mixture model of Eq. 4.8, where $J = 2$ and $a_1 = a_2 = 1$.

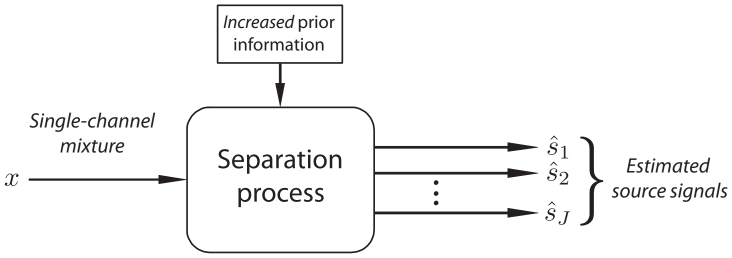


Figure 4.4: Illustration of the general form of a single-channel separation process.

$\{s_j\}_{j=1}^{J}$, in other words:

$$\boldsymbol{\mathcal{S}}_j = \begin{bmatrix} \mathcal{S}_j(1,1) & \mathcal{S}_j(1,2) & \cdots & \mathcal{S}_j(1,R) \\ \mathcal{S}_j(2,1) & \mathcal{S}_j(2,2) & \cdots & \mathcal{S}_j(2,R) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{S}_j(K,1) & \mathcal{S}_j(K,2) & \cdots & \mathcal{S}_j(K,R) \end{bmatrix} \tag{4.14}$$

4.3 General remarks on single-channel separation methods

The very fact that we have such a limited amount of audio data to work with in comparison with the rest of the cases in §4.2.1 shows the distinctive difficulty of working with mono. While multi-channel techniques have the advantage of exploiting the audio coming from different microphones to extract useful information, single-channel techniques have to rely only on one version of the mixture. This, in general, leads to a greater need for prior information to assist the algorithm (see Fig. 4.4).

Because of the high degree of complexity of this *one-to-many* problem, many techniques have been proposed from different areas of signal processing and computing. Some of them have focused, for example, on certain types of source signals, while others have concentrated on certain stages of the separation process. In general, though, there is a standard sequence of steps that is followed by all of the techniques listed below. While each technique may differ from the others in the way it proceeds at each step or by omitting a certain step completely, the sequence remains mainly the same apart from some minor exceptions discussed further below:

1. Build source models from prior learning procedures on solo recordings, or manually insert additional information.

2. Transform the mixture signal to a suitable TF representation, if needed, and incorporate additional models. If the matrix corresponding to the time-domain estimated sources is:

$$\hat{\mathbf{S}} \;=\; [\hat{\mathbf{s}}(1)\,\hat{\mathbf{s}}(2)\,\ldots\,\hat{\mathbf{s}}(T)], \tag{4.15}$$

   where $\hat{\mathbf{s}} \;=\; [\hat{s}_1(n)\,\hat{s}_2(n)\,\ldots\,\hat{s}_J(n)]^{\mathsf{T}}$, then a common example of their TF representation is their STFT:

$$\mathsf{STFT}^h_{\hat{\mathbf{S}}} = \begin{bmatrix} \mathsf{STFT}^h_{\hat{s}_1} \\ \mathsf{STFT}^h_{\hat{s}_2} \\ \vdots \\ \mathsf{STFT}^h_{\hat{s}_J} \end{bmatrix} \tag{4.16}$$

   where h is the windowing function applied to subsequent frames of the time-domain signals $\{\hat{s}_j\}_{j=1}^{J}$ (see §3.2.1).

3. Recognise structures using supervised or unsupervised techniques. Calculate the model parameters that describe these structures.

4. Extract structures from the mixture representation, or synthesise them using the calculated parameters.

5. Group structures into their respective sources using supervised or unsupervised techniques.

6. Transform each of the source signals back to the time domain, if they are not already (*i.e.*, end up with $\hat{\mathbf{S}}$).

It is noted step 6 can precede step 5, or be ignored completely (*e.g.*, in the case where a separation system targets SO applications). If step 6 is indeed part of the algorithm, what we usually end up with is the matrix corresponding to the estimated time-domain signals $\hat{\mathbf{S}}$ (if it is assumed that $\mathbf{a} = [1 \ldots 1]^{\mathsf{T}}$, or already known). Otherwise, we end up with a TF representation of every row of $\hat{\mathbf{S}}$, such as $\mathsf{STFT}^{h}_{\hat{\mathbf{s}}}$.

We will examine, now, these techniques focusing first more on steps 3 and 4.

4.3.1 Recognising source structures

Possibly the most important part in a source separation process is the *identification* of regions or shapes within the mixture representation which might correspond to a specific source (step 3 in the separation process). In order for this identification to take place, the *observation*, followed by the *interpretation* of these structures has to be carried out.

The ways source components can be observed and interpreted differs, depending on which point of view they are examined from, and why. Below is a 'taxonomy' of different ways of carrying out an interpretation of an observed signal component (which is shown in parenthesis):

- value above a predefined or pre-estimated threshold (TF point);
- peaks of a predefined or pre-estimated shape (groups of TF points) – the shape includes height (*i.e.*, magnitude);
- waveform of a predefined or pre-estimated shape (time-domain signal portion);
- predefined or pre-estimated peak structures (TF frame);
- predefined or pre-estimated peak structure pattern (peak structure extended in the time domain) (groups of TF frames).

The use of the phrase 'predefined or pre-estimated' is used repeatedly to emphasise the fact that either some degree of *a priori* explicit (*i.e.*, predefined) or model-inferred (*i.e.*, pre-estimated) knowledge is required in order for the search for source components to give meaningful results. As will be seen below, some ways of interpreting information (especially for semi-blind and blind methods) are by making use of auditory cues (AM, FM modulation, common onset/offset of partials, harmonicity of partials), statistical properties (*e.g.*, statistical independence) or more sophisticated spectral and spectrotemporal/timbral models.

4.3.2 Categorisation of source separation methods

A preliminary introduction to the way in which source separation methods can be categorised (at least broadly) according to some of their basic characteristics was carried out in §1.2. Here, the categorisation will become more explicit. This will be useful for navigating through the review of a large number of processing methods, and drawing the connections between them, necessary for establishing the current research context, as well as pinpointing where the work presented in this thesis fits in. Because of the interdisciplinarity of the field of source separation and its growing variety, it has to be noted that in no way should this categorisation and relationship between different classification methods be regarded as definitive.

A first way to distinguish between musical source separation methods is to identify whether they are related more to the field of CASA, or the field of BSS. Although in recent years these fields have been coming closer to each other more than ever, it will be seen that there are still quite clear differences in their general philosophy and it still makes sense to categorise methods in this way. Here, a broad definition of the CASA-related/inspired methods is used for facilitating categorisation of methods falling between the two categories: a method can be classified as related to or inspired by CASA if it employs any processing means that are directly or indirectly inspired by the way humans are believed to perceive and understand auditory and musical scenes. In other words, if a method includes any means of observing, interpreting and identifying source structures that is inspired by the ways in which humans do it, it will be classified as a CASA method. As a result, for example, methods that combine the use of psychoacoustic findings with techniques typically associated with BSS, will be classified as CASA-related/inspired.

The relationship between 'understanding' and 'separation' – as described in §1.2.1) – can also be of great use in differentiating between separation methods. To be specific, it highlights where the priority in a certain method is placed: the separation process, or processes that seek to infer application-specific information from the mixture. The difference between BSS-related and CASA-related/inspired methods in that respect is quite clear. Understanding generally involves the use of music or psychoacoustically specific models, this is why CASA methods are classified as UFS and SFU, while BSS ones are classified as SWU.

Another important parameter that helps to differentiate between source separation methods is the amount and type of prior information that they use. This parameter is related to the degree of *blindness* and to the degree/type of *supervision*. With regards to blindness, methods that use minimal prior information and assumptions about the sources and the mixture are classified as blind;[8] predictably this is where all BSS methods can be grouped.

At the other extreme from BSS are the non-blind methods. It can be said that these methods use an 'excessive' amount of high-level prior information. A common example of this is the supply of a MIDI-like score or the ground truth F0 tracks for assisting with source identification. In cases like these, the level of required human intervention and technical or musical expertise is usually high.

All the rest of the methods that do not belong to the two extremes regarding blindness, are the semi-blind methods. These methods make use of sufficiently generic advanced models that are 'hard-wired' into the algorithms of the system or provided through training or learning procedures. In addition, the models may or may not be offered the ability by the system to adapt appropriately to the mixture. CASA-related/inspired methods span the semi-blind/non-blind continuum: the use of advanced source/mixture models or explicit high-level information is tied,

[8] As will be seen in §4.4, blind methods do use some prior information in the form of specific assumptions about the sources and the mixture, such as such as source independence, sparsity and non-negativity.

in one way or another, to the way humans make sense of a musical scene. Finally, the degree of supervision is another useful way of distinguishing between methods, because it focuses on the amount of preparation and user intervention involved in the incorporation of models: at the one end we have supervised methods which are associated with a high amount of preparation and user intervention, while unsupervised methods represent the extreme opposite case. According to this, semi-blind methods can well be supervised or unsupervised: models learnt through learning/training techniques require a level of preparation and user intervention that 'hard-wired' models do not. Also, non-blind methods are supervised exactly because the inclusion of a considerable amount of *a priori* information usually involves a high degree of user intervention.

It is important to note, here, that a clear consensus does not yet seem to exist regarding what differentiates 'supervised' from 'unsupervised' methods in the context of source separation. In fact, the most common usage of these terms originates from the fields of machine learning and pattern recognition – fields that have shaped the directions that source separation research has been taking over the years. Within this context, some kind of prior training procedure usually takes place, that tunes the parameters of the system in question. This procedure can be characterised either as 'supervised learning' or 'unsupervised learning'. In supervised learning a set of inputs and a set of outputs is available; the learning process involves learning the mapping function from the inputs to the outputs, whose correct values are supplied by a 'teacher'. Because of the mapping process, the inputs are characterised as *labelled*. On the other hand, in unsupervised learning this 'teacher' does not exist and only the input data is available. In this case, the process of learning consists of finding statistical regularities in that input data, which, since no mapping to outputs is pre-defined, they are deemed unlabelled.[9]

By using the above definition of supervision to describe source separation systems, one can classify the ones that involve prior learning procedures based on solo segments (*e.g.*, [141, 124]) as unsupervised systems, since they involve input data

[9] For a general introduction to machine learning methods, see, *e.g.*, [4].

	S/U relationship	Blindness	Supervision
BSS-related	SWU	Blind	Unsupervised†
CASA-related/ inspired	UFS, SFU	Semi-blind	Unsupervised‡
			Supervised*
		Non-blind	Supervised**

Table 4.1: Relation between the different ways for categorising musical source separation systems. The superscripts are used for distinguishing between different approaches. †: Using basis decomposition methods, ‡: Using advanced models, ∗: Using training/learning procedures, ∗∗: Using explicit high-level prior information.

in their learning process, but not a mapping function from inputs to outputs.[10] This thesis, however, describes these systems (all systems that use prior training processes) as 'supervised', because training procedures are deemed to entail a high level of preparation and a high level of user intervention.

Finally, it is worth noting that a view on supervision that is similar to the one presented here is also expressed in [52] and [23]. Table 4.1 summarises the various categories and the ways in which they are related to each other. Following this categorisation, a review of single-channel audio and primarily musical source separation methods is carried out in detail. Also, since this thesis concentrates on pitched harmonic or near-harmonic sounds, most of the methods presented here are primarily suited for these kinds of sounds.

4.4 Blind source separation-related methods

These methods (which can be also referred to in the literature as 'unsupervised learning methods' or 'spectral decomposition methods') generally do not get inspiration from auditory models, *i.e.*, they do not try to imitate the HAS, nor use any other specific signal models. Instead, they employ data-adaptive techniques based on information-theoretic principles in order to separate meaningful structures directly from the input data, and usually without the need of any prior

[10] To the best of the knowledge of the author, no musical source separation system has been proposed so far which makes use of supervised learning techniques as commonly defined by the machine learning community.

information. This is done by factorizing the spectrogram, or some other chosen representation. The criteria used for this factorization define the main differences between the methods presented below. In all cases, an additive model is used for the mixture signal. In matrix notation this is written as:

$$\mathbf{X} \;\approx\; \mathbf{B}\,\mathbf{G}, \tag{4.17}$$

where $\mathbf{X} = [\mathbf{x}_1\,\mathbf{x}_2\,\ldots\,\mathbf{x}_R]$ is the *observation matrix*, $\mathbf{B} = [\mathbf{b}_1\,\mathbf{b}_2\,\ldots\,\mathbf{b}_N]$ is the *mixing matrix* and $\mathbf{G} = (g_{nr})_{N\times R}$ is the *gain matrix.* N is the number of the basis functions used, and R is the number of time frames. Although the estimation can be done using various representations, usually $\mathbf{X}$ denotes the magnitude spectrogram. In this context, $\mathbf{x}_r$ is the short-time DFT magnitude spectrum at time frame $r = 1, 2, \ldots, R$, and $\mathbf{b}_n$ are constant basis spectra with time-varying weights[11] g_{nr}.

Time-domain representations have also been used [77, 86, 14] for source separation. However, working with time-domain basis functions is quite tedious: it is impossible for a time-domain basis function to represent a single source: because of the highly nonstationary behaviour of phase, the time waveforms of the source are often not identical from frame to frame. Besides, as we have already mentioned (§2.2.3), a single note is a time-varying structure; hence, it can only be represented using multiple components. In general, a large number of components increases the difficulty of separation, especially when there are a relatively large number of sources to be separated.

4.4.1 Independent subspace analysis

In particular, for single-channel recordings, one of the first approaches by Casey and Westner [27] used *Independent Subspace Analysis* (ISA). The term ISA has been used to refer to techniques which apply ICA to factor the spectrogram (or any other representation) of a mono signal to separate sound sources [170]. ICA is a common technique used for blind signal separation. However, this method

[11] These weights are also called ‘gains’ or just ‘coefficients’.

normally requires determined or overdetermined mixtures [83]. A way to apply it to mono mixtures would be to assume that each frequency bin in the DFT magnitude short-time spectrum (for example) can be considered to be a different sensor [170]. In effect ISA represents the spectrum as the sum of basis spectra (which are statistically independent to each other) with time-varying weights. The separated spectra are then automatically clustered by grouping together the weight series that show the highest dependencies [27]. Molla and Hirose [117] proposed recently a method with a similar philosophy which, according to the authors, avoids some of the limitations of using the STFT, (see p. 36). As an alternative they decompose the *Hilbert spectrum*. Indeed, when compared with an STFT-based method, an improvement of separation performance was demonstrated in two-source mixes where one of the sources was always a speech signal.

4.4.2 Sparse coding

One other alternative from this family of techniques is *sparse coding*. Its goal is to build distributed representations of the mixture signal in which relatively few elements are active. One could argue that it makes sense to apply this technique to musical signals, since they appear to have similarly sparse behaviour: if we consider, for example, the majority of piano musical pieces, only up to around 6 notes are normally played simultaneously, out of the possible 88 notes of the keyboard [131]. In the model of Eq. 4.17 this restriction is applied to the weights in $\mathbf{G}$; for each time frame only a small number of weights are allowed to have a non-zero value. As a consequence, each component (a basis function together with its corresponding gain) is only active in a small number of frames. Work on separation using sparse coding has been presented, for example, by Virtanen [168], who presented some good results for the separation of drum instruments.

4.4.3 Non-negative matrix factorization

In cases where spectral magnitude is used for representing the signal (as, for example, a spectrogram) it is reasonable to set a restriction for the basis spectra and their gains to be non-negative. This restriction is used in Non-negative Matrix Factorization (NMF) algorithms. They were first introduced by Lee and Seung [104] in image processing and then in music transcription by Smaragdis and Brown [153]. Promising results have been reported for separating percussion from pitched instruments [73], simple mixtures of pitched sounds [175, 91][12] and random mixtures of pitched and drum sources [171]. In fact, the method in [171] is probably the current state-of-the-art. NMF has a downside, though: when two sources appear to be active simultaneously at all times in the mixture the algorithm will probably associate them with a single component [170]. This, often, appears in cases where the sources have an ideally static spectrum and they are perfectly synchronized (as, for example, in MIDI-synthesized mixtures). However, according to [165], NMF also produces artefacts in real mixtures, where the sources are not perfectly synchronized. Nevertheless, it is fair to say that, in general, the use of NMF for audio separation has not been tested enough, and further developments on this field could lead to more successful results.

4.5 General issues with blind source separation-related methods

In order for the above techniques to be implemented, two major assumptions are usually made about the source spectra: that (a) they are *statistically independent*,[13] and (b) they are sufficiently sparse. For the case of real musical signals, however, these assumptions do not often hold. For example, the independence assumption disregards the fact that it is common in music for instruments to play at the same tempo or with harmonically related melodies. As for the sparsity assumption, it could be less limiting, but this still depends on the degree of sparsity that can be accepted by a certain algorithm (see also §4.12.1).

[12] It is important to add, here, that the technique in [91] requires training based on solo excerpts for generating the initial basis spectra.

[13] Two events A and B with respective probabilities $\mathsf{Pr}(A)$ and $\mathsf{Pr}(B)$ are said to be statistically independent if and only if $\mathsf{Pr}(A \cap B) = \mathsf{Pr}(A)\mathsf{Pr}(B)$.

Because they do not often make explicit musical assumptions, these algorithms can be applied to a wide variety of signals. In this sense, they can deal more naturally with nonstationarities and nonlinearities, which are an integral part of real-world musical signals. For example, they do not explicitly assume that the source signals consist of notes, nor do they assume harmonicities, although they can converge to note-like or chord-like structures. Despite this flexibility, though, their performance is still currently limited to some degree for the reasons given above. As ways forward from this category of methods, alternatives have been proposed which combine unsupervised learning methods with source-specific models and techniques usually related to CASA (*e.g.*, [174, 54]).

4.6 Computational auditory scene analysis

The research field of CASA is inspired by psychological research on the way humans infer meaning from their auditory environment. Working with 2 mixture channels (originating from 2 spatially separated microphones) with the purpose of simulating binaural hearing would seem a natural thing to pursue, and this has indeed stimulated some research towards that end [43, 99, 120]. However, humans have no difficulty in separating sounds perceived by just one ear. Thus, the majority of the CASA approaches have worked on the single-channel case.

For the case of music, CASA is basically concerned with the conversion of a musical signal into higher-level musical information, such as notes, chords and rhythms using computational algorithms. For this conversion, psycho-acoustical cues are used [55]. The main ones are harmonicity, common onset and offset, amplitude and frequency modulation, and temporal and spectral proximity of partials. A three-dimensional, perceptually-motivated representation, called the *correlogram*, is often used [151], while the spectrogram is also used for simplicity of interpretation. Data-driven (*e.g.*, [115, 34, 21]) and prediction-driven [55] approaches have also been introduced, with the former being less robust than the latter, because of their inability to deal with masked auditory objects (*i.e.*, when multiple partials coexist within the same critical band). The prediction-driven approach, on the other hand, employs some advanced object models of the waveform and a blackboard model, where several competing hypotheses are tested about these objects. This approach is related to Bayesian approaches in CASA. These approaches employ probabilistic priors about the object models and they estimate how well these models fit the observed data using a likelihood function. Moreover, some of them even make effective use of musical knowledge and timbre models [90, 92]. However, while these approaches may lead to good separation results because they use these advanced timbre models, these models are still based just on solo segments. For a more detailed account on the issues involved regarding the CASA systems, the reader is referred to [176].

According to Wang and Brown, the distinction between CASA approaches and other sound separation approaches is analogous to the subtle distinction between the terms ‘computer vision’ and ‘computational vision’ in visual processing: while the former focuses more on applied image processing research, the latter deals primarily with the modelling of human vision [177, p. 29]. However, because of its ability to offer great insights into how to model the way humans separate sounds (or at least their mental representations), there are obvious benefits to employing aspects of the CASA framework in a wider range of applications. This has been done, so far, in two main strands of work:

- by regarding CASA from the viewpoint of UWS approaches[14] targeting applications within the AMT and MIR continuum (*e.g.*, [89, 68]);
- by adopting the organisational principles (grouping and segregation cues), analysis front-ends, mid-level representations, extraction and evaluation procedures used in CASA for separating audio sources, without necessarily having the actual goal of accurately describing the audio (or music) scene.

This chapter continues by reviewing source separation work that belongs mainly to the second strand.

[14] For a discussion regarding the classification of certain audio processing methods according to how they approach the interrelated concepts of ‘separation’ and ‘understanding’, see §1.2.1.

4.7 Non-blind methods

In this category of methods, the user input plays a significant part in assisting the separation process. In particular, the user is required to supply the MIDI-type score of the musical mixture as one of the system's main sources of prior information.

In an early approach, Shalom *et al.* [9] used already aligned MIDI score as the starting point for a predominant/accompaniment separation process that operates in the time domain. Strict harmonicity is assumed for the pitched part of the sources, and the accompaniment is modelled as white Gaussian noise, which although it is not a realistic assumption for a musical mixture, gave relatively good results for a few specific mixtures.

Itoyama *et al.* [85] also assume an ideally aligned score for their system that is designed as a pre-processing stage for a remixing application. Their proposed approach introduces mixture-adaptive models for both harmonic and nonharmonic source structures. The models are represented in terms of the power spectrogram and initially trained using template sounds. One of their limitations, however, is that a particular source (or notes from the same source) must be active sufficiently often for the model to be adequately learned.

Raphael [136] uses the score for a separation approach aimed at removing the dominant melody from its accompaniment. As the approach in [85], it is focused more on AQO applications (such as karaoke). He casts the problem of associating each of the TF points to either the dominant melody or the accompaniment as a classification problem (which involves a prior training procedure) and extracts the sources using binary TF masking. Some good separation results are presented on concertos, which means primarily mixtures of instrument sounds with no strongly percussive elements.

Every [58, 60] combines his method for aligning the MIDI score and the mixture with a subsequent F0 refinement process. The system is not limited to predominant/accompaniment-type separation, in other words it is more generally designed for multi-source separation. The system does not deal with strict harmonicity, and the source extraction process is carried out using adaptive spectral filtering, which, as discussed below, can be an effective way to deal with overlapping harmonics and produces a residual that is relatively free from extraction artefacts. Additionally, no prior training procedures are employed. This work forms the basis of the proposed approach, so its merits and limitations will be described in more detail below, throughout this and the next chapter.

Finally, Duan and Pardo [50] recently presented a method that uses the MIDI score to implement a single-frame source separation system. The F0 values acquired by the score (which has been aligned by an algorithm proposed in the paper) are refined with the help of a multiF0 estimator. After identifying the frequency bins that are believed to be corresponding to overlapping harmonics, their amplitudes are set to the inverse square of their harmonic number. The extraction of non-overlapping harmonics is performed assuming that their main lobe has a specific width affected only by the windowing process, and the nonharmonic part of the remaining residual is, finally, equally distributed between the sources. It is also worth noting that, since this system carries out source separation in one frame at a time, it can be employed as a real-time system.

Issues that may arise with assumptions and limitations associated with the detection and extraction parts of the separation process, such as the assumption of strict harmonicity, the ideal peak shape and types of TF masking are discussed more thoroughly in the next sections of this chapter.

4.8 MultiF0 estimation

MultiF0 estimation is a particularly useful means for the structured identification of source components (it corresponds to Step 3 of the separation process on p. 56), since what would be considered to be an important *a priori* knowledge in musical mixtures – the harmonicity assumption – is already ‘hardwired’ into the process.

F0 estimation methods (in general, both multi- and single- ones) make use of specific signal models, all of which include the required assumption of periodicity for the signal in question. The periodicity assumption is 'translated' within the frequency domain as the harmonicity assumption: if the signal is considered from the point of view of the frequency spectrum there is energy located in positions associated with specific frequency values that are integer multiples of a certain F0. It is, hence, expected that the signal model in question will be parameterised by the F0 (or, equivalently, the signal period if the method works in the time domain).

Furthermore, many of the algorithms also include – implicitly or explicitly – the realistic expectation that the F0 will be a function of time. This implies an expectation of a certain degree of temporal and spectral continuity in the signal. It also permits the inclusion of higher-level information in the model, such as instrument models and note onset/offset timings (for the case of music), or grammars (for the case of speech/singing), often presented in a probabilistic context. In this way, more complex situations, such as FM behaviour for example, can be taken into account.

Nonlinear processing is often used for enhancing cues which could point towards the correct value of the F0. This can include forcing the F0 to appear (in the case where the F0 is originally missing), or reintroducing additional harmonics that are not originally there (in the case where the method relies on inter-partial spacing for determining the F0).

For a good overview of existing multiF0 estimation approaches, the interested reader is referred to [40] and Part III of [96]. It is not the purpose of this work to design a multiF0 estimator, as this is a very difficult problem in itself. Instead, it would be enough for the separation system to be provided with a method that calculates automatically and quite reliably multiple F0s in a frame-wise basis. If the method is adequately robust, the post-processing stage of §5.4 can then hopefully correct most of the errors arising from the process. The method proposed by Klapuri [98] fits this criterion well. It will be described, along with the way it is employed by the proposed system, in §5.3.

4.9 Semi-blind unsupervised methods involving F0 estimation

Since the majority of the following approaches use varieties of the harmonic model for the source signals, one could relate them to the multiF0 estimation problem, *i.e.*, the problem of estimating the F0s of multiple source sounds which appear simultaneously in polyphonic mixtures. Indeed, many of these techniques employ (either implicitly or explicitly) a multiF0 estimation process somewhere within the separation algorithm. The inclusion of such a process is useful because the knowledge of the location of the F0s at a specific moment allows the correct identification and grouping of the spectral components (*i.e.*, peaks) belonging to the same source, through the harmonicity assumption.

However, the problem of accurate multiF0 estimation is still far from solved, and for this reason is often avoided. This task is either reduced to single-F0 estimation (a process which generally exhibits less errors), or multiF0 estimation is relied upon only partially, with the capability for refinement.

Since the use of F0 estimation carries the implication of the harmonicity assumption for an observed signal, the sinusoidal harmonic model (§3.3) has been employed by the majority of the approaches for the identification of source components. One of the early attempts of this sort, was the system by Parsons [125]. Its primary target for separation was speech sounds, and especially their harmonic part. An important feature of this system is that it deals with the problem of resolving overlapping partials. By using criteria of peak symmetry, inter-peak distance and phase stability, the peaks are first labelled as overlapping or not; then, in order to resolve the overlapping ones, it is assumed that there is enough amplitude and frequency difference between the peaks, so that the estimation of the strongest will be considered as more reliable. The strongest one is, hence, extracted first, with the remainder then extracted separately. Parameter estimation for unresolved (described as "shared") peaks is carried out via linear interpolation using information from adjacent harmonics. For the extraction, a spectral filter is used that has a fixed shape based on the windowing function and the pre-estimated FM rate of the peaks.

Another early system for the separation of two musical sources was proposed by Maher [109]. After an initial multiF0 step, a multi-strategy approach is carried out for resolving the overlapping partials, depending on their TF characteristics: either by solving a linear system of equations (assuming 'well behaved' sinusoids), by exploiting their *beating* characteristics (assuming that no significant vibrato or tremolo is exhibited by the individual sources, and that their duration is more than their beat period),[15] or by applying linear interpolation (similarly to [125]). The system is still primarily suited best to mixtures of only two sources.

Virtanen [169] proposed a system that is based on sinusoidal modelling for dealing with mixtures that not only can handle more than two sources, but also does not require one or more sources to dominate the others. First, the multiF0 estimator proposed in [97] is used for providing initial sinusoidal parameter estimates. This is followed by iteratively estimating the amplitudes and phases and refining the frequencies. The estimation of overlapping harmonics is dealt with using a linear model for the amplitudes of each harmonic structure. A variety of basis functions is used, such as polynomials, frequency-warped cosines, fixed and adaptive frequency bands. The choice for the use of a linear model for the amplitudes is an application of the spectral smoothing principle, (an ASA cue that helps build mental representations of sound sources) and it guarantees that the sum of the amplitudes of the resolved partials will equal the amplitudes of their composite peaks. The smoothness principle was also extended in terms of temporal evolution [167].

[15] When two sinusoidal components are close in frequency, they exhibit AM in the time domain; the modulation frequency is equal to their frequency difference, and this is called a beating phenomenon.

An alternative to the linear models for resolving the problem of overlapping partials is the nonlinear smoothing method first proposed in [97] and used in source separation in [173]. It is a post-processing operation (*i.e.*, taking place after identifying the individual source spectra) that applies an estimated envelope to the whole amplitude spectrum of a source to smooth out the effects of overlapping. This, however, compromises the accuracy of all the amplitude estimates; further, this simple idea of spectral smoothness often does not hold for real instrument sounds.

More recently, and as will be shown below, the attention has turned to the estimation and extraction of single sources – often assumed as predominant – from musical mixtures. The interest for this kind of mixture is fuelled by the need to directly target SO applications (*e.g.*, MIR tasks) or AQO applications (*e.g.*, karaoke). A method of this sort was proposed by Ryynänen *et al.* [143], specifically for suppressing, rather than extracting, the main (most often vocal) melody in musical material, so that it can then be used for karaoke. As with Maher and Virtanen above, sinusoidal modelling is used for both parameter estimation and resynthesis. The melody extraction method (*i.e.*, the F0 track corresponding to the melody) by [142] is used as the first step. After that, the amplitudes of the harmonic components corresponding to the estimated F0s are estimated by simply cross-correlating the windowed sinusoids at the specified frequencies with the windowed signal on a frame-by-frame basis. It is showed that their system can achieve quite significant suppression levels for a variety of accompaniment-to-vocals ratios.

A different (and more common in the literature) way of handling mixtures containing predominant melody and accompaniment, is to focus on the extraction of the predominant melody. Li and Wang [105] do that with a particular interest in the singing voice. They improve and extend a separation method by Hu and Wang [80] which originally dealt with speech. Their method includes the stages of singing voice detection (identifying the segments that are believed to contain the vocals), estimation of the F0 contour, deciding which TF units are singing dominant and extraction of those TF units. Also, unlike the methods discussed above, an auditory representation, the cochleagram [108] is used to identify the sung segments, and the extraction is carried out via binary TF masking (instead of additive sinusoidal synthesis), which involves selecting the TF units from the mixture which have been labelled as being dominated by that source.

A method with a similar philosophy, in terms of the representation and extraction process, was proposed by Hsu *et al.* [78], where the estimation of the voice pitch track is carried out using a multi-resolution STFT method by Dressler [48]. However, apart from detecting only the voiced (*i.e.*, pitched) part of the vocals, this method also includes a detection process for the unvoiced part. They report an improvement of quality in terms of voice/accompaniment separation, compared to [105], primarily because of the effective extraction of the unvoiced content.

The previous two methods use binary TF masking as the method for source extraction. As will be seen in §4.12, although this technique does not try to resolve overlapping content, it can be more robust compared to sinusoidal extraction in terms of the resulting intelligibility of the extracted sources (assuming speech or, in a stricter musical sense, singing). The work by Virtanen *et al.* [174] carries out a comparison between these two methods, in addition to proposing an improvement from binary TF masking. This improvement becomes possible by estimating and making use of a model for the accompaniment. The extraction of the melody proposed in [142] provides a binary template that defines the TF regions where the voice is present. NMF is then used on the spectrogram area that does not belong to the binary mask, to estimate the overall accompaniment spectral characteristics. When these are estimated, their spectrogram representation is subtracted from the mixture spectrogram to yield the vocal signal – hopefully isolated from interference. Thus, robust estimation of the accompaniment spectral characteristics can help to resolve overlapping content. This is one example of how the use of a spectral magnitude model of a source (here, the accompaniment is considered as one composite source) can assist in reducing its interference on another source.

As with [174], Durrieu *et al.* [54] also use pitch-based inference combined with an NMF-based model of the accompaniment in order to separate the dominant source from a musical recording. It is, in this way, an attempt to combine the framework used by semi-blind supervised methods (see §4.11) with an approach to source estimation coming from semi-blind unsupervised methods.[16] This approach parameterises the power spectral densities of the main melody and the accompaniment by using different models: a source-filter parametric model for the melody (tuned particularly for the spoken and singing voice) and a model that emphasises temporal repetition of the notes' spectra for the accompaniment. The separation involves an iterative, 2-step process: an initial model parameter estimation is followed by the estimation of the F0 contour of the dominant source, *i.e.*, the melody sequence. A second parameter estimation follows where, this time, the parameter estimation of the dominant source is refined by constraining it to follow the pre-estimated melody sequence. Finally, Wiener filters are used to separate the sources, assuming they are statistically independent from each other. Their experimental results in terms of extracting the main melody are promising: keeping in mind that this is an unsupervised method, its performance was highly comparable to a semi-blind supervised system [124]. Also, there were indications that it can perform better than a sinusoidal modelling system such as [143] which, according to the authors, is due to the use of Wiener filters. Lastly, one of the drawbacks of this method is that the F0 tracking of the leading voice is carried out using only energy cues. This means that the F0 of the dominant instrument is estimated at every time frame, even if that instrument is not the one responsible for the dominant melody.

The above five methods concentrated on extracting the predominant instrument (which is usually, but not always, the voice). In contrast to this set of works, Li and Wang [106] proposed a system whose target is the separation of two concurrent melodies of similar mean energies, played by popular musical instruments. In this scenario, source overlapping cannot be easily ignored, since it can be a more perceptually noticeable phenomenon. Also, the fact that the method in [174] can provide an improvement against binary TF masking does not necessarily mean that binary TF masking alone cannot be a sensible choice for overlapping source

[16] The method is an extension and improvement of [53], the goal of which was predominant melody extraction (an SO application), but not separation.

separation. If an estimation of the multiple F0 tracks and the spectral models of the sources from the mix alone have been robust, TF masking based on binary decisions that are informed by contextual information can lead to output signals that are perceptually acceptable. The system in [106] follows this logic. Ground-truth pitch is first used for the labelling of TF points as belonging to one of the sources, or both, and also for segmenting the mixture into note events. The amplitudes of the overlapping harmonics are estimated using the amplitude information of the non-overlapping harmonics and the assumption that note events of the same instrument within a certain time segment in the mix are spectrally similar. The last assumption can hold provided there is no large variation in pitch or dynamics taking place within the specified time segment. The final amplitude estimates are then used to inform the binary masking decisions. It is shown that their system can yield better separation results compared to the aforementioned methods by Parsons [125], Virtanen [169] and the spectral smoothing method applied in [167] (after providing the ground-truth F0 tracks).

Regarding the previous method, its goal is not to resolve the sources to the best possible degree, but to take the best binary decision for masking, *i.e.*, estimate the Ideal Binary Mask (IBM) (defined in p. 85). Also, the method cannot be extended to more than two sources. An alternative was proposed in [107], a method that seeks to actually resolve the sources and it is not limited to 2-source mixtures. As will be seen in §4.12.2, although it can be convenient, attempting to resolve overlapping peaks by working only on their amplitudes does not lead to optimal results. This is because the phases of the peaks are usually not taken into account. The method in [107] accounts for both the amplitudes and the phases, which are estimated within a least-squares framework. It is assumed that the phases of the peaks can be predicted using the knowledge of the ground-truth F0s associated with those harmonics; likewise, for the amplitudes, the assumption is that the amplitude envelopes of the associated partials are correlated with each other, *i.e.*, they exhibit common AM (an important ASA cue). As mentioned, the starting point for the identification of sources and labelling of peaks as overlapped or non-overlapped is the ground-truth F0 information. The extraction of the identified sources is carried out by overlap-add inversion of the STFT representation, after replacing the amplitudes and phases of the overlapped peaks with the newly estimated ones (the non-overlapped peaks are extracted directly via binary TF masking, *i.e.*, keeping the mixture phases).

It can be seen from the above that for predominant source estimation and extraction, usually the estimated F0 contour is enough for grouping the identified components into sources (*i.e.*, employing the harmonicity cue and the fact that the desired source usually has higher average energy compared to the accompaniment). Any methods that do not concentrate on separating just a single source, carry out the grouping process either by using simple heuristic F0 continuation rules (when $J = 2$), ground-truth F0 information, or performing it manually. Duan *et al.* [52] proposed a method that separates and groups sources from a mixture using only automatic means, while not being explicitly limited on the number of sources to extract. This method relies on the assumption that each instrument sound has a relatively *constant* harmonic structure if it is confined to a limited pitch range, and that this structure (because of its relation to the idea of timbre, see §2.2.4) is distinctive for each sound. Based on this, models for each source are learnt directly from the mixture. After a peak picking process in the magnitude STFT domain, an initial multiF0 estimation is performed in each frame. The harmonics corresponding to the detected F0s are then clustered into sources using the *NK* algorithm [185]. Next, a model called *Average Harmonic Structure (AHS)* is learned for each source. As its name suggests, AHS is the average of the magnitudes of all the harmonic structures associated with the same source in the mix. After learning the models, an F0 estimation refinement is carried out, the harmonic source components are re-identified and transformed back to the time domain. It is worth noting here that, regarding the inference of the amplitude of an overlapped harmonic, this method follows a similar philosophy to [107] and [106]: non-overlapped spectral information from different time-frames in the mix is employed as a model for this purpose. The method in [52] may however be somewhat limited compared to the other two, because of its implicit need for a source to have a 'significant presence' in the mixture: in order for the AHS model of a certain source to be learned effectively, a sufficient pitch variation of that source (*i.e.*, playing several notes) has to be taking place. Furthermore, it has to be said that the assumption of relatively constant harmonic structures excludes sources with highly unstable harmonic spectra (*e.g.*, voice) or structures that deviate from strict harmonicity (*e.g.*, piano, guitar, percussive instruments). However, these sources can be separated from the others as a by-product of the whole process, because they will remain within the residual.

Duan *et al.* compare their method to the NMF-based method by Wang and Plumbley [175]. Their separation results for 2-source and 3-source[17] mixtures show that for sources with quite stable harmonic structure, this approach outperforms the NMF method.

4.10 Semi-blind unsupervised methods not involving F0 estimation

Knowing the temporal variation of the F0 of a source in mixtures of harmonic or near-harmonic sources is very useful for assisting the identification and estimation of components belonging to that source. However, if this needs to be carried out automatically (*i.e.*, having an unsupervised system in mind), an automatic multiF0 estimation is not always error-free. Primarily for this reason, some separation methods offer alternative approaches that do not involve multiF0 estimation for searching for source structures.

Just as [105, 78, 143, 174] do, Lagrange *et al.* [100] focus on mixtures with a predominant melody, which is usually the singing voice. To be specific, their system involves the extraction of the predominant source from music recordings as an intermediate stage towards MIR tasks, such as melodic pitch extraction and voicing detection, *i.e.*, it targets SO applications.

[17] The 3-source mixtures in that paper include a vocal signal which stays in the residual after the extraction of the other 2 sources.

The method uses sinusoidal modelling as a means of mid-level representation and estimated source resynthesis. Since the F0s are not known beforehand or they have not been pre-estimated, detecting peaks which are potentially harmonically related becomes a rather complex problem. Hence, the harmonicity cue is included by introducing the *Harmonically Wrapped Peak Similarity (HWPS)* measure; this measure computes the harmonic similarity between peaks.[18] In addition to that, the ASA-inspired organisational cues of amplitude and frequency proximity are also used. It is important to note that, unlike the previous automatic methods (which usually either perform source formation and tracking in a sequential manner and in different hierarchical levels, or incorporate a predominant melody estimator to help them out) this method utilises a single stage that combines both source tracking (using temporal information) and formation (grouping of the partial tracks). This takes place in a TF region called a “texture window”. The contribution of each cue to the grouping process is reflected in the different similarity weights which are estimated between the peaks. The peaks are then clustered together using the “normalized cuts” criterion, an algorithm previously used in computer vision for image segmentation [149]. Their results show that the use of the HWPS measure along with the new clustering algorithm can lead to an improvement of separation quality when compared with two other techniques using the sinusoidal model [172, 156]. More importantly, when the system is used as an intermediate stage for predominant melody extraction and voicing detection, its performance is very promising.

This approach, however, has the downside that it is computationally intensive. For this reason Lagrange *et al.* presented an alternative, more efficient algorithm [102]. This time they consider only the HWPS cue, in conjunction with a method that selects for synthesis only the first 10 peaks which are most likely to belong to the dominant source. Although the source separation performance is lower compared to the original method, the algorithm still shows potential as a pre-processing step for MIR-related tasks, such as F0 estimation and voicing detection accuracy in a computationally efficient way.

[18] See [112] for a detailed discussion on the HWPS cue.

4.11 Semi-blind supervised methods

The methods that fall within this category make use of statistical source models that have been trained or pre-stored before source separation takes place. If the models can provide good representations of the sources that comprise the mixture, this can usually lead to an improvement in separation performance compared to unsupervised methods, since a considerable amount of source-specific information is incorporated.

One type of statistical model used in these cases is the Hidden Markov Model (HMM). Roweis [141], for example, uses solo segments of speech to train HMMs, which are combined to form a factorial HMM. This model is then used to predict the sources in a mixture of two spoken voices, generating in this way the binary masks required to reconstruct the sources (WDO is assumed). As an alternative to this approach, Benaroya *et al.* use an extension of Wiener filtering to locally stationary, non-gaussian signals [11] and nonstationary signals [10]. Specifically, for the second case, they use Gaussian Mixture Models (GMMs) for the learning process, in a Bayesian framework. Their preliminary results indicate improvement against traditional Wiener filtering.

The methods in [11] and [10] are compared by Blouet *et al.* [13] together with two other codebook strategies for speech [155, 154]. Specifically, they are assessed by their ability to separate a mixture of speech and piano. Their results indicate that the codebook strategy introduced in [11] appears to be more suitable for representing music, while the autoregressive-based model [155, 154] can, instead, capture better the features associated with speech.

In order for GMMs to describe musical signals accurately, a very large number of Gaussian functions are usually needed. This raises a variety of difficulties, such as, *e.g.*, trainability and computational complexity issues. As a way to overcome these difficulties, [124] describes a method for separating the vocals from the musical background in a statistical framework that involves Bayesian model adaptation. They propose starting with general models that are then adapted to the properties of the sources in the mix, using the mixture segments where the sources appear isolated. However, model adaptation from the mix has its own drawbacks: for an acceptable model adaptation, there still have to be long enough non-vocal segments, and the music in the non-vocal segments has to be adequately similar to the music in the vocal segments.

Most of the above methods train their models using solo segments of the sources in the mix. This can be quite restrictive when it comes to generalising a method to a variety of mixtures. An approach that attempts to challenge this restrictiveness was introduced by Vincent and Plumbley [165], and it is based on *Bayesian harmonic models*. The main difference of their system from other Bayesian approaches for the estimation of harmonic components (*e.g.*, [39]) is that this one employs (1) a perceptually motivated residual and (2) a learning procedure based on isolated notes. The prior learning is done, however, using a large enough database of notes so that the resulting model, in conjunction with harmonicity, can be considered *generic*. It is reported that their approach performs generally better than NMF.

In contrast to the philosophy of the above methods, Burred and Sikora [25] (with extensions in [23]) proposed an approach that uses the sinusoidal model for the mid-level representation and source extraction. In this sense, it bears similarities to many of the semi-blind unsupervised methods that use the sinusoidal model, or the semi-blind supervised method in [165]; however, here harmonicity is not assumed, and *a priori* learned source-specific models are included. In particular, these models describe the evolution of source spectra more accurately than, for example, the multiplication of static spectral envelopes with time-varying gains [165]. Since the assumption of harmonic sounds is not made, mixtures containing nonharmonic sources can be considered. Also, contrary to most of the other sinusoidal modelling-based methods, it does not rely on the prior knowledge or pre-estimation of the F0 contours. The method results in good separation quality in a variety of mixtures containing up to 4 notes from up to 5 instruments. One of its drawbacks, however (which is a result of the omission of the harmonicity cue), is that it cannot handle the separation of same-onset sources because they are recognised as a single source.

Since the method proposed in this thesis is not a supervised one, this section does not attempt to give a full overview of the semi-blind supervised methods. Rather, the purpose was to show some of the basic points of these methods. While they can have quite good performance in terms of separation, they still rely on prior model-learning procedures for the sources. This compromises their ability to be applied to a wide range of instrument/source types. However, by using more generic models, or models that are adapted to the mixture, the methods can be less restricting and closer to an unsupervised philosophy.

4.12 Extraction of the source estimates

Every separation system discussed here – as well as the one proposed in Ch. 5 – includes a stage where the identified structures are isolated from the rest of the original mixture to form the estimated sources. This stage is often called the extraction, synthesis, or output generation stage, and is associated with Step 4 in the general source separation framework of p. 56. Having said that, it is worth noting that there are many cases where this stage is not clearly separated from the estimation process but, rather, it is a part of it.

4.12.1 Source disjointness and signal representations in musical mixtures

The degree to which the sources overlap with each other within a chosen mixture representation is a key feature with regards to their correct estimation and extraction. Indeed, if the sources were not overlapping at all in the chosen signal domain of the mixture, then, as long as all the source components had been identified correctly, their estimation and extraction process would be relatively straightforward. However, this is an ideal situation, and for the case of typical musical mixtures is quite rare. Using multi-channel mixtures containing sources coming from distinctly different spatial positions can also increase disjointness (see below for a definition of disjointness); in this work, however, only one mixture channel is considered available, and spatial information does not exist. In this case, there is another process which can modify and help reduce the amount of source overlapping to some degree: this is the transformation of the mixture signal to different kinds of mid-level representations.

As seen in §3.2.3, sparsity is an assumption associated with designing compact representations for signals. Since signals that appear sparse in a particular domain can be described with just a few non-zero coefficients, it would be reasonable to expect that they would not overlap significantly in that domain if they were mixed together. This expectation, though, assumes that the sources do not have similar probability distributions; if they do, no matter how sparse they are, they will still overlap with each other. A (loosely) similar concept to sparsity which in a sense includes this consideration, while being more suitable for mixtures instead of monophonic signals, is *disjointness.*

The strict condition for absolute source disjointness within a mixture representation is WDO. Initial formulations of this condition were introduced by Jourjine *et al.* [87] and Yılmaz and Rickard [183]. In order for the sources (as expressed in the TF domain[19]) $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_J$ that coexist in a mixture to be W-disjoint orthogonal to each other, they have to satisfy:

$$\mathcal{S}_1(k,r)\,\mathcal{S}_2(k,r)\,\ldots\,\mathcal{S}_J(k,r) \;=\; 0, \quad \forall k, r, \tag{4.18}$$

or, equivalently,

$$\boldsymbol{\mathcal{S}}_1 \circ \boldsymbol{\mathcal{S}}_2 \circ \ldots \circ \boldsymbol{\mathcal{S}}_J \;=\; 0. \tag{4.19}$$

Eqs. 4.18 and 4.19 express the perfect disjointness situation: each point in the TF representation of the mixture is occupied by just one source. Although real-world musical mixtures almost never satisfy this equation, it would be very useful to be able to assess how close a particular mixture is to this ideal case; The validity

[19] The definition of WDO is made in the TF domain for reasons of generality. WDO can certainly be defined in the time domain, as well.

and strength of the disjointness assumption in different mixing conditions and representations can be assessed by using an appropriate measure associated with WDO. Burred [23] introduced the measure of *approximate WDO* for this reason. For a specific source j, the approximate WDO is defined as:

$$\mathsf{WDO}_j = \frac{\|\mathbf{H}_j \circ \boldsymbol{\mathcal{S}}_j\|_{\mathsf{F}}^2 - \|\mathbf{H}_j \circ \mathbf{U}_j\|_{\mathsf{F}}^2}{\|\boldsymbol{\mathcal{S}}_j\|_{\mathsf{F}}^2}, \tag{4.20}$$

where $\mathbf{H}_j$ is the ideal binary TF mask applied to the instantaneous mixture $\boldsymbol{\mathcal{X}}$ for extracting the source $\hat{\boldsymbol{\mathcal{S}}}_j$ (see Eqs. 4.23 and 4.25) and $\mathbf{U}_j$ is the sum of all the source signals in the mix that possibly interfere with source j:

$$\mathbf{U}_j = \sum_{\zeta \neq j} \boldsymbol{\mathcal{S}}_\zeta, \quad \zeta \in [1, J]. \tag{4.21}$$

Eq. 4.20 expresses the approximate WDO for source j as the normalised difference between what is called the "preserved energy" and the "interference energy". Also, as a more global measure that can characterise the whole mixture, the average WDO ($\overline{\mathsf{WDO}}$) can be defined [23]:

$$\overline{\mathsf{WDO}} = \underset{j \in [1,J]}{\mathsf{mean}} \, \mathsf{WDO}_j. \tag{4.22}$$

In the ideal situation of absolute inter-source disjointness, it is $\overline{\mathsf{WDO}} = 1$.

Experiments were then carried out to examine the effect of four different representations on $\overline{\mathsf{WDO}}$ in music and speech mixtures. For musical mixtures in particular (and in order to reflect the wide variety in polyphonic music[20]), the music corpus was split into two distinct classes: mixtures containing correlated melodies and uncorrelated ones. The division was carried out according to whether the melodies were supposed to sound musically coherent when mixed (*i.e.*, note events appearing frequently simultaneously and in consonant pitch intervals) or not, leading to an expectation of high or low spectral and temporal overlap, respectively.

[20] It is reminded that the term 'polyphonic' is defined, here, in a less strict sense from the usual one (see §4.1).

The representations under test were the pure time domain and three TF representations in varying frequency resolutions: the STFT, the Equal Rectangular Bandwidth (ERB) [119] and the *Bark* representation [187]. The TF representations were realised using 50% overlapping Hann windowing. First of all, it was shown that, as might be expected, a transformation to any of the aforementioned TF domains increases significantly the $\overline{\mathsf{WDO}}$, when compared with the time-domain representation. This is another reason that justifies working in a TF domain for the purpose of source separation. Secondly, uncorrelated music showed higher disjointness in all cases, compared with correlated music. This can justify how in many works the related assumption of statistical independence between the sources can be connected with WDO, especially when it is considered in conjunction with source sparsity.

Thirdly, the gain in disjointness achieved by frequency-warped representations (in other words, the benefit of using them), is higher for the case of correlated music and low resolutions. At the same time, however, for musical mixtures (as opposed to speech) $\overline{\mathsf{WDO}}$ increased with the number of frequency bands in all TF representations. This means that when we move to sufficiently higher resolutions in order to increase the disjointness in music mixtures, the benefit of using frequency-warped representations diminishes: the $\overline{\mathsf{WDO}}$ of the STFT and the auditory representations become statistically equivalent. According to this, the use of a 8192-point STFT as the mid-level representation of the proposed separation system (Ch. 5) is a sensible choice as a means of increasing disjointness (and, hence, reducing overlapping) between the sources.

Methods which use music-inspired assumptions have two basic options for retrieving the estimated source signals: (a) TF masking, and (b) synthesising using the estimated sinusoidal model parameters and a bank of sinusoidal oscillators.

4.12.2 Time-frequency masking

The alternative to sinusoidal synthesis for retrieving the estimated source signals is to *extract* the source estimates from the mixture. This is generally performed by what we call *TF masking.* TF masking is a type of filtering which is performed in an invertible TF domain (such as the STFT). If the complex STFT representations of the original source j and the mixture signal in a specific TF point (k, r) are $\mathcal{S}_j(k, r)$ and $\mathcal{X}(k, r)$ respectively, then the TF masking process at that point can be simply defined as:

$$\hat{\mathcal{S}}_j(k, r) \;=\; H_j(k, r)\, \mathcal{X}(k, r), \tag{4.23}$$

where $\hat{\mathcal{S}}_j(k, r)$ and $H_j(k, r)$ are the j-th extracted source and the *TF mask* used to extract that source, respectively, at a specific TF point (k, r). Eq. 4.23 can also be written in matrix form as:

$$\hat{\boldsymbol{\mathcal{S}}}_j \;=\; \mathbf{H}_j \circ \boldsymbol{\mathcal{X}}. \tag{4.24}$$

The set of masks $\{H_j\}_{j=1}^{J}$ provide weightings on the TF points according to whether, and to what degree, it is believed that any source is present at those points. They generally take real values in the interval $[0, 1]$ and they can be classified into binary (taking either the values 0 or 1) or real-valued ones.

The use of binary masks often assumes that the sources are highly disjoint from each other in the TF space so that 'hard' masking decisions do not result in estimated sources with a degree of interference that is perceptually significant. From a perceptual point of view, employing binary decisions for isolating sources from the mix is driven by the phenomenon of masking encountered in auditory perception, according to which if a sound is within a critical band from a louder sound, the first sound is rendered inaudible [118]. In fact, with regards to the field of CASA, the IBM has been proposed as its computational goal [79, 80]. For the j-th source at point (k, r), the IBM can be defined as:

$$H_j(k, r) \;=\; \begin{cases} 1, & \text{if } |\mathcal{S}_j(k, r)|^2 - \sum_{\zeta \neq j} |\mathcal{S}_\zeta(k, r)|^2 > \theta \\ 0, & \text{otherwise.} \end{cases} \tag{4.25}$$

θ is a parameter typically chosen to be 0. The problem of binary masking, thus, equates with obtaining accurate source parameter estimates (the magnitudes in particular), with the purpose of optimally allocating the TF points to the sources according to Eq. 4.25.

A binary-decision masking system is expected to be more robust than sinusoidal resynthesis against background noise and to show a tolerance against room reverberation effects [106]. Also, judging from the results obtained on musical mixtures, it can also be argued that a successful binary mask – *i.e.*, one that is close to the IBM – can capture all the important characteristics of the extracted musical sources (at least from a SO-application point of view). However, unless strict WDO is assumed, binary masking is still not really capable of recovering the original sources. As alternatives to binary masking, methods adopting real-valued masks work on partitioning the TF point energy into more than one sources, thus providing a 'soft' mask. Adaptive Wiener filtering is commonly used as a way to perform real-valued masking. The Wiener filter is actually the optimal linear filter in the minimum mean-square sense [178]. However, in order for it to be used as a ratio mask of the type:

$$H_j(k,r) \;=\; \frac{|\mathcal{S}_j(k,r)|^2}{\sum_{\zeta}^{J}|\mathcal{S}_{\zeta}(k,r)|^2}, \tag{4.26}$$

uncorrelatedness between the sources is often assumed. Sources co-existing in musical mixtures very often do exhibit correlated behaviour with each other; this can lead to a masking filter that is no longer optimal. In addition, Wiener filter-based masks work in the magnitude or power domain. If phase reconstruction considerations are not explicitly made, they do not successfully resolve the overlapping source content.

Indeed, the problem of resolving overlapping spectral peaks is the problem of estimating both the amplitudes and the phases of all the TF points associated with those peaks. This is not a straightforward task: even if the amplitude of one of two overlapping peaks is known beforehand, the amplitude of the other peak cannot be guaranteed to be estimated correctly [40]. It is the situation encountered in §4.2.3 regarding the preservation of linearity in the mixture model.

Overlapping content can potentially lead to nonlinearities during the extraction process. There are two main ways of going forward from this:

- Assume that the nonlinearities in the extraction process do not affect the desired results negatively.
- Assume that the nonlinearities in the extraction process do indeed affect the desired results negatively, but then proceed to reduce those effects.

The first case is employed by binary TF masking. Indeed, allocating a TF point to only one source means that the phase of the mixture can be used at that point, thus the mixture model linearity is preserved.[21] Real-valued masking can also use the mixture phases. In a heavy source-overlapping situation this is less justifiable compared to binary masking.

Methods following the latter approach employ phase generation or reconstruction techniques for the overlapped content. These techniques can be 'global' (working on the reconstruction of the entire signals from their magnitude or power spectra) [69, 122, 3], or just focus on specific overlapped regions employing F0 contour information for inferring the phases [107] and synthesising the rest of the signal using the mixture phases.

Unitary sum constraint

An aspect of TF masking techniques which is important for this work is whether they satisfy the unitary sum constraint [164]. In order for a TF masking technique to satisfy the unitary sum constraint, the following has to be true:

$$\sum_{j=1}^{J} H_j(k,r) \;=\; 1, \quad \forall r, k. \tag{4.27}$$

[21] This can be also seen by Eq. 4.23: the fact that masking is taking place in the complex domain (and not in the magnitude domain, for example), implies that the original mixture phases are used for reconstructing the time-domain signal.

Not all TF masking methods satisfy the unitary sum constraint; its importance for this work is that for points (k, r) where $\sum_{j=1}^{J} H_j(k, r) < 1$ the existence of the residual is implied.

Adaptive spectral filtering of the harmonics

Spectral filtering can be seen as a special case of TF masking, since estimated masks are applied on the mixture TF representation to yield the sources. Regarding the estimation of the mask, it has a different estimation philosophy than the general type of TF masks. In all other masking situations, it is a matter of estimating the values of $\{H_j(k, r)\}_{j=1}^{J}$ for all or specified TF (k, r) points using a 'global' procedure. By 'global' it is meant that every TF point where a source is believed to be present is treated equally in the estimation process. Spectral filtering, on the other hand, moves the focus to the spectral peak maxima and their shape. In particular, the estimation of the mask is a 2-step process: first only the points corresponding to the maximum amplitude of the peak are identified; second, if the peak is believed to be overlapped with another peak, an attempt to estimate their approximate shape is performed, and the energy in the shared bins is allocated accordingly to each of the sources [60].[22] This difference allows adaptive spectral filtering to be considered as a possible middle-ground between TF masking and sinusoidal modelling: while extraction takes place by multiplying the STFT with a mask, the mask itself takes values according to the estimated maximum, width and (for the case of overlaps) shape of the sinusoidal peaks. When no overlapping takes place, spectral filtering reduces to binary TF masking with no assumption of ideal peak shape.

The system by Parsons [125] is similar to this method in the sense that it uses spectral filtering for the extraction, although the filters remain fixed. The ability of the filter to adapt to the peaks is important when highly nonstationary signals (such as musical ones) are represented using a fixed-resolution representation (such as the STFT). Indeed, as was shown in [60], adaptive spectral filtering ex-

[22] Here, we refer particularly to the "Filter a" energy-based approach presented in [60], since it is the method used in the proposed system.

hibited higher performance in terms of Signal-to-Residual Ratio (SRR) on mixes of synchronous notes when compared to Parsons' method. For more detail on the adaptive spectral filtering approach, see §5.7.

On the whole, it can be argued that, when compared to conventional TF masking, adaptive spectral filtering can offer a degree of flexibility that adds to the accuracy of extraction. In addition, it offers a more robustly estimated residual compared to sinusoidal resynthesis, as will be explained in §4.12.3. The residual channel plays a key role in the system proposed in this thesis. This is why this method of source extraction will be chosen for implementation.

It has to be noted here, that, as with many other TF masking methods, the mixture phases are used for the transformation to the time domain. Although this can sometimes lead to perceptually noticeable artefacts in the output signals, these effects will have less impact if the signals are used in a remixing scenario or for feature extraction. Also, the use of triangular windows during the overlap-add reconstruction helps to reduce the effect of sudden phase changes at frame boundaries (see §3.2.1).

4.12.3 Sinusoidal synthesis and comparison to adaptive spectral filtering

Although techniques based on sinusoidal modelling (*e.g.*, [172, 173, 167, 100]) can produce fairly realistic results for a limited number of instruments and complexity of the signal, they are still restricted by the sinusoidal model itself. Using this model it is difficult to perform perfect subtraction of the partial content from the mixture. This is because of the nonstationary nature of musical signals and the inaccuracy in estimating time-varying partial parameters. As a direct consequence of the non-perfect subtraction of the partial content, the residual produced is not artefact-free.

Adaptive spectral filtering is a TF masking method that does not satisfy the unitary sum constraint and hence produces a residual. Instead of subtracting

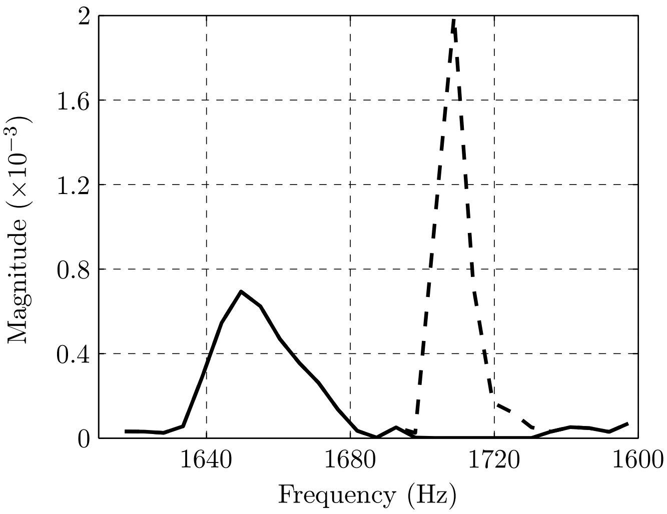


Figure 4.5: The residual before (dashed line) and after the extraction of the partial peak on the right (solid line). The main lobe of the peak has been completely removed after the application of adaptive spectral filtering.

single sinusoids, adaptive selective filters are constructed in the frequency domain that filter frequency content around the location of the note harmonics. This frequency content will correspond to the spectral peaks of a single note, that have been broadened, not only by the windowing process of the STFT and the nonconformity of the signal to the bin frequencies, but also due to the time-varying behaviour of the peaks. Unlike sinusoidal extraction, the filtering process is hence capable of extracting both a more realistically sounding note content and a residual which is relatively free from artefacts (see Fig. 4.5). Of course, a possible downside to this approach to extracting frequency content is that everything else that falls within the filtered region will be filtered as well (something that is true, also, for standard TF masking). Indeed, this happens when a high amount of broadband noise energy is present (*i.e.*, when the Signal-to-Noise Ratio (SNR) is low), which unavoidably leaks into the extracted harmonic content. As a consequence, it was reported in [58] that sinusoidal extraction leads to better results when the SNR is sufficiently low (and no overlapping harmonics are present), while spectral filtering should be preferred for the remaining cases.[23] This shows

[23] This is still subject to favourable STFT settings: the DFT length N and the hop size L have to be sufficiently small in order for the sinusoidal model to track the nonstationary behaviour of the peaks.

the complementary nature of these two approaches, something that could be exploited for improving the overall separation performance by employing multiresolution representations. However, it is not of primary concern for the proposed system to output source signals that are necessarily clean from leaked broadband noise content. Instead, it currently concentrates on making sure no remnants of the estimated content (in particular the main lobe of the partials associated with that content) has remained at the residual.

4.13 General issues with computational auditory scene analysis-inspired methods

This section summarises the main points of interest regarding the CASA-inspired methods; in this way, their similarities and differences between them and the system proposed in Ch. 5 will be highlighted.

Most of the methods, (*i.e.*, all of them apart from the ones using an auditory front-end [106, 105, 78]) use the STFT as a means of representation.[24] Parameter estimation is usually carried out on the amplitude, or magnitude domain. If overlapping content considerations are made, they are mostly limited to the estimation of the peak amplitudes; the mixture phases are often used for the estimated source synthesis/extraction, instead of explicitly attempting a phase estimation procedure (except [107], which addresses that). For reasons outlined in §4.12, the proposed approach also works on the STFT magnitude domain using the mixture phases, relying on an overlap-add step with triangular frame-windowing for dealing with abrupt phase changes at the frame boundaries.

Regarding the estimation of amplitudes of overlapping peaks, the use of various ASA and related cues (and their combination) has been adopted and is certainly helpful, to varying degrees. Since the proposed system does not have as its current priority a source extraction performance that is necessarily perceptually optimal,

[24] The methods using the sinusoidal model are included here, except [143], which works in time-domain frames.

a linear amplitude interpolation method from neighbouring harmonics is used. Despite its simplicity, this has been shown to yield good quality separation, assuming the F0s are known [60].

For the majority of methods discussed above, grouping of the estimated components into their respected sources is not needed. Since these techniques are trying either to extract a single predominant source or, broadly, to 'disentangle' mixtures of two sources, grouping of the estimated components is equivalent to source grouping. With the exception of Duan *et al.* [52], where methods do include a grouping stage, that stage is not automatic, because it is assisted by *a priori* F0 and timing information.

Many of the existing methods that use pitch-based inference are intended to employ MIDI-like score or ground-truth pitch contour information as a way to focus on stages of the separation other than F0 estimation. As this information is very strong, it enables working with considerably complex situations. On the other hand, the majority of those using an automatic F0 estimation method work primarily on predominant or 2-source separation. This is arguably a more favourable situation than using multiF0 estimates for the separation of multiple sources, since the latter is often more prone to error. The separation methods that go the 'difficult' route either reject the erroneous estimates using knowledge of the original F0s [169], or carry out refinement strategies that, although improving the separation quality, do not involve significant error checking and correction. The system presented in this thesis proposes to do that in two different ways.

Furthermore, strict harmonicity is assumed by most systems. Because of this, the partial peak energy is expected to be located near the predicted harmonic frequency locations, without always needing to verify whether the predicted frequency values correspond to observed energy peaks. Even if a search among observed peaks is being carried out, if the F0 contour information is not the ground-truth, the possibility that an observed peak may potentially be associated with more than one predicted harmonic components has to be taken into account.

The parameter estimation stage of the system of [60] makes this consideration, and this approach is also used in the system proposed here.

In addition, the existing methods concentrate, for the most part, on the separation from a musical mixture of the predominant melody, which is generally louder than the rest of the sources. Although this can often be a sensible assumption, such an amplitude difference between sources does not always hold in musical mixtures; even when it does, though, the proposed system is, in theory, not limited to the isolation of just the main melody. Dominant source extraction – a process that is part of an iterative extraction framework – is used as an option in the proposed system of Ch. 5, but is certainly not a requirement: the system still has the goal of multiple source separation, which can well be carried out jointly for all the sources.

Finally, although the systems discussed above – and in particular the semi-blind unsupervised ones – produce a residual, this is not exploited further in significant ways except from it being the channel containing an unmodelled source [52] or generally the background/accompaniment [143]. This work (Ch. 5) explores some of the various additional ways in which the residual can be exploited further.

4.14 Performance measurement and evaluation

The purpose of engineering is, in general, to propose solutions to problems under specified conditions. These solutions are considered to have, generally speaking, the form of a system, which produces an output, given a number of inputs. Analysis of the performance of such a system has to be carried out at various stages of its design, both in order to gain insight into its effectiveness in solving a certain problem[25] and to assist with decisions about further possible improvements and alterations. In order for this analysis to be carried out via the comparison of a

[25] Performance analysis is usually carried out also for gaining insight about the effectiveness of the constitutive parts of that system (assuming that the system can be broken into separate, quite autonomous sub-system elements, *i.e.*, it is based on a modular architecture).

corpus of results (which are usually the outputs of the system) with some sort of reference data, there are several basic requirements:

- Define the system, part of a system, or the process that we want to evaluate.
- Set the reference data.
- Choose the signal domain in which the measurement and evaluation will be carried out. (It may be required that the data have to be transformed to an appropriate domain, where comparison is deemed more useful.)
- Define a way to compare reference and output data.
- Decide on how to come to a conclusion about the effectiveness of the system/process by judging from the evaluation results.

It is important to clarify the distinction between the processes of *measurement* and *evaluation*. Within the context of performance analysis, they constitute two successive steps:

1. *Measurement:* Calculation of a value that indicates how close the output data is to the reference data;
2. *Evaluation:* Examination of the meaning and the significance of the measured result. This can be done by comparing it with another measured result that plays the role of the reference.

Further, there are three main kinds of reference that can be used within the evaluation step:

- *Oracle* estimates; these results set the upper bound for a certain class of algorithms that includes the system under consideration [164].
- Results from another specific, state-of-the-art system, defined as the 'reference system';

- Results from the 'best possible' performance of the system under evaluation. This provides the upper bound of the system itself.

The first problem is how to define the reference data. First of all, they will have to be of the same nature as the corpus of output data for the measurement step. The output of a system is related to the *tasks* this system is expected to carry out. In the case of source separation systems, examples of these tasks can vary from straightforward ones (*e.g.*, counting the number of sources in the mix) to more complex ones (*e.g.*, generating the musical score or a remixed version of the original recording).[26] The complexity of the task in the context of performance analysis is defined in terms of the number of parameters associated with it (the *dimensionality* of the task) and the way these parameters are associated with the success of this task. The lower the dimensionality, usually the easier it is to evaluate the performance of these tasks; for instance, if the task is to count the number of sources in the mix, the evaluation of its success would involve the comparison of an output scalar value with a reference value, most probably by calculating the scalar difference (*i.e.*, a one-dimensional operation).

The system tasks are obviously dictated by the current application that we want our system to be used for. From this point of view, it could be said that tasks dictated by SO applications exhibit lower dimensionality than the ones dictated by AQO applications (see §1.1). Also, following this type of distinction, a performance comparison of a large number of systems under a common evaluation framework could be established more easily. Towards this end, public evaluation initiatives, such as the Signal Separation Evaluation Campaign (SiSEC) [1, 161] and the Music Information Retrieval Evaluation eXchange (MIREX) [2, 47], have been introduced during the past few years; the SiSEC has been concentrating more on the AQO side of source separation, whereas the MIREX on MIR-oriented algorithms, most of which fall naturally within the category of SO applications for source separation.

[26] To be more precise, a task is also related to the *a priori* information involved and, as a consequence, the degree of system blindness [162]. However, if one can assume that this information remains the same across all systems under evaluation, a more direct connection can be drawn between the complexity of a task and the difficulty of evaluating its success.

The SiSEC, in particular, has carried out some promising steps towards establishing evaluation strategies for blind underdetermined separation systems. However, an agreed evaluation framework has not been proposed yet for single-channel audio separation, specifically regarding the task more closely related to AQO applications: the estimation of the signals $\{\hat{s}_j\}_{j=1}^{J}$.[27] As this task is a highly challenging one to evaluate, this section will concentrate on offering further insights into some of the current tools and strategies used for this purpose.

There are two main approaches to carrying out performance analysis. The first one assumes that the ground-truth is known. For the specific task of estimating $\{\hat{s}_j\}_{j=1}^{J}$, the ground-truth are the original source signals $\{s_j\}_{j=1}^{J}$. Since in 'real-world' practical situations these signals are not available, this type of performance analysis will be called *theoretically-based*, while the second case, where the ground-truth is not known, will be called *non-theoretically-based* analysis.

4.14.1 Theoretically-based performance analysis

Measurement

As discussed above, the measurement step carries out a calculation of the 'closeness' of the output data to the reference data with the help of a comparison measure. This measure can be quantitative (based on mathematical tools) or qualitative (based on perceptual notions of audio distortion and similarity). Since the quantitative measures have so far been more popular within the source separation community, the focus here will be primarily on them; the use of qualitative measures will be briefly discussed in §4.14.2.

The comparison measure typically involves the output data (the estimated source signals $\hat{s}_j$ or a specified function of them) and the reference data (the original source signals s_j or a specified function of them). Sometimes additional contextual information may be utilised (such as the mixture x), but $\hat{s}_j$ and s_j are the two

[27] Finding the correct order of the estimated signals (*i.e.*, the correct map from s_j to $\hat{s}_j$) is not considered here as part of the task. In other words, the estimation of source signals is considered to be correct up to a permutation.

core entities always involved in some way within the measurement and evaluation processes. For this reason, and for the purpose of presenting the general framework of the theoretically-based performance analysis, the comparison measure will be denoted as $\mathsf{measure}_{\mathit{ref}}(\mathit{out})$, where *ref* is the reference data and *out* is the output data. There are two main forms in which it appears in the literature:

- $\mathsf{measure}_{s_j}(\hat{s}_j)$: $\hat{s}_j$ is compared with s_j. This is the most basic way to use a comparison measure; in this case no distortions are allowed for the estimated signals.
- $\mathsf{measure}_{\tilde{s}_j}(\hat{s}_j)$: $\hat{s}_j$ is compared with $\tilde{s}_j$, where

  $$\tilde{s}_j = \mathsf{dist}(s_j) \tag{4.28}$$

  and $\mathsf{dist}(\cdot)$ is the function that represents the allowed distortions. An example of allowed distortion is the one potentially introduced by the transformation from the time domain to the TF domain and back. Li and Wang [106] account for this distortion by applying an "all-one" mask to s_j. This involves the transformation of s_j to the TF domain, followed by multiplying all the TF unit values with 1 and resynthesis (transformation back to the time domain). Another kind of allowed distortion that is employed in this way is the class of simpler, linear distortions. For example, Vincent *et al.* took into account the distortions caused by time-invariant gain, time-invariant filters, time-varying gain and time-varying filters [163].

Evaluation

- $\mathsf{measure}_{s_j}(\hat{s}_j)$ with $\mathsf{measure}_{s_j}(\breve{s}_j)$. $\breve{s}_j$ has been obtained either by an oracle estimator (making $\mathsf{measure}_{s_j}(\breve{s}_j)$ correspond to the upper bound of a certain class of algorithms), of by a reference system. The use of oracle estimators for BSS-related systems was presented in [164], while for CASA-related systems the optimal Wiener filter (*e.g.*, [54]) or the IBM (*e.g.*, [106]) have been used as an oracle: $\breve{s}_j$ are the original signals after the optimal Wiener filter or the IBM has been applied to them and they have been transformed back to the time domain.

- $\mathsf{measure}_{s_j}(\hat{s}_j)$ with $\mathsf{measure}_{s_j}(\tilde{s}_j)$. $\tilde{s}_j$ is the 'best possible' extraction, meaning that $\mathsf{measure}_{s_j}(\tilde{s}_j)$ corresponds to the upper bound performance of the system. This is useful in cases where the absolute recovery of the original source signals is not the main goal of the system and some kind of distortion is allowed. Indeed, note that $\tilde{s}_j$ is of the same type of signal as the ones subjected to allowed distortions according to Eq. 4.28 during the measurement step. The difference is that in this case the introduced distortion can often be characterised as rather more complex and nonlinear. This is because it involves the actual separation process (or parts of it) of the system under evaluation.

  This kind of evaluation is applicable especially in systems with a modular architecture (because the performance analysis can be focused on different parts of the system) and which produce a residual. The proposed system belongs to this category.

Defining the signal domain

This section considers only the use of the time-domain data for measuring and evaluating separation performance. Because of the difficulty in determining accurately the phase information in overlapping content (which could lead to a more distorted signal), it could be argued that carrying out performance analysis in, say, the spectral or TF magnitude domain would be more desirable because the phase indeterminacy errors would not be included. However, it is still not clear how significant the effect of partly erroneous phase information would be for analysing separation performance. Furthermore, AQO applications do expect to receive time-domain extracted signals, which means that their extraction quality should be studied in that particular domain. Finally, performance analysis in the time domain makes the comparison with previously reported results easier, since the vast majority have been carried out in the time domain.

Measures used in separation performance analysis

A number of basic measures that are frequently employed when analysing the performance of single-channel source separation systems will be presented below. It should be noted that in some cases the names and notation may vary between authors.

Perhaps the most common of the quantitative measures used for the analysis of single-channel source separation performance is the SRR (*e.g.*, [58, 169, 100, 107]). This is expressed as the energy ratio in dB between s_j and the difference between s_j and $\hat{s}_j$:

$$\begin{aligned} \mathsf{SRR}_{s_j}(\hat{s}_j) \;=\; \mathsf{SNR}_{\mathsf{out}j} &= 10\log_{10}\frac{\sum_n \big(s_j(n)\big)^2}{\sum_n \big(s_j(n)-\hat{s}_j(n)\big)^2}\ \text{dB} && (4.29)\\ &= 10\log_{10}\frac{\|s_j\|^2}{\|s_j-\hat{s}_j\|^2}\ \text{dB}. && (4.30)\end{aligned}$$

As Eq. 4.29 shows, the $\mathsf{SRR}_{s_j}(\hat{s}_j)$ can also be interpreted as the SNR measured at the output of the system, $\mathsf{SNR}_{\mathsf{out}}$. This is a sensible way of measuring extraction performance, since the “noise” in this case is the difference between the original and extracted signals. The smaller this difference is, compared to the original signal, the better extraction performance it will indicate. The measure does not provide us, though, with an idea of the difficulty of the problem, *i.e.*, something that would help us make a judgement about the extraction quality in proportion to the difficulty of the extraction. In order to do that, contextual information can be incorporated by employing the mixture signal x; the $\mathsf{SNR}_{\mathsf{in}}$ can, thus, be calculated:

$$\begin{aligned} \mathsf{SRR}_{s_j}(x) \;=\; \mathsf{SNR}_{\mathsf{in}j} &= 10\log_{10}\frac{\sum_n \big(s_j(n)\big)^2}{\sum_n \big(x(n)-s_j(n)\big)^2}\ \text{dB} && (4.31)\\ &= 10\log_{10}\frac{\|s_j\|^2}{\|x-s_j\|^2}\ \text{dB} && (4.32)\end{aligned}$$

and a new measure can be defined by taking the difference of the SNRs in Eqs. 4.29 and 4.31:

$$\Delta\mathsf{SNR} \;=\; \mathsf{SNR}_{\mathsf{out}} \;-\; \mathsf{SNR}_{\mathsf{in}}. \qquad (4.33)$$

$\Delta\mathsf{SNR}$ compares the output not just with the original source, but also with the mixture as well ($\mathsf{SNR}_{\mathsf{in}}$). This is an important consideration when dealing with mixtures in which the sources are mixed with different energy ratios. Intuitively, the lower the interference compared to a particular source, the easier the extraction of that source would be. Using $\mathsf{SNR}_{\mathsf{in}}$ is a way of incorporating in the measure the degree of difficulty that the mixture poses on the extraction of a particular source.

It is worth noting that what is referred to as the 'residual' here is in some ways different from the usage of the word elsewhere within this thesis. In this section the residual is defined for the purpose of performance analysis and can take the forms of $s_j - \hat{s}_j$ or $x - s_j$; this implies that s_j is available. Throughout the thesis, however, s_j is not considered to be available, and the residual has a somewhat different purpose: it is generally defined as the remainder after the extraction of $\hat{s}_j$ from the mixture x, in other words, $x - \hat{s}_j$ (while an alternative, more enhanced version of it is introduced in §6.1).

The $\mathsf{SRR}_{s_j}(\hat{s}_j)$ is a measure that takes into account all the possible distortions and errors introduced into $\hat{s}_j$ (for example, inter-source interference, sensor noise, artefacts due to the extraction process, or the transformation back to the time domain). A more educated study of the performance would involve a closer look at each of the different types of errors or distortions and their role in the separation performance. Vincent *et al.* [163] proposed a number of alternative measures with that aim in mind. They decompose $\hat{s}_j$ as

$$\hat{s}_j = s_{\mathsf{target}} + \underbrace{e_{\mathsf{interf}} + e_{\mathsf{artef}}}_{e} \tag{4.34}$$

where s_{target} is the reference signal calculated as the orthogonal projection of $\hat{s}_j$ onto s_j, *i.e.*,

$$s_{\mathsf{target}} = \frac{\langle \hat{s}_j, s_j \rangle}{\|s_j\|^2} s_j, \tag{4.35}$$

and e is the total 'residual', which is broken into two parts: e_{interf} and e_{artef} are, respectively, the results of source interference and extraction artefacts.[28] e_{interf} is calculated using

$$e_{\mathsf{interf}} \;=\; \mathbf{S}^{\mathsf{T}}\,\mathfrak{c} \;-\; s_{\mathsf{target}} \tag{4.36}$$

where

$$\mathfrak{c} \;=\; \mathfrak{G}^{-1}\left[\langle \hat{s}_j, s_1\rangle\;\langle \hat{s}_j, s_2\rangle\;\ldots\;\langle \hat{s}_j, s_J\rangle\right]^{\mathsf{T}} \tag{4.37}$$

and $\mathfrak{G} = (\mathfrak{g}_{j\zeta})_{J\times J}$ is the Gram matrix of the source signals with elements:

$$\mathfrak{g}_{j\zeta} \;=\; \langle s_j, s_\zeta\rangle. \tag{4.38}$$

From Eqs. 4.34 and 4.36, e_{artef} can then be obtained:

$$e_{\mathsf{artef}} \;=\; \hat{s}_j \;-\; \mathbf{S}^{\mathsf{T}}\,\mathfrak{c}. \tag{4.39}$$

Using the quantities s_{target}, e_{interf} and e_{artef}, three measures can be defined. The SDR is the 'global' measure, taking into account all possible kinds of distortion:

$$\mathsf{SDR}_j \;=\; 10\log_{10}\frac{\|s_{\mathsf{target}}\|^2}{\|e_{\mathsf{interf}}+e_{\mathsf{artef}}\|^2}\;\text{dB}, \tag{4.40}$$

while the Signal-to-Interference Ratio (SIR) concentrates on the effect of source interference:

$$\mathsf{SIR}_j \;=\; 10\log_{10}\frac{\|s_{\mathsf{target}}\|^2}{\|e_{\mathsf{interf}}\|^2}\;\text{dB}, \tag{4.41}$$

and the Signal-to-Artifacts Ratio (SAR) on the effect of the estimation and extraction artefacts:

$$\mathsf{SAR}_j \;=\; 10\log_{10}\frac{\|s_{\mathsf{target}}+e_{\mathsf{interf}}\|^2}{\|e_{\mathsf{artef}}\|^2}\;\text{dB}. \tag{4.42}$$

The SIR and the SAR provide the means of acquiring additional insight into the performance of a particular system in terms of source separation. Further, the SDR has a particular importance because it involves the overall distortion, hence

[28] A third part, e_{noise}, corresponding to the additive sensor noise, is also included in [163]. In this thesis, however, the employed mixing model is the instantaneous one of Eq. 4.5 which ignores this noise element.

offering a ‘summary’ of the ability of that system to extract a source signal from a mixture to an acceptable degree. In this sense, it belongs in the same ‘category’ as the $\mathsf{SRR}_{s_j}(\hat{s}_j)$. However, there does not seem to be a clear agreement, particularly within the single-channel musical source separation community, regarding the use of any one particular measure. For example, the SiSEC [1, 161], which is a positive step towards a community-based agreement for an evaluation framework, has not yet introduced proposals for the single-channel case. Hence, one priority is to study the differences between these measures in terms of their behaviour with various forms and amounts of distortion, so that a stronger judgement can be made about their effectiveness. A contribution to this study is offered in §5.8.1, via a combination of two analysis frameworks, a theoretical one and a practical one. It is noted that the practical framework does not belong to the non-theoretically-based performance analysis discussed next. This is because, unlike the non-theoretically-based type of analysis, the original sources are considered available in the practical framework.

4.14.2 Non-theoretically-based performance analysis

This type of analysis does not assume the availability of the original sources, so the reference data will have to be obtained using a reference system. (In a sense, the fact that the system is regarded as ‘reference’ automatically renders its output as reference data.) This system has to share a common framework (*i.e.*, common or similar assumptions and conditions of operation) with the system under test. The measurement stage involves the calculation of $\mathsf{measure}_{s'_j}(\hat{s}_j)$, where s'_j is the reference data. Because of the unavailability of s_j, this type of analysis does not include an evaluation stage. Separation systems can still be compared, though, through a direct comparison of their respective $\mathsf{measure}_{s'_j}(\hat{s}_j)$. Having said that, it does not seem that this method of performance analysis has been preferred by researchers so far.

A different way of comparing signals is by carrying it out qualitatively: the comparison measure is a perceptual one (*i.e.*, related to human perception). This can be particularly useful when the original sources are not available. Although perceptual measures have been proposed for the quality assessment of coded audio and speech (see, *e.g.*, [6, 7]) they have specifically focused on types of distortion errors of a different nature to those often encountered in ASS; using such perceptual measures as a way to evaluate separation quality could, hence, potentially give misleading results. However, Fox *et al.* [64] showed that the linear combination of four quantitative measures (including the SIR and the SAR discussed above) could correlate highly with similarity assessments carried out by humans, at least when tested on woodwind instrument 2-channel mixtures processed by BSS-related algorithms. Further study of this type of comparison on a wider variety of separated music and/or speech signals would be needed in order to devise more robust measures that correlate well with assessment by humans.

4.15 Summary

This chapter introduced the problem of single-channel source separation from mixtures of music sounds and carried out a review of the methods that have addressed this problem so far. After clarifying a few of the related terms often used in the literature and describing the variety of mixing models, we focus on the instantaneous mixtures. This is followed by outlining the main stages of a separation system and introducing a classification framework for the existing methods.

The methods can be generally classified as BSS-related or CASA-related/inspired ones. Their general difference lies in the use of human perception-associated models: the first category does not use them, while the second one does. This is why CASA-related/inspired methods are not blind, while the BSS-related ones obviously are. Also, an additional way to classify them is by the degree of supervision (*i.e.*, the degree of human intervention). The BSS-related methods have the advantage of being always unsupervised and they are usually not restricted on the type of mixtures. However, their often-applied assumptions of statistical independence and high sparsity can be unrealistic for real-world musical mixtures.

On the other hand, the methods related to or inspired by CASA can incorporate advanced models for the sources, or models of the hearing and scene analysis processes. This enables more flexibility but also increased challenges, such as how to combine those models and processes effectively. The methods belonging to this category can be supervised or unsupervised. The supervised methods make use of learning procedures before the separation in order to construct statistical source models. While satisfactory results are reported, the models are usually quite restricted in the type of musical sources they can accommodate; in addition, most of the time, appropriate training material encompassing a variety of musical sounds is not readily available.

Unsupervised CASA-inspired methods do not involve restrictive training procedures. Instead, they are based on more generic models (such as the harmonic model) and psychoacoustic cues. Most of the current methods use pitch-based inference for the extraction of two sources or the predominant one. They usually assume strict harmonicity and ideal peak shape. The proposed method does not make those assumptions and is not limited to the extraction of just one or two sources.

Also, with regards to non-blind supervised methods, because of their MIDI front-end, they rely heavily on user intervention and expertise. These can be very restricting characteristics for a system, and that is why it was chosen to replace the MIDI front-end in Every's system with an automatic stage in Ch. 5.

The chapter continues with a discussion on source overlapping, its relation to sparsity and disjointness, and how this is dealt with by various extraction methods. It is concluded that for the case of musical mixtures, the choice of a 8192-point STFT (assuming a sampling frequency of 44.1 kHz) is a reasonable one. A variety of TF masking methods is then discussed, ending with the adaptive spectral filtering technique that is used in the proposed system, amongst other reasons because it is a balanced alternative between binary TF masking and sinusoidal modelling. Particularly when compared to sinusoidal modelling, the spectral filtering method is less likely to produce a residual with considerable extraction artefacts, since no ideal peak shape is assumed, and can tolerate estimation inaccuracies that the sinusoidal model would not. Finally, there is no method that uses the residual as an enhanced source of information in the way it is used in this thesis.

The final part of the chapter gave an introduction on the different ways available for carrying out performance analysis for source separation systems. Normally this analysis takes place in two stages, measurement and evaluation. The first stage uses an available measure for comparing the extracted signals with some reference (the original signals or a function of them). The evaluation stage compares the measured value with a measurement made for a reference system or a best-case scenario of the system under question.

The next chapter describes the proposed source separation approach in detail.

CHAPTER 5

One-pass proposed approach

The previous chapter reviewed a variety of approaches towards the separation of musical sources from single-channel polyphonic recordings, and established a context for the proposed approach to this problem. This chapter provides an overview of the whole system and introduces some of the definitions needed for the purposes of description and analysis. Then a one-pass implementation of the proposed approach is discussed in detail.

The extensions, and improvements relating to the move to a more automatic front-end are described and experiments involving mixtures of varying complexity are carried out. The analysis of the performance results reveals the viability of employing the automatic front-end versus its MIDI-based alternative.

5.1 Overview of the complete proposed approach

One of the main goals of this work is the design of a system that performs source separation from single-channel musical mixtures. Although being a highly underdetermined task that would require significant prior information, depending on its intended application, it would still be desirable for such a system to be automatic. Musicality constraints imposed on the sounds and mixtures have been helpful in such cases, while retaining a useful degree of generality. Here, the basic musicality constraints are the assumptions that the signals to be identified and extracted can be characterised adequately by a sinusoidal harmonic model, along with the fact that the mixtures contain detectable musical events. It should be emphasised, however, that no strict harmonicity assumption is imposed.

The proposed system uses an existing one [60, 58] as its starting point. The particular system was selected for two primary reasons:

- it showed higher separation performance compared to other methods on a variety of musical mixes containing synchronous single-note sources [60, 58];
- because of its particular extraction stage (an adaptive spectral filtering method applied to the harmonic or near-harmonic content), it produces a residual that can be relatively free from extraction artefacts (*i.e.*, no main-lobe spectral energy of previously estimated content is expected to be found in the residual, see Fig. 4.5). This opens up a number of possibilities for the use of the residual, as will be also explained below.

The previous system made use of prior pitch and timing information in the form of of a user-supplied MIDI-type score as a first basic step for identifying the source structures. However, this implies the need of a trained musician, the medium through which this information will be inserted. By replacing the MIDI front-end with an automatic method of providing the F0 tracks (and in some cases the timings as well) the need for a trained musician is removed. At the same time, the proposed work offers an improved accessibility for the end user – in a sense it provides an advancement towards a new tool which people from the wider music technology community (engineers/scientists, but also people with a stronger emphasis on the music/creative side) can experiment with and enjoy the benefit of.

The approach presented here pays particular attention to the residual channel and its potential in providing information within an iterative framework, with the purpose of realising a SAU system. One of the assumptions related to this idea is that, if the identification and extraction step are robust enough and relatively

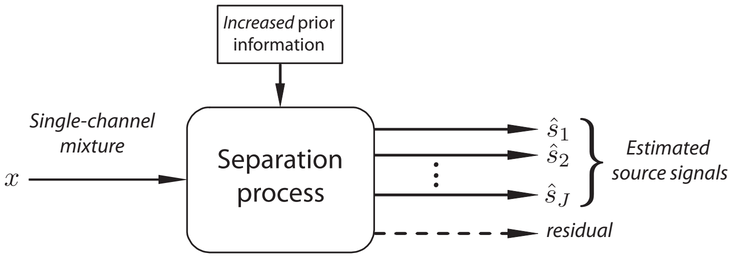


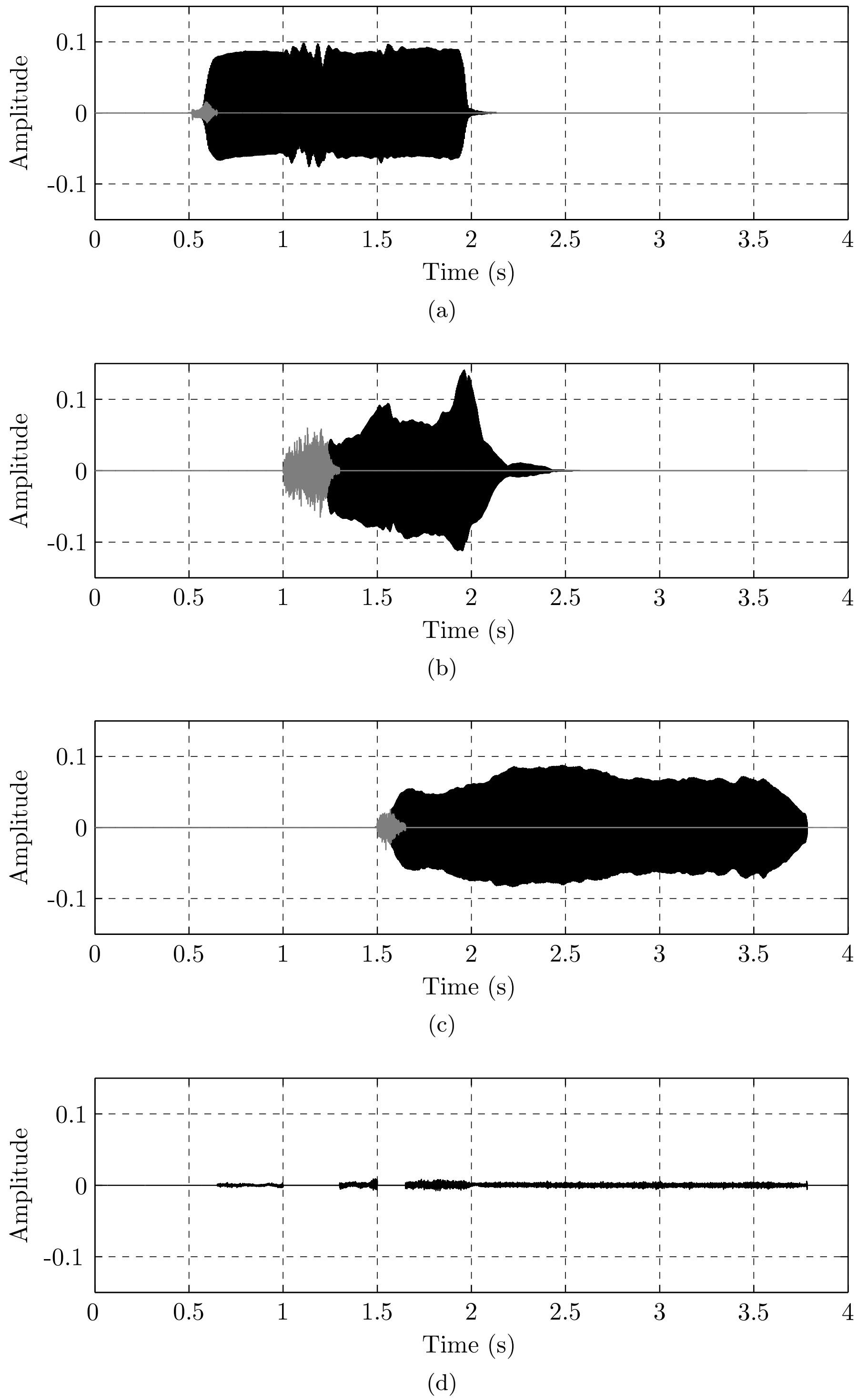
Figure 5.1: Single-channel source separation that produces a residual channel.

artefact-free, the combination of a multiF0 estimator with a residual feedback loop can lead to a robust source separation system (see §6.1 and §6.2).

The residual signal can take multiple forms, according to its different roles. In fact, the very concept of the residual allows for the partitioning of the single-channel mixture into rather more signals than there are sources. For example, in the most common case, if the number of original sources is J, the basic separation system can produce $J + 1$ signals: J extracted source signals, and a residual (see Fig. 5.1). But since the initial residual contains the unmodelled content belonging potentially to all of the source signals, this content may be further separated – potentially into J separate attacks and the remaining unmodelled content. Hence, most generally, there can potentially be up to $2J$ signals associated with note events, resulting in $2J + 1$ output channels, including the 'final residual' (see Fig. 5.2).

The focus of this thesis is primarily in the separation of pitched (harmonic and near-harmonic) content. Also, because of the use of the multiF0 estimator, the system shows its strong reliance on the harmonicity psychoacoustic cue; this is the main reason why it can be classified as a CASA-inspired approach. Having said that, nonharmonic content will still be considered here but mainly as a source of information. Specifically, nonharmonic energy existing in the residual and associated with the note onsets will be used for devising a residual-based note onset detector (see §6.4).

There is no actual restriction regarding the number of sources. In addition, although parts of the presented algorithms might benefit by the predominance of



(a)

(b)

(c)

(d)

Figure 5.2: An explanation/illustration of the fact that the proposed system can produce up to $2J + 1$ output signals. The original mixture contains three notes ($J = 3$), appearing in 0.5 s intervals with the following order: soprano saxophone A3, cello F4 and violin E4. (a)-(c) show the J extracted harmonic parts (in black) and J nonharmonic parts (in grey); (d) shows the remaining signal, the 'final residual'.

	System by Every	Proposed approach
Degree of blindness	Non-blind	Semi-blind
Improved accesibility	×	✓
Iterative use of the residual	×	✓
Iterative extraction	×	✓
Spectral peak picking and estimation	as in [60, 58]	same
Source parameter estimation	as in [60, 58]	improved
Extraction	as in [60, 58]	same
Residual-based onset detection method	×	✓
Near real-time	×	×

Table 5.1: A summary of the main differences of the proposed approach to, and commonalities with Every's approach [60, 58]. The distinction "same/improved" refers to the proposed approach as compared to Every's system.

certain sources over others, they are certainly not limited to this particular case. Finally, because of the multi-goal nature of this system, an implementation that is fast, or even close to real-time, is not of a primary concern here. Of course, by choosing to focus on more restrictive goals, a faster system is certainly possible. Table 5.1 provides a summary of the main differences of the proposed approach to, and commonalities with Every's approach [60, 58].

5.2 Some additional definitions

As discussed in §4.3.1, every source separation method incorporates its own way(s) of identifying, organising and allocating mixture energy that may belong to distinct sources. This organisation is often carried out in a number of steps, effectively in a *hierarchical* manner. Hierarchical considerations that are imposed on a signal are useful in two ways:

- By enabling a breakdown of the problem into distinct processing steps related to the hierarchical levels. In this manner, a systematic examination of the variety of challenges related to different types of mixtures can be carried out.

- By enabling a methodical comparison between methods that use similar hierarchical organisations.

In source separation, the chosen hierarchical organisation depends on the way the mixtures that we are interested in separating vary from each other. Here, the difference between the types of mixtures lies in the place that they occupy in the homophony/polyphony continuum.

Since the focus of this thesis is on the treatment of mainly Western-type musical signals, the use of a MIDI description can be convenient for describing levels in the grouping process of already identified musical structures. Although the MIDI description is probably not the only way to do this, it fits with the philosophy of the proposed approach and its popularity permits a more general consideration of other methods as well. Fig. 5.3a shows a 'pianoroll-like' representation of a musical piece where three interweaving melodies are played simultaneously by three instruments. The musical structures are MIDI note events with associated pitch values which vary with time, and time is divided into STFT frames. The grouping of these structures is expressed in this thesis in terms of three factors:

- J, the number of source signals, as they are defined in §4.1.4.
- P, the total number of note events appearing in the total mixture. It can be defined as:

  $$P = \sum_{j=1}^{J} |P_j|, \tag{5.1}$$

  where P_j is the set of indices $\{p_j \in [1, P]\}$ that show which of the notes belong to source j.
- O, a quantity which can be called the *polyphony*. It is generally a function of time, because it corresponds to the number of different identifiable structures in each time instant, and since these structures belong to different note events, O cannot be larger than P.

When time is expressed in terms of the STFT frames, the relationship between the three quantities can be expressed as:

$$0 \leqslant O_r \leqslant J \leqslant P, \quad \forall r, \tag{5.2}$$

where $r = 1, 2, \ldots, R$ is the index of the STFT frames. The above equation makes clear that the present work does not deal with synchronous notes coming from the same source, *i.e.*, chords are not assumed here. Fig. 5.3b shows the variation of O_r for the mixture in Fig. 5.3a. There are four cases of particular interest, with regards to the relationship between J, P and O_r, shown here in order of complexity from low to high:

Case 1: $O_r = J$, $\forall r$ and $P = J$. In this case every source contains a single note event, *i.e.*,

$$|P_j| = 1, \quad \forall j, \tag{5.3}$$

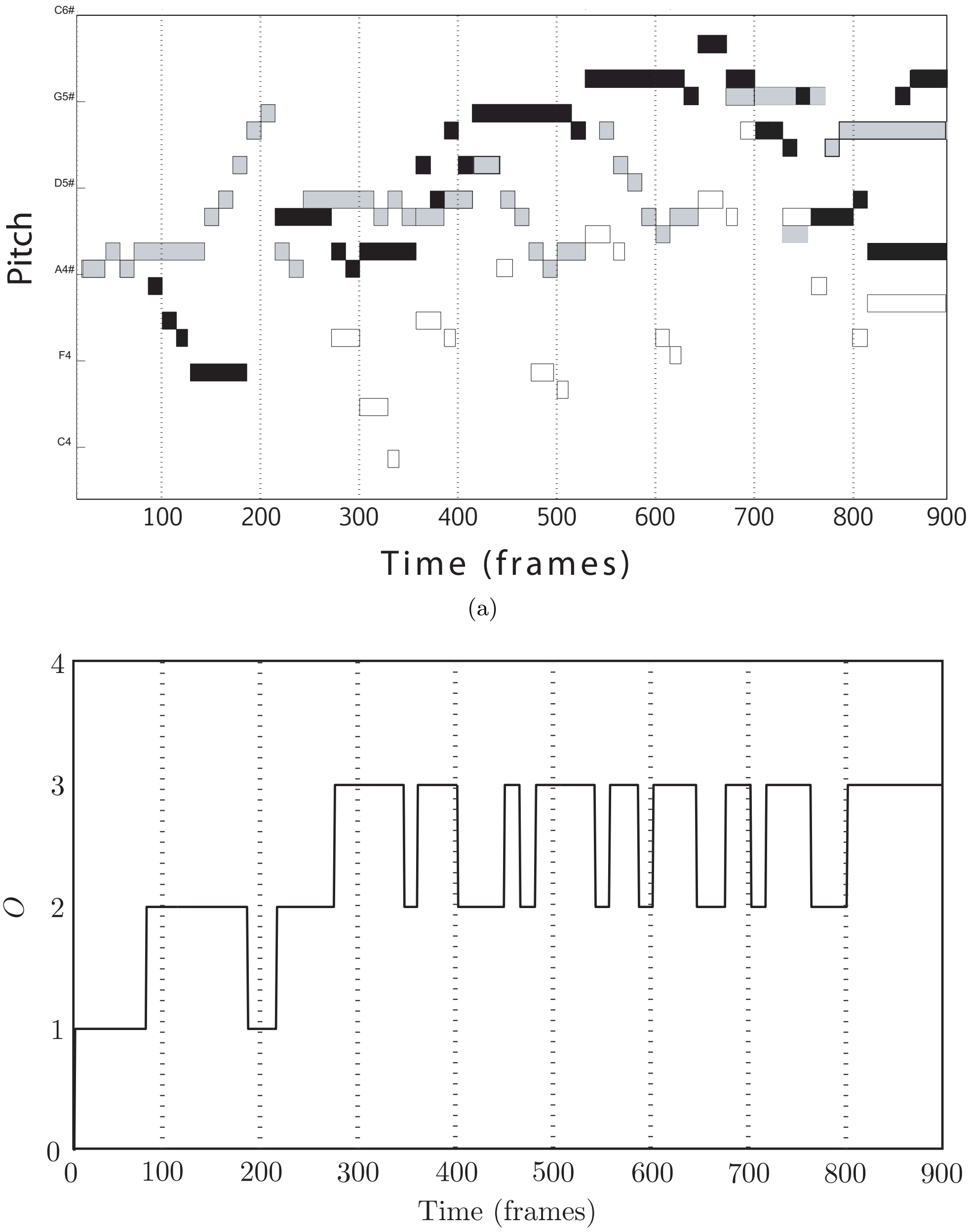
and these note events are synchronous and of the same duration (see the time-frame intervals $[240, 270]$ and $[870, 900]$ in Fig. 5.3).

Case 2: $O_r = J$, $\forall r$ and $P > J$. This happens when we have all the melodies that exist simultaneously in every time instant (see the time-frame interval $[800, 900]$ in Fig. 5.3).

Case 3: O_r is not constant and $P = J$. As with the first case, every source contains a single note event, but these note events are not necessarily synchronous or of the same duration (see the time-frame interval $[70, 100]$ in Fig. 5.3, supposing that $J = 2$).

Case 4: O_r is not constant and $P > J$. This is when there is no restriction regarding the duration and quantity of note events in relation to J, as well as their evolution in time and whether they appear in every time instant or not (see Fig. 5.3).

Fig. 5.4 shows the assumed hierarchical organisation. The formation of complete sources – which, because of the particular hierarchical organisation, is achieved by



(a)

(b)

Figure 5.3: (a): A pianoroll representation of a musical piece excerpt, where three distinct melodies coexist (shown in white, grey and black colour). The melodies appear as sequences of note events. (b): the polyphony O of the piece shown in (a) as a function of the time.

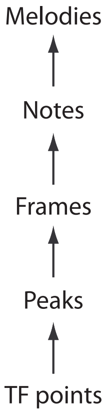


Figure 5.4: The hierarchical organisation of source structures moving from the lowest (TF points) to the highest (melodies) level. In this sense, peaks are considered as groups of TF points, the frames as groups of peaks, the notes as groups of frames and the melodies as groups of notes. Here, the melody represents a complete high-level source representation.

the grouping of extracted note events – will have to be preceded by the formation of note events (see Fig. 5.4).

This work operates at a level between the frames and the notes which can be called *sustained note interval.* These kinds of intervals are expected to be containing only sustained parts of notes. For Case 1, the sustained note interval is the same as the length of the synchronous notes (and the whole mixture); for the rest of the cases, the mixture is segmented into a series of sustained note intervals, using the note onset/offset timing information, and each interval is processed individually. No automatic tracking or grouping process is included in this system when melodies are concerned, so the final step is to ideally group the extracted sources. By *ideal grouping* in this thesis is meant that the extracted signals are compared on a windowed frame-by-frame basis with the original sources, and thus rearranged according to the best-possible matches of these comparisons.

The rest of this chapter goes through a detailed analysis and discussion of the one-pass version of the proposed separation system.

5.3 MultiF0 estimation stage

The multiF0 estimation algorithm used in the proposed method was proposed by Klapuri, originally in [98] and in an extended version that employs an auditory model as a front-end in [95]. This frame-based method is employed here because its good performance with multiple sources makes it a pretty solid starting point for an automatic separation system that separates accurately multiple sources. Provided the note event timings can be known and the method is correct for the majority of the time frames, the F0 track disentangling operation can help to correct any possible errors. The rest of this section will summarise Klapuri's multiF0 estimation algorithm. The parameter values are the ones suggested by Klapuri.

All of the processes outlined below take place in a single Hamming-windowed frame r. First of all, the DFT of the frame is taken, followed by an application of a bandpass filterbank. This is carried out through the use of C filters with centre frequencies of the form:

$$f_c \;=\; 229 \cdot \left(10^{\frac{0.39c+2.3}{21.4}} - 1\right), \tag{5.4}$$

where $c = 0, 1, \ldots, C-1$ and $C = 70$. Each subband has a triangular power response $H_c(k)$ extending from f_{c-1} to f_{c+1} and zero elsewhere. After applying the subband response to the power spectrum of the mixture, the standard deviations σ_c are calculated:

$$\sigma_c \;=\; \left(\frac{1}{N}\sum_k H_c(k)\,|\mathcal{X}(k)|^2\right)^{1/2}, \tag{5.5}$$

where N is the length of the DFT. The compression coefficients can be then defined as:

$$\gamma_c = \sigma_c^{\nu-1}, \tag{5.6}$$

where ν controls the amount of spectral whitening. Here, we use $\nu = 0.33$. Finally, a compression function $\gamma(k)$ is derived for all the frequency bins by linearly interpolating between the values $\{\gamma_c\}_{c=0}^{C-1}$. The whitened mixture spectrum can be then estimated as:

$$\mathcal{Y}(k) \;=\; \gamma(k)\,\mathcal{X}(k). \tag{5.7}$$

Source detection is carried out by either an 'iterative detection and cancellation' technique, or joint estimation. Here, the iterative method has been used. Klapuri showed that both of them perform equally well in multiF0 estimation, with the iterative method being favoured because of its lower computational complexity. The iteration procedure operates by first detecting the most prominent period, cancelling it (*i.e.*, subtracting the peaks corresponding to this period from the overall magnitude spectrum), and proceeding to detect the next prominent period. In order to determine the prominence of a certain period, the measure of *salience* $\lambda(\tau)$ is defined:

$$\lambda(\tau) \;=\; \sum_{m=1}^{M} w(\tau,m) \max_{k\in\kappa_{\tau,m}} |\mathcal{Y}(k)|, \tag{5.8}$$

where $w(\tau,m)$ is a weighting function and $\kappa_{\tau,m}$ is the set of frequency bins around the m-th partial of the F0 candidate that has frequency F_s/τ. (F_s is the sampling rate.) Finally, $w(\tau,m)$ has to be optimised for minimising the multiF0 estimation error rate. The behaviour of two factorised forms of the function was studied on a database of training material consisting of random mixtures of sounds and varying polyphony. According to the observations it was decided that the weighting function should take the following form (see [98] for more details on the procedure):

$$w(\tau,m) \;=\; \frac{F_s/\tau+\alpha}{mF_s/\tau+\beta}, \tag{5.9}$$

where $\alpha = 52\,\text{Hz}$ and $\beta = 320\,\text{Hz}$ for a 93 ms frame.[1]

Regarding the use of this method in the proposed system, two final points need to be made: (1) the polyphony value O_r has to be provided as prior information; (2) the method assumes that the sounds we are dealing with have F0s between 40 Hz and 2.1 kHz. These frequency limits correspond to the lowest and highest detectable musical notes of the total separation system, E1 and C7 respectively.

[1] The same values where also used for the current implementation that uses 186 ms frames.

5.4 F0 track disentangling

Since the F0 estimates are produced in a frame-by-frame basis by the multiF0 estimator of the previous stage, no tracking or sorting process is involved. Assuming synchronous single-note sources of the same length (Case 1-type mixtures), or that the note timings are known (for the rest of the mixture cases), the stage of F0 track disentangling receives as input the frame-wise F0 values and sorts them into separate tracks; at the same time an error correction process is carried out.

For each Hamming-windowed frame $r \in [1, R]$, the multiF0 algorithm calculates two quantities: the F0s in Hz $\{f_0^{(p)}\}_{p=1}^{P}$ and the salience function $\{\lambda(f_0^{(p)})\}_{p=1}^{P}$. These quantities correspond to the two $R \times P$ matrices:

$$\mathbf{F} \;=\; [\mathbf{f}^{(1)}\;\mathbf{f}^{(2)}\;\ldots\;\mathbf{f}^{(P)}] \tag{5.10}$$

and

$$\mathbf{\Lambda} \;=\; [\lambda(\mathbf{f}^{(1)})\;\lambda(\mathbf{f}^{(2)})\;\ldots\;\lambda(\mathbf{f}^{(P)})], \tag{5.11}$$

respectively, where

$$\mathbf{f}^{(p)} \;=\; [f_{01}^{(p)}\;f_{02}^{(p)}\;\ldots\;f_{0R}^{(p)}]^{\mathsf{T}}, \tag{5.12}$$

and it represents the p-th F0 track. For every frame r, the row elements $\{\lambda(f_{0r}^{(p)}) \equiv \lambda_r^{(p)}\}_{p=1}^{P}$ are arranged from left to right in decreasing order of salience. In this way, the F0s are allocated to their respective sources; this, however, operates under a best case scenario.

The best case scenario regarding the estimation of the F0 tracks is when the F0s have been estimated correctly, and have also been grouped into tracks using only λ as a distinguishing feature. As can be seen in Eq. 5.8, the magnitudes of the harmonic partials are directly proportional to λ. An effective use of λ can thus be enabled for F0 tracking in cases where there is a sufficient difference between source energies in the mix. This can be quite safely assumed when one of the sources is expected to always be the predominant one in the mix, particularly in a 2-source scenario (predominant melody/background, see Fig. 5.5a). Also, since

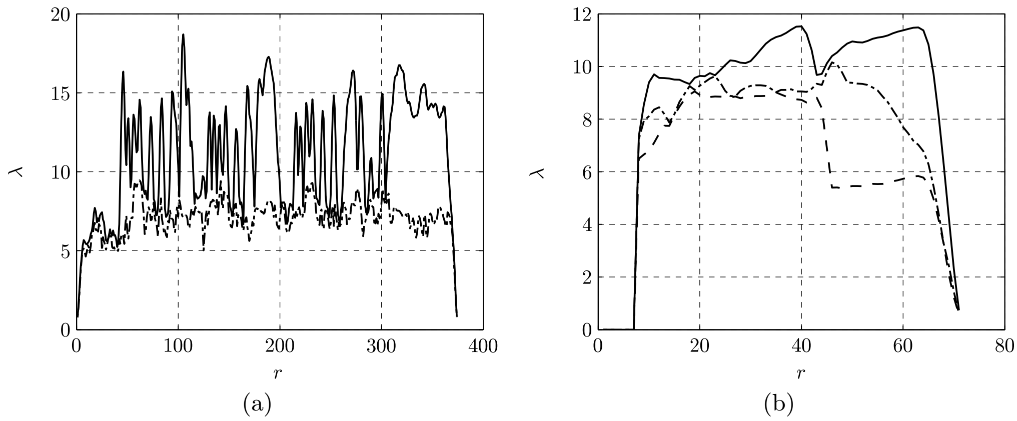


Figure 5.5: Examining the possibility of distinguishing between sources using λ as a distinguishing feature. Both figures display λ for each source before the disentangling process.(a) a mixture of two interweaving melodies; (b) a mixture of 3 synchronous notes of the same length. Lines of different type correspond to different sources. (Note the scale change at the axes.)

the only requirement for using λ as a feature for F0 tracking is the predominance of one of the sources (or a clear energy difference between the sources if there are more than 2) it can be applied to all cases, not just Case 1 (with Cases 3 and 4 needing supplementary timing information).

However, when the sources are mixed with equal or similar energies, or with energies that change over time, salience cannot be used reliably as a feature for F0 grouping. Fig. 5.5b shows such an example. In this situation, a heuristic grouping rule can be used instead. For example, the use of a rule that is based on research in music perception and composition and states that pitch tracks coming from different instruments should not cross each other [82] can be a quite reasonable choice for a variety of musical mixtures [107]. This rule, nonetheless, assumes correct multiF0 estimation, or knowledge of the ground-truth F0 contours. When automatic multiF0 estimation is employed, it is often not error-free – there will be time instances where it will fail for one or more sources. For this reason, a different approach is adopted and described below. It is a two-part process and deals with both the issues of track disentangling and correction of multiF0 estimates.

5.4.1 Identification of mis-labelled silences

There are times where silence segments exist in the beginning and/or the end of a note event. While for those particular frames the polyphony value that drives the multiF0 estimator should be $O = 0$, there are cases where this may not be true (for example, because of wrong timing information). In those cases the multiF0 estimator will be misled and will produce obviously erroneous estimates $\{f_{0r}^{(p)}\}_{p=1}^{P}$ for all the sources that were supposedly present in the mixture.

For this reason, a correction stage seeks to identify those time-frames and set all their associated F0 values to zero. It was found that the salience measure λ can be used effectively for this purpose, since it is directly related to the magnitude of the STFT spectrum: the frames r where $\lambda_r^{(p)}$ is below a certain threshold ρ and $O_r \neq 0$ are re-labelled as silenced and the F0s associated with those frames are set to zero. More formally:

$$\{f_{0r}^{(p)}\}_{p=1}^{P} \leftarrow 0, \quad \forall p, \forall r \in [1, R] : O_r \neq 0 \wedge \lambda_r^{(p)} < \rho. \tag{5.13}$$

The value of $\rho = 1$ was found to be satisfactory after considerable experimentation. As with λ (which is derived from the weighted sum of spectral magnitude values), ρ signifies a spectral magnitude threshold.

5.4.2 F0 swapping/correcting process

This process relies on the assumption that the multiF0 estimation stage has delivered correct estimates for each of the sources in at least 51% of all the frames. Using this assumption, the P most probable F0s in the whole of $\mathbf{F}$ (in other words, in the whole mix) can be found.

First of all, its columns are concatenated in a single vector $\mathbf{f}^{\mathsf{conc}}$ of size $RP \times 1$. The concatenation operation on $\mathbf{F}$ is equivalent to:

$$\mathbf{f}^{\mathsf{conc}} \;=\; [\mathbf{f}^{(1)^{\mathsf{T}}}\; \mathbf{f}^{(2)^{\mathsf{T}}}\; \ldots\; \mathbf{f}^{(P)^{\mathsf{T}}}]^{\mathsf{T}}. \tag{5.14}$$

```
begin
    Input $\mathbf{f}^{\mathsf{conc}}$
    $\mathbf{h} \leftarrow \mathsf{hist}(\mathbf{f}^{\mathsf{conc}})$
    for $p \leftarrow 1$ to $P$ do
        $f^{(p)}_{\mathsf{pitch}} \leftarrow \mathsf{arg\,max}(\mathbf{h})$
        $h_{f^{(p)}_{\mathsf{pitch}}} \leftarrow 0$
        $f^{(p)}_{0\,\mathrm{best}} \leftarrow$ mean of all values $\{f^{\mathrm{conc}}_m\}_{m\in[1,RP]}$ that satisfy
        $|f^{\mathrm{conc}}_m - f^{(p)}_{\mathsf{pitch}}| < \mathfrak{c} f^{(p)}_{\mathsf{pitch}}$.
end
```

Figure 5.6: Algorithm for finding the P most probable F0s in the mix, $\{f^{(p)}_{0\,\mathrm{best}}\}^{P}_{p=1}$.

Next, a 69-bin histogram of $\mathbf{f}^{\mathsf{conc}}$ is calculated. Fig. 5.6 shows the algorithm for the calculation of the P most probable F0s. The histogram is denoted by a 69-element vector $\mathbf{h}$:

$$\mathbf{h} \;=\; [h_1\, h_2\, \ldots\, h_{69}]^{\mathsf{T}}, \tag{5.15}$$

and its bins are centred at the frequencies corresponding to the pitch of notes E1 to C7. The P largest values of $\mathbf{h}$ are then identified – these are the P most probable pitches, $\{f^{(p)}_{\mathsf{pitch}}\}^{P}_{p=1}$. From this, a 'best' F0 value $f^{(p)}_{0\,\mathrm{best}}$ is derived for each p by searching $\mathbf{f}^{\mathsf{conc}}$ for frequencies that are less than half a semitone apart from $\{f^{(p)}_{\mathsf{pitch}}\}^{P}_{p=1}$ and taking their mean. This is why the value of $\mathfrak{c} = 0.03 \simeq \frac{2^{1/12}-1}{2}$, corresponding to the half semitone frequency ratio (see §2.2.1), is used in the algorithm.

The $f^{(p)}_{0\,\mathrm{best}}$ values are used, next, for disentangling and correcting the F0s. Fig. 5.7 shows the algorithm of this process. For each r, the value of $f^{(p')}_{0r}$ that is closest to $f^{(p)}_{0\,\mathrm{best}}$ is placed at the p-th track. However, it is not allowed for $f^{(p')}_{0r}$ to be the closest frequency for another $f^{(q')}_{0\,\mathrm{best}}$; if this happens, $f^{(p')}_{0r}$ is set to zero as an erroneous multiF0 estimate to be inferred at the F0 correction stage that follows.

A simple approach is taken for the correction of the F0 estimates that have been labelled as erroneous. Using the values that have been labelled as correct, *i.e.*,

```
begin
    Input $\{f_{0r}^{(p)}\}_{p=1}^{P}$
    $\{\tilde{f}_{0r}^{(p)}\}_{p=1}^{P} \leftarrow \{f_{0r}^{(p)}\}_{p=1}^{P}$
    for $r \leftarrow 1$ to $R$ do
        for $p \leftarrow 1$ to $O$ do
            $q \leftarrow \mathsf{arg\,min}_{p' \in [1,P]}(|f_{0\,\text{best}}^{(p)} - f_{0r}^{(p')}|)$
            if $q \neq p$ then
                $\tilde{f}_{0r}^{(p)} \leftarrow f_{0r}^{(q)}$
            if $\exists q' \in [1,P] : |f_{0\,\text{best}}^{(q')} - \tilde{f}_{0r}^{(p)}| < |f_{0\,\text{best}}^{(p)} - \tilde{f}_{0r}^{(p)}|$ then
                $\tilde{f}_{0r}^{(p)} \leftarrow 0$
        $\{f_{0r}^{(p)}\}_{p=1}^{P} \leftarrow \{\tilde{f}_{0r}^{(p)}\}_{p=1}^{P}$
end
```

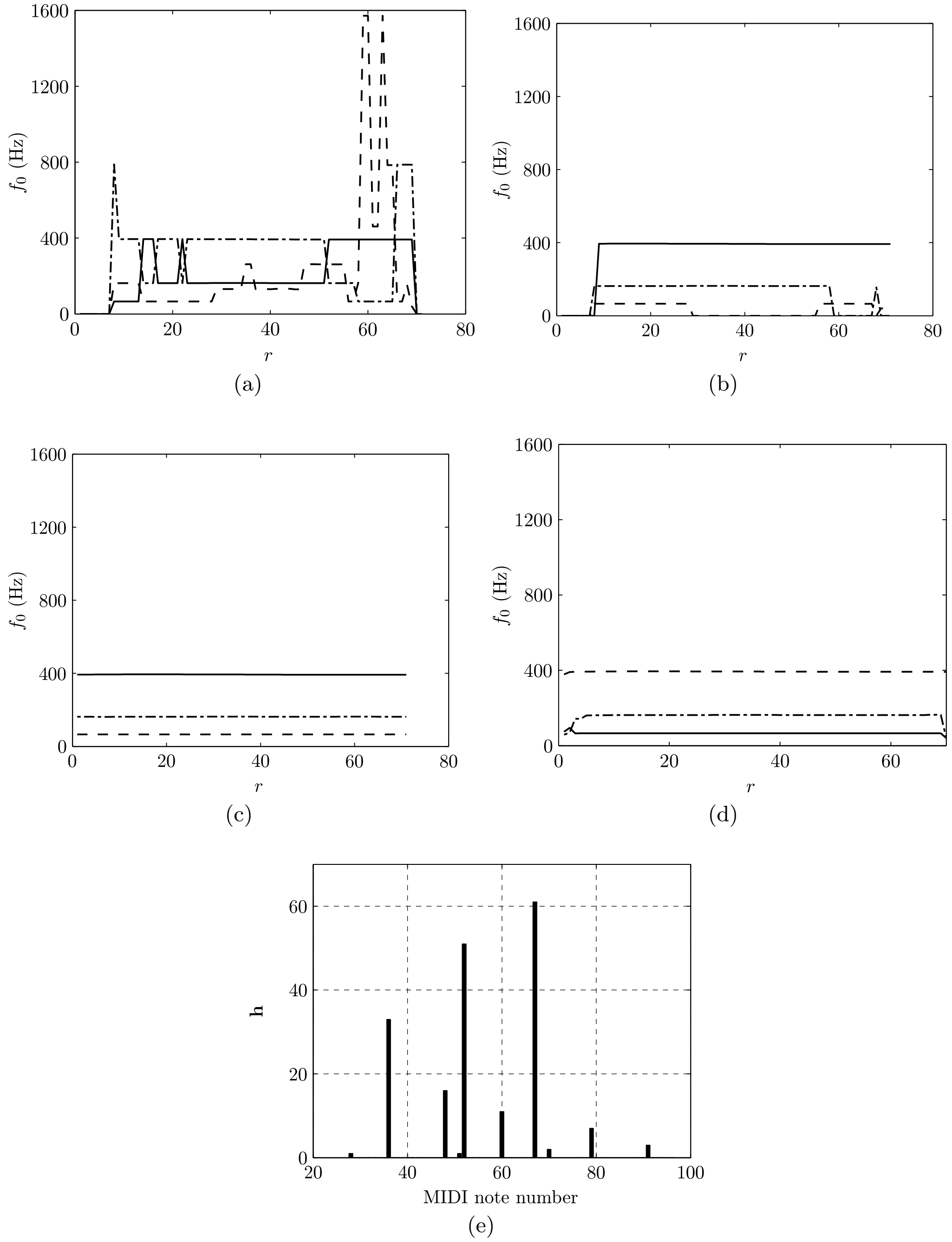
Figure 5.7: Algorithm for swapping the F0 estimates between the F0 tracks, in order to disentangle them.

the ones satisfying the inequality:

$$|f_{0\,\text{best}}^{(p)} - f_{0r}^{(p)}| \; < \; \mathfrak{c}\, f_{0\,\text{best}}^{(p)}, \quad r \in [1,R], \tag{5.16}$$

nearest neighbour interpolation across r is performed for calculating the rest of the values. Also, note that FM modulation of more than 3% is not allowed here, since the maximum allowed deviation is $\mathfrak{c} = 0.03$, corresponding to a half semitone frequency difference. Fig. 5.8 shows the frequencies before and after the swapping and correcting processes, compared with the ground-truth values, along with a graph of the histogram for the particular mix (in this case, mix 2 from Table 5.2).

In order to compare the error rates before and after the F0 track disentangling stage, the F0 estimates were obtained for 11 different mixtures. Each of the mixtures consists of three synchronously played notes having equal RMS energy. The audio samples for this case (as well as in every other case where non-synthetic audio has been used in this thesis) were acquired from the *University of Iowa Musical Instrument Samples* database [159]. The mixtures were created so as to enable a variety of mixing situations to be tested:



(a)

(b)

(c)

(d)

(e)

Figure 5.8: The disentangling process for mix 2 of Table 5.2. (a): before disentangling; (b): after track swapping; (c): after correction; (d): ground-truth F0s; (e): the histogram associated with this mix. The three highest values correspond to the most probable F0 values in the mix.

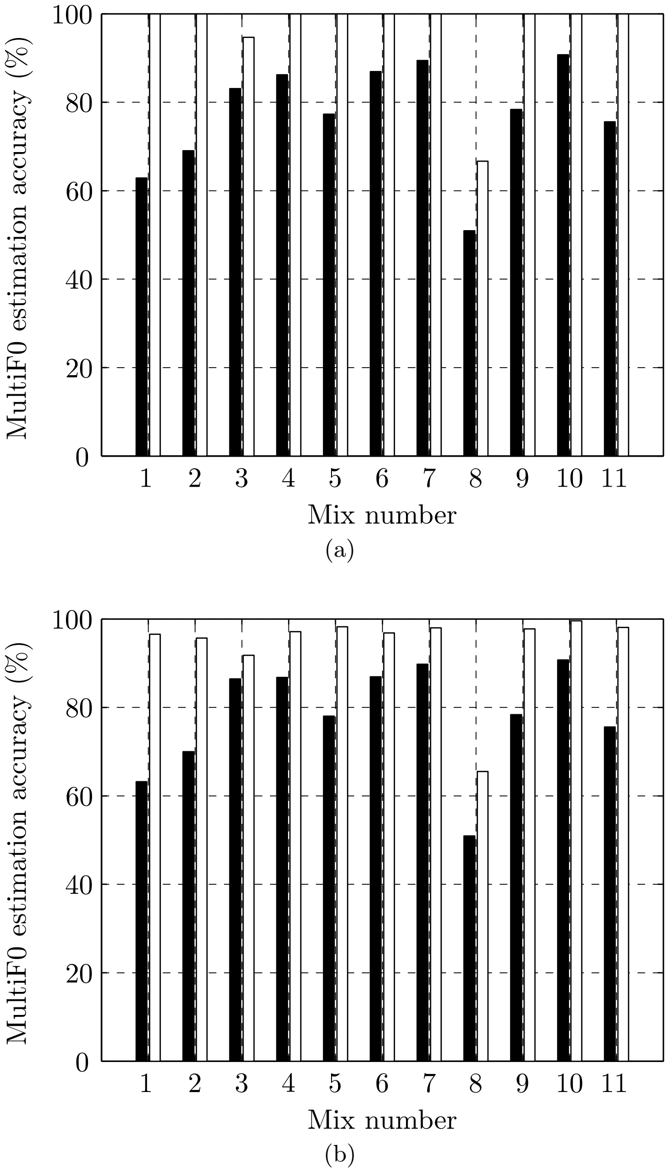
Mix number	f s_1	s_2	s_3
1	Piano A1	Cello C2	Violin E4
2	Piano C2	Cello E3	Sax G4
3	Flute D6	Bassoon A4	Cello F4
4	Flute G6	Cello E4	Sax B5
5	Piano D♭3	Bassoon F3	Sax A♭3
6	Flute G♭4	Cello B♭3	Violin D♭4
7	Piano C6	Flute B♭6	Bassoon A♭4
8	Piano G♭5	Flute E5	Cello E♭4
9	Bassoon B2	Cello B♭3	Violin A♭4
10	Flute F6	Bassoon G4	Sax E♭6
11	Cello B3	Violin B♭4	Sax A3

Table 5.2: The musical notes corresponding to the original sources contained in each of the 11 mixtures used here. In particular, the instrument and pitch associated with the notes are shown.

- harmonically-related notes (resulting in low mixture disjointness) versus non-harmonically-related ones (resulting in high mixture disjointness);
- variety of pitch combinations ranging from very low to very high frequencies;
- variety in types of harmonic instruments (brass and woodwind instruments, bowed string instruments and percussive instruments);
- inclusion of a source that exhibits inharmonicity (the piano).

Table 5.2 lists the instrument and pitch associated with these notes.

A comparison of the F0 estimation accuracy before and after F0 track disentangling is presented in Fig. 5.9. Two sets of references were used, the MIDI-pitch frequencies and the single-F0 estimates of the original sources using the same algorithm. An estimated F0 was labelled as erroneous if it deviated more than 3% from the reference. It can be seen that in both cases, and for all mixtures, the use of the F0 track disentangling stage led to an improvement of the F0 estimation accuracy. The fact that the improvement appears somewhat lower in Fig. 5.9b



(a)

(b)

Figure 5.9: Average accuracy of the multiF0 estimation for 11 3-source mixes before (black bars) and after the F0 track disentangling process (white bars). The accuracy is measured in comparison to (a) the MIDI pitch; and (b) the single-F0 estimates using the isolated sources.

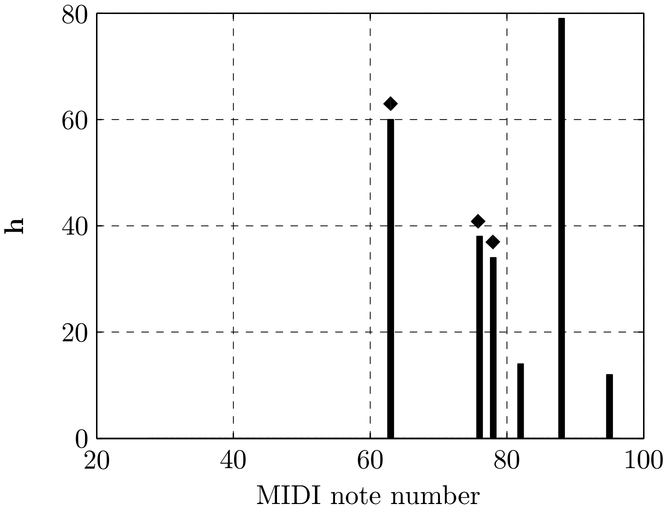


Figure 5.10: The histogram calculated in the F0 track disentangling process for mixture 8. The ground-truth MIDI pitch values are indicated with a diamond. In particular, the one corresponding to the undetected piano source (see text) corresponds to the ground-truth pitch with the lowest histogram count (at MIDI note number 78).

when compared to Fig. 5.9a is because the error correction method does not allow F0 deviations of more than 3%. This results in a relatively unchanging F0 contour that naturally matches the MIDI reference better.

Furthermore, the low performance observed for mixture 8 is a result of a complete failure of the multiF0 estimator to detect one of the sources (the piano). This is a case where the F0 track disentangling stage cannot improve the F0 estimates, since there are not enough correct estimates to work with for the duration of the note. The situation can be illustrated by the histogram calculated during the disentangling stage (Fig. 5.10). Its maximum value does not correspond to any of the ground-truth pitches. Since the algorithm selects only the J highest histogram values for further processing, the one corresponding to the piano is not selected.

5.5 Mixture pre-processing

Once the F0s have been estimated, the mixture is fed to the mixture pre-processing stage. Here, the harmonic (or near-harmonic) components associated with each source are modelled prior to separation of each source. A sinusoidal model analysis is carried out that will be used for the identification of source components corresponding to the estimated F0 tracks. The analysis is the same as in [60] and it is summarised below.

All the audio involved (sources which are going to be mixed, or the already available mono mixture) are sampled at $F_s = 44.1\,\text{kHz}$. The STFT is employed as the selected TF representation. The FFTs are calculated on $N = 8192$-sample Hamming-windowed frames, with 87.5% overlapping.[2] The amplitude spectrum is denoted as $A = |\mathcal{X}|$. First of all, the spectral peaks that are likely to have been produced by source partials are located. Peak selection is performed using a frequency-dependent threshold $\eta E(k)$, where

$$E(k) \;=\; (\tilde{A}(k))^c, \quad \forall k \in [0, N/2] \tag{5.17}$$

is the shape of the threshold, η a frequency-independent amplitude threshold height, and $\tilde{A}$ is the smoothed amplitude envelope produced by the convolution of A with a normalised Hamming window of length $1 + N/64$ samples. A suitable range for c is $[0.5, 1[$ and for this case the value $c = 0.8$ is used. Also, for making sure the envelope is adaptable to fixed-gain scaling, $\eta \propto (\mathsf{mean}(A))^{1-c}$.

The next step is to identify all the local maxima. A frequency bin k corresponds to a local maximum if

$$A(k) \;>\; b(|k-j|)\,A(j), \quad \forall j \in [k-d, \ldots, k+d] \wedge j \neq k, \tag{5.18}$$

where $b(|k-j|)$ takes values in the range of $]0,1]$ and d is the length of vector $\mathbf{b}$. This vector helps construct an adaptive threshold on j bins at either side of the k-th bin with the purpose of deciding whether it corresponds to a local maximum or not. For the specific case of $N = 8192$, $\mathbf{b}$ has been empirically chosen to be equal to $[1\,1\,0.5]^{\mathsf{T}}$. Finally, the DFT[1] method is used [44] for refining the frequency

[2] A justification for the use of a 8192-point STFT as a suitable mid-level representation for work was given in §4.12.1.

and amplitude estimates of the peak maxima, and a zero-padded version of the window's spectrum was used for calculating the amplitude refinements.

5.6 Source parameter estimation

This stage locates the harmonic partials associated with the estimated F0s and calculates their amplitudes. The search for the partials is made amongst the spectral peaks selected in the mixture pre-processing stage (§5.5). It is based on the method described by Every [58] and is basically a way of matching the observed spectral peaks with the predicted positions of the harmonics. One important distinction of this stage from other source identification techniques that use F0-inferred information is that here strict harmonicity is not implied.

We define as $m_r^{(p)}$ the index of the m-th harmonic of source p which will be ideally matched with its predicted frequency value during the r-th time frame. These will belong to the set of all the matched harmonics $M_r^{(p)} \subset \mathbb{N}_0$, the cardinality of which is set to be:

$$|M_r^{(p)}| \leq \mathsf{min}(\lfloor 0.5F_s/\underset{p}{\mathsf{min}}\,\mathsf{mean}\,\mathbf{f}^{(p)}\rfloor, 50), \quad \forall p, \tag{5.19}$$

i.e., $|M_r^{(p)}|$ is determined to be the maximum possible number of harmonics up to $F_s/2$ (the Nyquist frequency) corresponding to the minimum mean F0 in the mixture or 50, whichever is the smaller. ($\mathbf{f}^{(p)}$ is the p-th F0 track from Eq. 5.12.)

The frame index can be now dropped, since the following are frame-by-frame operations. The method is an iterative process. During each iteration, the note corresponding to $\mathsf{min}_p\,\mathsf{mean}\,\hat{\mathbf{f}}^{(p)}_{m^{(p)}+1}$ is selected, where $\hat{f}^{(p)}_{m^{(p)}+1}$ is the predicted centre frequency of the $(m+1)$-th harmonic of source p. The algorithm is initialised with $m^{(p)} = 0$, $\forall p$, with $m^{(p)}$ being incremented with every iteration of note p.[3] The predictions of frequencies are made either using the harmonic model

$$\hat{f}^{(p)}_m \;=\; m\, f_0^{(p)} \tag{5.20}$$

[3] From now on, and for the sake of readability, we set $m^{(p)} \equiv m$.

or a model of inharmonicity, such as the one used for piano notes (Eq. 2.2).

According to the method, we first find the largest peak v whose centre frequency f_v is within a range of $\delta f_0^{(p)}$ from $\hat{f}_m^{(p)}$ (satisfying the inequality $|\hat{f}_m^{(p)} - f_v| < \delta f_0^{(p)}$), where $\delta = 0.03$. Next, we examine if v is within the range of another predicted frequency, $\hat{f}_n^{(q)}$. If that is true, we search for a second largest peak u with frequency f_u that is within the matching range of $\hat{f}_m^{(p)}$. If u can be found, we have two options:

- If $\hat{f}_m^{(p)} < \hat{f}_n^{(q)}$, $\hat{f}_m^{(p)}$ is matched to the peak with the lower frequency and $\hat{f}_m^{(p)}$ is matched to the remaining peak.
- If $\hat{f}_m^{(p)} > \hat{f}_n^{(q)}$, $\hat{f}_m^{(p)}$ is matched to the peak with the higher frequency and $\hat{f}_m^{(p)}$ is matched to the remaining peak.

If a peak u cannot be found, then $\hat{f}_m^{(p)}$ is matched to f_v only if the following inequalities hold:

$$|\hat{f}_m^{(p)} - f_v| < 0.5\,|\hat{f}_n^{(q)} - f_v| \quad \text{and} \quad |\hat{f}_m^{(p)} - \hat{f}_n^{(q)}| > \frac{\mathfrak{w} F_s}{N}. \tag{5.21}$$

For the case of the Hamming window, $\mathfrak{w}$ is set to 2 bins. If the harmonic in question is matched to one of the peaks, say u, then $f_m^{(p)} \leftarrow f_u$ and $A_m^{(p)} \leftarrow A_u$ (where $A_m^{(p)}$ and A_u are the respective amplitudes of peaks $m^{(p)}$ and u). Furthermore, whenever a matching is successful, $f_m^{(p)}$ is used to refine the F0 estimate associated with that peak by minimising the least-squares error fit to the harmonic frequencies.

In the case when the harmonic in question cannot be matched to any peak, its predicted value is assigned to it: $f_m^{(p)} \leftarrow \hat{f}_m^{(p)}$. The unmatched peaks normally occur because of peak overlapping. For these peaks, the amplitude estimation is carried out via linear interpolation using the parameters of the nearest-neighbouring peaks in the time and frequency domain. Although this way of parameter correction is quite adequate for the type of system that is the goal of this thesis, it was found that the process can be enhanced by adding the ability to perform linear *extrapolation* of the amplitudes. Indeed, there are cases where the fundamental (and maybe subsequent lower harmonics) of a note can be overlapped with other peaks throughout the total duration of the note. If extrapolation using the higher harmonics is not included in the algorithm, the unmatched harmonics will acquire zero amplitudes. Additionally, since this enhancement stops some of the F0 peaks from being rejected, it can lead to a significant improvement in separation performance. Fig. 5.11 illustrates the effect of linear amplitude extrapolation for the sax note G4 of mix 2.

5.7 Source extraction

The previous stage provided the central frequencies $\{f_m^{(p)}\}$ and amplitudes $\{A_m^{(p)}\}$ of the identified source partials. These partials will be now extracted from the mixture using the process described below. It is an adaptive filtering method according to which a spectral filter provides the *mask* which will be multiplied with the DFT spectrum of each frame, for isolating the desired partials. In particular, it is the "Filter a" energy-based approach presented in [60], and was chosen among the other extraction approaches presented there because it showed the most satisfying separation performance in terms of SRR (as was shown in [60]).

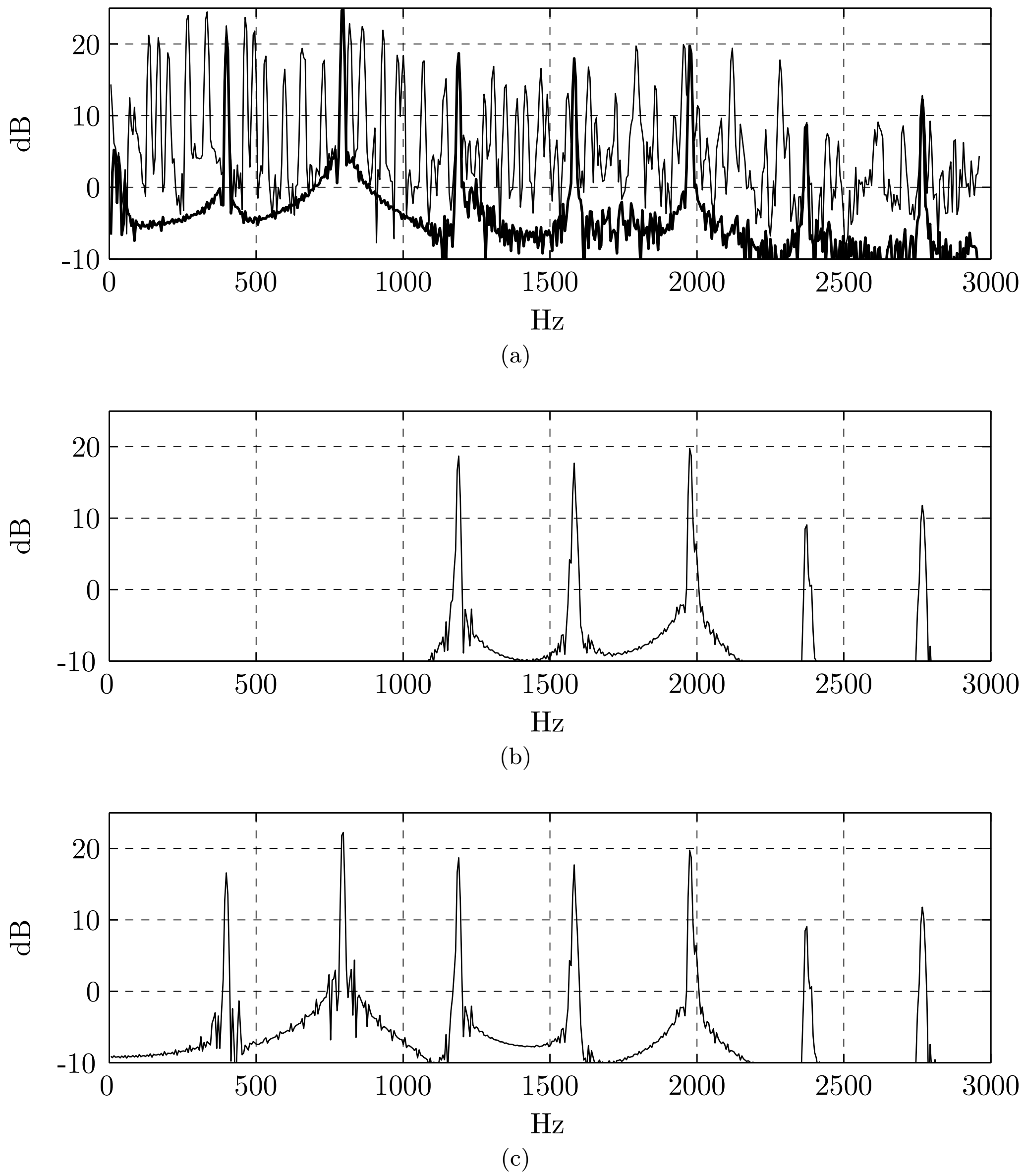
Let $H^{(p)}(k)$ be the frequency response of the filter for source p. First, we initialise it to 0 for every $k \in [0, N/2]$. The amplitude of this filter has the form:

$$\tilde{H}^{(p)}(k) \;=\; A_m^{(p)} \exp\left(-\frac{|f_k - f_m^{(p)}|}{\sigma} \right), \quad m = 1, 2, \ldots, M^{(p)}. \tag{5.22}$$

This is followed by the normalisation:

$$H^{(p)}(k) \;=\; \frac{\tilde{H}^{(p)}(k)}{\sum_{p'} \tilde{H}^{(p')}(k)}, \quad \begin{array}{c} p = 1, 2, \ldots, P \\ k_{m\mathrm{L}}^{(p)} \leq k \leq k_{m\mathrm{R}}^{(p)} \end{array}, \tag{5.23}$$

where $k_{m\mathrm{L}}^{(p)}$ and $k_{m\mathrm{R}}^{(p)}$ are the first minima in A below $[k_m^{(p)} - 3]$ and above $[k_m^{(p)} + 3]$, respectively. A suitable value for σ is $0.25 F_s/N$. According to Eqs. 5.22 and 5.23,



(a)

(b)

(c)

Figure 5.11: The effect of linear extrapolation on the spectrum of the sax G4 of mix 2 at frame $r = 30$. (a): The original mix (thin line) with the original source (thick line); (b): the extracted sax before amplitude extrapolation; (c): the extracted sax after amplitude extrapolation.

if a harmonic overlaps with, *e.g.*, two harmonics in any bin k, the spectral energy is split into three parts depending on the values of $f_m^{(p)}$ and $A_m^{(p)}$. Thus, the filtering masks are real-valued with gains $0 \leqslant H^{(p)}(k) \leqslant 1$, as opposed to binary time-frequency masks which assign the energy of each time-frequency point to only one source. With a help from the above we can define:

$$H(k) \;:=\; \begin{cases} \sum_{p=1}^{P} H^{(p)}(k) \;=\; 1, & \text{for } k \in [k_{m\mathrm{L}}^{(p)}, k_{m\mathrm{R}}^{(p)}] \\ 0, & \text{otherwise.} \end{cases} \tag{5.24}$$

The part of the original spectrum which is filtered out corresponds to the frequency bin numbers of the set $\{k \in [0, N-1] : H(k) = 1\}$.[4] Any amount of energy located in all the remaining frequency bins will not be filtered, and consequently end up in the residual signal. Provided, however, that the peak parameters have been estimated correctly, and since the filters operate within the specified regions $[k_{m\mathrm{L}}^{(p)}, k_{m\mathrm{R}}^{(p)}]$, no main-lobe harmonic content will be left in the residual after the filtering operation (see Fig. 4.5 for an illustration of this). Also, since, as Eq. 5.24 shows, $H(k)$ is not unity for all k, the unitary sum constraint [164] is not satisfied.

The resynthesis of the separated source p is achieved by first obtaining its filtered spectrum:

$$\hat{\mathcal{S}}^{(p)}(k) \;=\; H^{(p)}(k)\,\mathcal{X}(k), \quad k = 0, 1, \ldots, N-1. \tag{5.25}$$

The time-domain frame-level source signals $\hat{s}^{(p)}$ are produced by applying the Inverse FFT (IFFT) on $\hat{\mathcal{S}}^{(p)}$. The algorithm preserves the mixture phases at the frame level. Thus, the frame-level residual can be simply defined as:

$$x_{\mathsf{res}}(n) \;:=\; x(n) \;-\; \sum_{p=1}^{P} \hat{s}^{(p)}(n). \tag{5.26}$$

Using an overlap-add method with triangular windows (see §3.2.1) the complete signals x_{res} and $\hat{s}^{(p)}$ are finally acquired.

[4] The filter values at the interval $[N/2+1, N-1]$ are symmetric to the values at $[0, N/2]$ with regards to the Nyquist frequency.

Further discussion regarding the use of x_{res} will be carried out in the following sections. Before we do that, though, a performance comparison will be presented first between the proposed system so far and its MIDI-informed version.

5.8 Performance comparison to the previous system

A comparison is now carried out between the system by Every [60, 58] and the modified implementation of that system as it has been presented in this chapter. The comparison will be in terms of both separation and multiF0 estimation performance. First, however, the relative merits of using the SRR and the SDR is explored, in order to make an educated choice of a separation performance measure.

5.8.1 Comparing the signal-to-residual ratio and the signal-to-distortion ratio

Theoretical analysis

We can start the comparison analysis between the SRR[5] and SDR by rewriting Eq. 4.40: because of Eq. 4.34, it can be written as:

$$\mathsf{SDR}_j \;=\; 10\log_{10}\frac{\|s_{\mathsf{target}}\|^2}{\|\hat{s}_j - s_{\mathsf{target}}\|^2}\ \mathrm{dB}. \tag{5.27}$$

As can be easily seen by comparing the above equation with Eq. 4.30, the difference between the two measures lies in the definition of the reference signal: the SRR uses s_j for this purpose, while the SDR uses s_{target}. Additionally, from Eqs. 4.34 and 4.35, it can be seen that (1) by employing s_{target}, the SDR allows for time-invariant gain distortion, *i.e.*, if $\hat{s}_j = as_j$ for $a \in \mathbb{R}$ then $s_{\mathsf{target}} = s_j$; and (2) that $e = \hat{s}_j - s_{\mathsf{target}}$ is orthogonal to s_j. This relationship between the involved quantities can usefully be visualised in vector form, as shown in Fig. 5.12. The distortion is expressed as the magnitude of vectors e (for the SDR) and e' (for

[5] In this section, the use of the acronym 'SRR' refers to $\mathsf{SRR}_{s_j}(\hat{s}_j)$ (see p. 99).

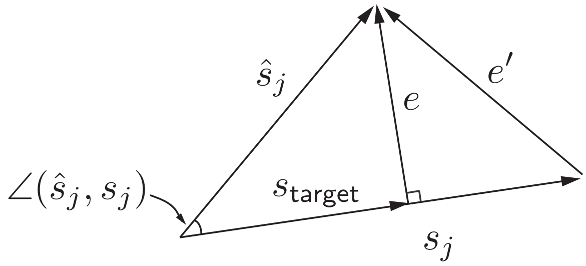


Figure 5.12: Figure illustrating the difference between the calculation processes of the SRR and the SDR: the SDR involves the ratio s_{target}/e, while the SRR involves the ratio s_j/e'.

the SRR). It can be observed that e' depends on the angle between $\hat{s}_j$ and s_j and the ratio of their magnitudes, while e depends only on the angle. In order to gain a better understanding of the way these characteristics influence the behaviour of the two measures, it is chosen here to compare them as functions of the angle $\angle(\hat{s}_j, s_j)$ (measured in rad) and the ratio

$$\Delta E(\hat{s}_j, s_j) \;=\; 10 \log_{10} \frac{\|\hat{s}_j\|^2}{\|s_j\|^2} \tag{5.28}$$

(measured in dB). The *cosine similarity* function

$$\mathsf{cossim}(\hat{s}_j, s_j) \;=\; \frac{\langle \hat{s}_j, s_j \rangle}{\|\hat{s}_j\| \|s_j\|} \tag{5.29}$$

will also be employed. This is just the cosine of $\angle(\hat{s}_j, s_j)$, but since this measure is a commonly used tool for measuring the similarity between vectors, it will be used to provide an additional viewpoint.

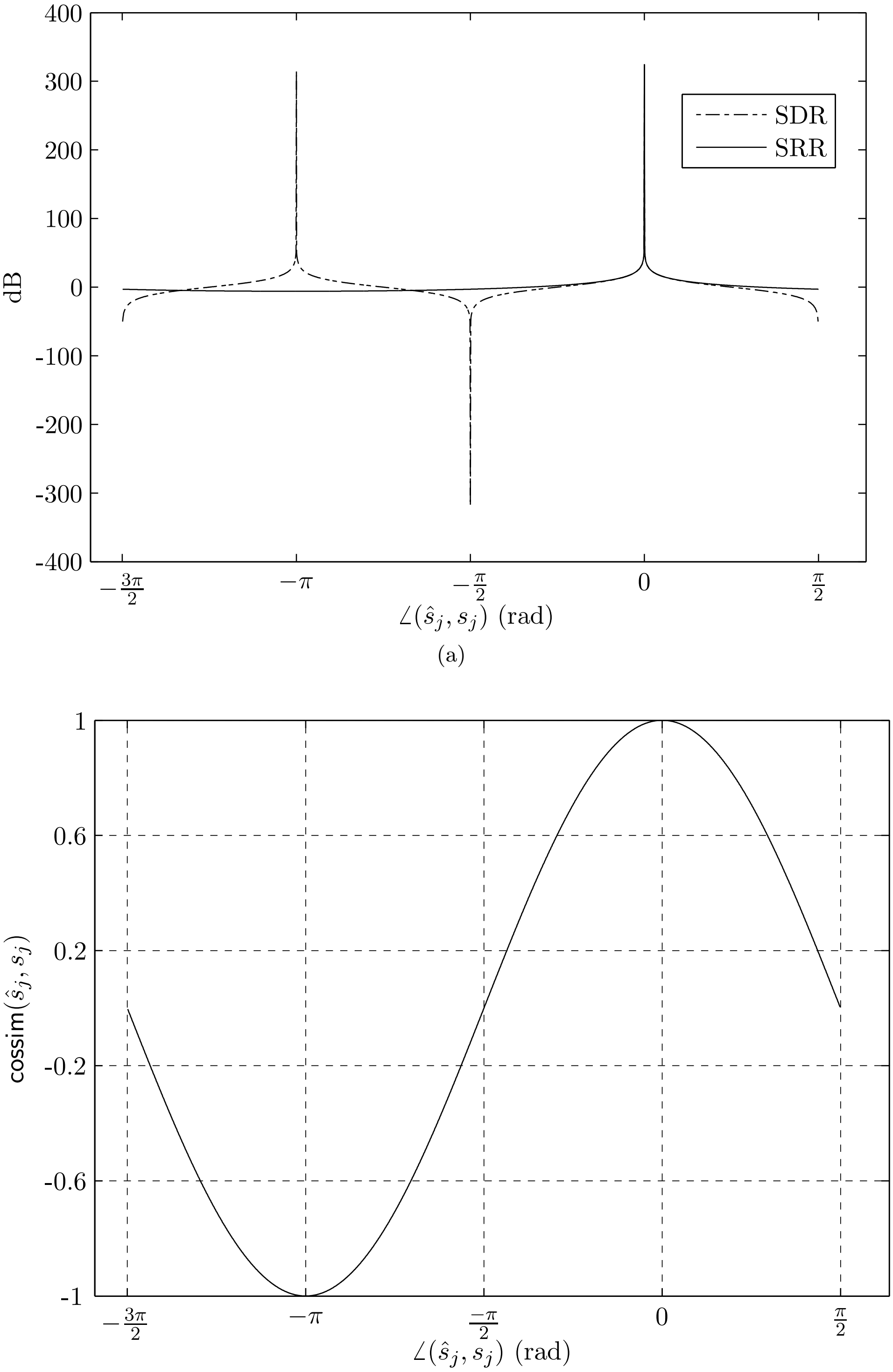
Since this analysis is carried out in a theoretical framework, we begin by a 2-element vector of the type $s_j = [\alpha\, 0]^{\mathsf{T}}$ as the reference, where α is a positive scalar (its value does not matter – in this analysis $\alpha = 10$ is chosen); from this, a variety of signals $\hat{s}_j$ can be then easily obtained according to different angles and energy differences, compared to s_j. Also, if it is assumed that the extracted sources have been ideally sorted based on a comparison to the original ones and the extraction/resynthesis process is linear, situations where $\Delta E > 0$ (*i.e.*, when the extracted source has higher energy than the original source) do not need to be considered here.

Fig. 5.13a shows the variation of the two measures with $\angle(\hat{s}_j, s_j)$ when $\Delta E = 0$, and this is compared with their cosine similarity (Fig. 5.13b). First of all, it can be seen that the SRR follows the SDR when $\angle(\hat{s}_j, s_j)$ is roughly in the range $[-0.4, 0.4]$; For the rest of the angles, the SRR does not deviate much from 0. On the other hand, the SDR is more consistently affected by $\angle(\hat{s}_j, s_j)$, exhibiting a periodic behaviour: as expected, it reaches a maximum when $\angle(\hat{s}_j, s_j)$ is around the values 0 ($\hat{s}_j$ is identical to s_j) or $-\pi$ (there is a π rad phase difference between $\hat{s}_j$ and s_j), while reaching a minimum when $\angle(\hat{s}_j, s_j) = -\frac{\pi}{2}$ ($\hat{s}_j$ and s_j are orthogonal to each other). A particularly interesting case is the maximisation of the SDR when $\angle(\hat{s}_j, s_j)$ is around $-\pi$, effectively mirroring the behaviour of the SDR around π. This means that the SDR does not consider a $|\pi|$ phase shift as distortion, whereas the SRR does. This is an advantage for the SDR since it makes sense for a performance measure in source separation to indicate successful separation when $\hat{s}_j$ and s_j are otherwise identical, apart from having opposite phases.

Next, it is worth focusing on a smaller angular range, in particular the range where the SRR and the SDR appear to behave similarly. Fig. 5.14 displays them in the angle interval $[\frac{\pi}{10^3}, \frac{3\pi}{4} - \frac{\pi}{10^3}]$, and on a set of $\Delta E = \{-20, -10, -5, -2, -1, -0.5, 0\}$ dB. It is shown that the SRR follows the SDR only within the small angle range from 0 to 0.4 rad and when $\Delta E > -0.5$ dB. It is also seen clearly, here, that the SDR has no dependence on ΔE, while this is not true for the SRR. These observations can also be made from an additional point of view: Fig. 5.15 shows that SRR increases in an almost linear fashion for ΔE up to around -10 dB. After that point it starts increasing at a higher rate, reaching the SDR values, as ΔE is tending to 0 dB (but only for the angles no higher than 0.4 rad $\{\frac{\pi}{200}, \frac{5\pi}{200}, \frac{10\pi}{200}, \frac{15\pi}{200}\}$).

Comparison in a practical scenario

Finally, a comparison of the two measures for ‘real’ signal separation is carried out. The measures in question are used for analysing the separation performance of the one-pass system on the set of mixtures of Table 5.2.



(a)

(b)

Figure 5.13: Variation of SRR and SDR when $\angle(\hat{s}_j, s_j) \in [\frac{-3\pi}{2}, \frac{\pi}{2}]$ (shown in (a)) and $\Delta E(\hat{s}_j, s_j) = 0$ dB. This is compared with the cosine similarity between the two vectors, which is shown in (b).

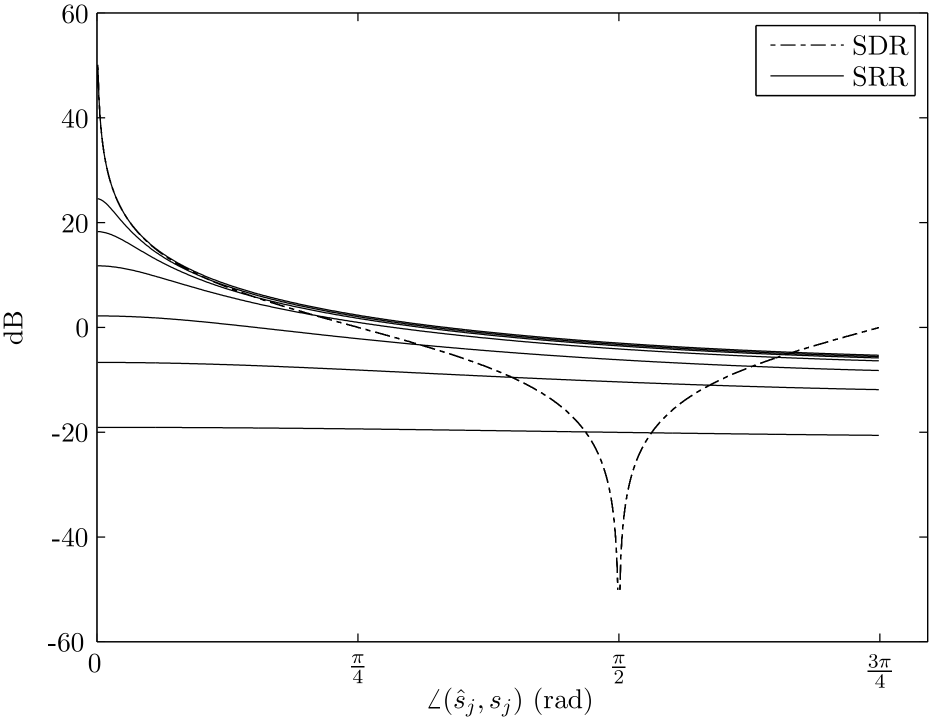


Figure 5.14: Variation of SRR and SDR when $\Delta E(\hat{s}_j, s_j)$ takes the set of values -20, -10, -5, -2, -1, -0.5 and 0 dB and $\angle(\hat{s}_j, s_j) \in [\frac{\pi}{10^3}, \frac{3\pi}{4} - \frac{\pi}{10^3}]$. Curves placed lower correspond to smaller ΔE.

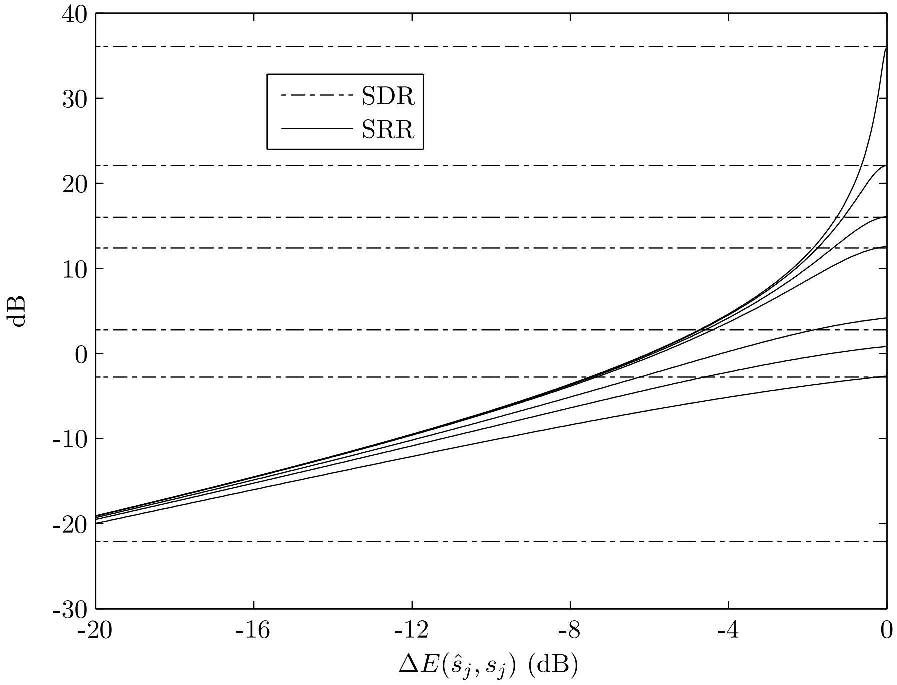


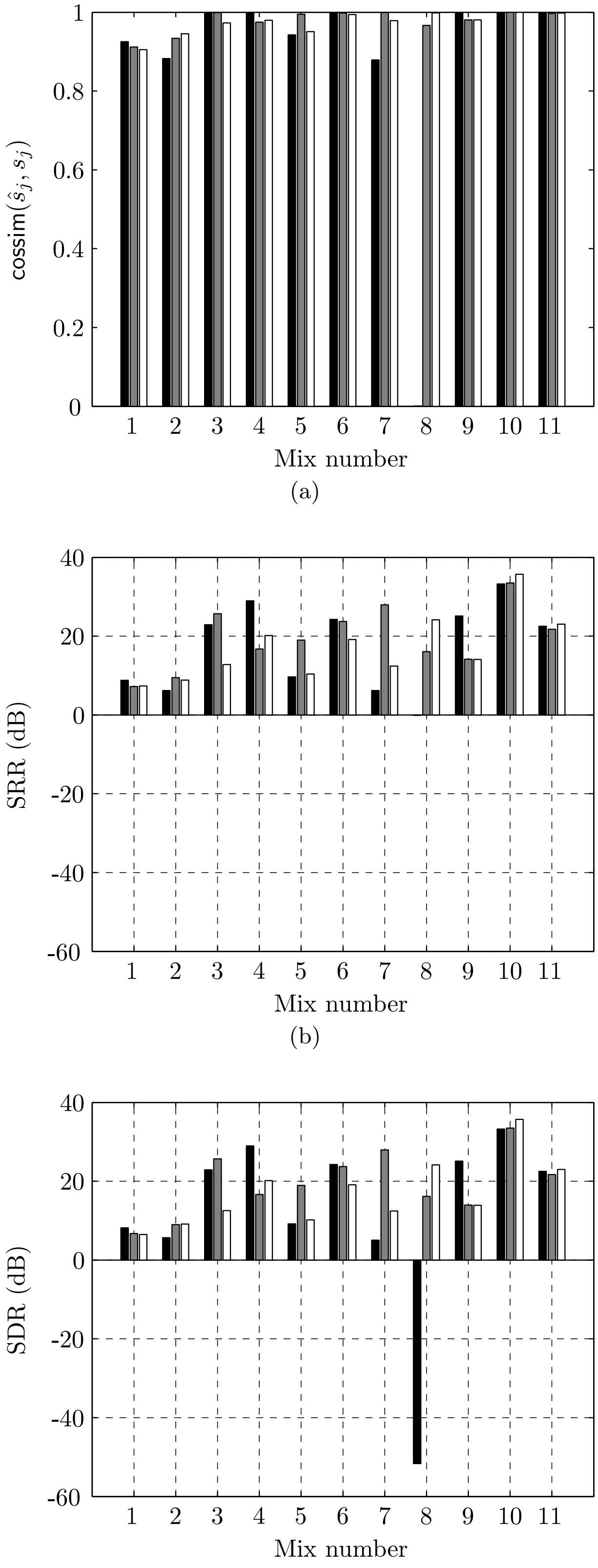
Figure 5.15: Variation of SRR and SDR with $\Delta E(\hat{s}_j, s_j)$ when $\angle(\hat{s}_j, s_j)$ takes the values $\frac{\pi}{200}$, $\frac{5\pi}{200}$, $\frac{10\pi}{200}$, $\frac{15\pi}{200}$, $\frac{40\pi}{200}$, $\frac{60\pi}{200}$ and $\frac{95\pi}{200}$. Curves placed higher correspond to smaller angles.

As Fig. 5.16 shows, the SRR and the SDR are practically identical to each other for almost all the extracted sources and mixtures apart from s_1 in mixture 8. According to the figures above, this means that for all the extractions in this group of mixtures apart from s_1 in mixture 8, $\Delta E(\hat{s}_j, s_j)$ and $\angle(\hat{s}_j, s_j)$ are roughly in the regions $[0.5, 0]$ dB and $[-0.4, 0.4]$ rad, respectively (this region of angles is also confirmed by Fig. 5.16a), where the SRR and the SDR show a practically identical behaviour. However, for the case where the system failed to extract the source, the SDR shows superiority. The measure indicates clearly that the system has failed in that particular situation. This is in contrast to the SRR which, although it has the lowest value in Fig. 5.16b (-0.1 dB), it is not very clearly differentiated from other extractions, such as the extraction of s_1 in mixtures 2 and 7; in those mixtures, while a large part of the source was identified and extracted correctly (the multiF0 estimator did not fail), the assumed source model was still not entirely appropriate for those sources (the sources were piano notes and inharmonicity was not incorporated into the model for those particular separations).

In conclusion, allowing a fixed-gain distortion is a sensible assumption for this group of mixtures and the particular separation system. It could be argued that this could be possibly extended to the a lot of the CASA-related systems because of their use of similar processing stages. Furthermore, $\angle(\hat{s}_j, s_j)$ is of a greater importance in terms of offering an insight on the amount of distortion. The SDR is totally insensitive to $\Delta E(\hat{s}_j, s_j)$ while being highly dependent on $\angle(\hat{s}_j, s_j)$. At the same time, the SRR depends on both, while being more sensitive to ΔE. These observations make the SDR a better overall choice as a measure for separation performance analysis compared to the SRR. Because of this, the SDR will be used for analysing the performance of the proposed system.

5.8.2 Performance comparison

A comparison between the two systems will be now made, in terms of their separation and multiF0 estimation performance. For validation purposes, this com-



(a)

(b)

(c)

Figure 5.16: (a) shows the cosine similarity between $\hat{s}_j$ and s_j in the 11 3-source mixtures of Table 5.2. This is compared with (b) the SRR and (c) the SDR, after having separated the source signals using the proposed system in its one-pass version. The black, grey and white colours correspond to s_1, s_2 and s_3, respectively.

parison is restricted to comparing Every's system [60, 58] with the assumption of harmonic peaks to the one-pass version of the proposed approach that uses the same assumption. (This assumption still allows for deviations from absolutely strict harmonicity in both methods, as shown in §5.6.) In order to explore a variety of situations, mixtures belonging to cases 1 (single notes) and 3 (melodies) will be used.

Case 1-type mixtures

The 11 3-source mixtures of Table 5.2 are used here. As a way of comparing 66 SDR measures in pairs within the same figure, a difference ΔSDR is introduced:

$$\Delta\mathrm{SDR} \;=\; \mathrm{SDR}^{\mathsf{prop}} \;-\; \mathrm{SDR}^{\mathsf{midi}}, \tag{5.30}$$

where $\mathrm{SDR}^{\mathsf{prop}}$ are the SDR values of the proposed system and $\mathrm{SDR}^{\mathsf{midi}}$ are the SDR values of the MIDI-informed system. Fig. 5.17a shows $\mathrm{SDR}^{\mathsf{midi}}$ for the 11 mixes, while Fig. 5.17b shows ΔSDR. (Note, also, that $\mathrm{SDR}^{\mathsf{prop}}$ has already been already been presented in Fig. 5.16c.)

It can be observed that the proposed system exhibits a more robust overall performance, compared to the MIDI-informed system, for this group of mixtures. Apart from mixture 8, where s_1 fails completely to be detected by the multiF0 estimation front-end (followed by a degradation in s_2 because a part of it was leaked to s_1), the system performs equally well, or even better than its MIDI-informed version. At first, this might seem surprising, given that the knowledge of the MIDI note pitch and timings is quite a powerful means for correctly identifying and separating source structures. Since the basic difference of the two systems is in the way the pitch and timing information is obtained, these results show that the use of initial estimates of this information has its limitations. Indeed, user-improvised MIDI information can often be misleading; most often it is assumed that this information is roughly correct, but because it is based on the capability of the user, it can deviate unpredictably from the ground-truth. Because of this, the overall success of the separation system depends largely on the

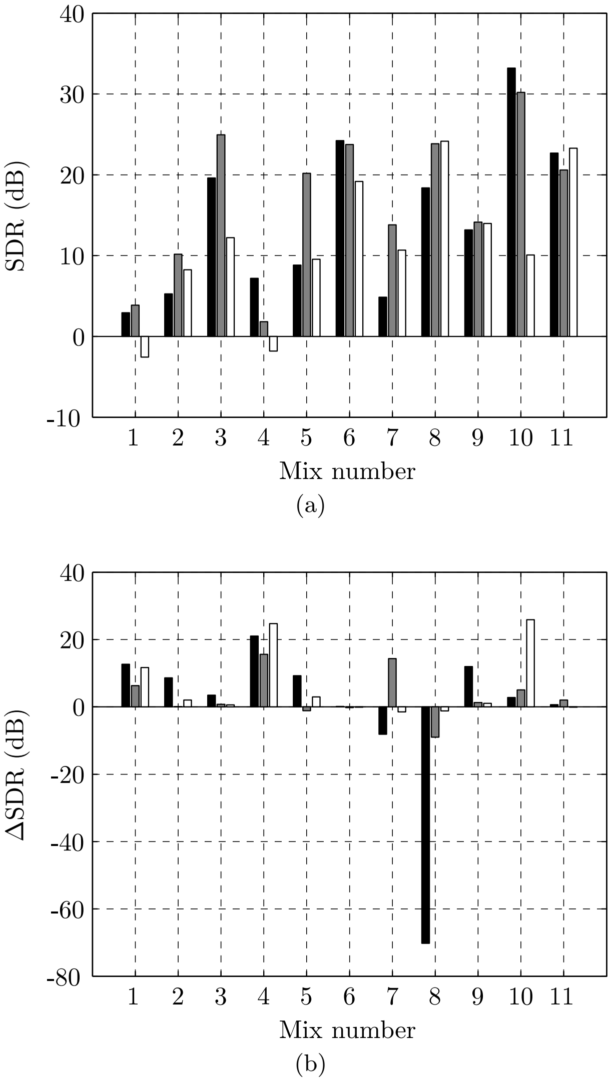


Figure 5.17: (a) SDR performance of the MIDI-informed system by Every. A comparison of this to the performance of the proposed system by using ΔSDR is shown in (b). The 3-source mixtures of Table 5.2 are used here. Positive ΔSDR values indicate superiority of the proposed system. The black, grey and white colours correspond to s_1, s_2 and s_3, respectively. (Note, also, the vertical axis scale change between the two figures.)

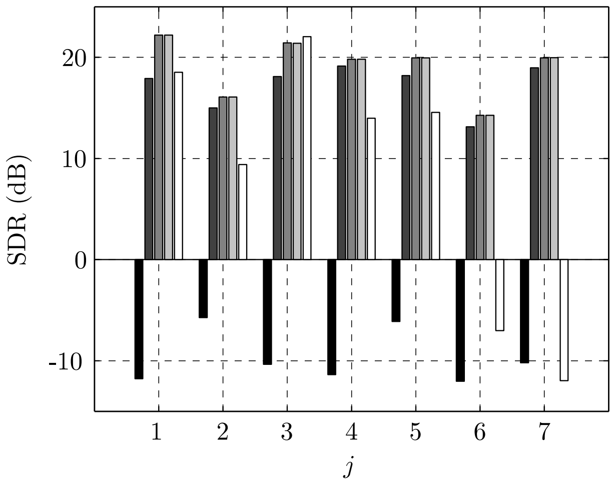


Figure 5.18: Separation performance for a mixture of $J = 7$ synchronous violin notes of the same duration (F5, A♭5, A5, B5, D♭6, E6 and G♭6). The following five cases are compared: no F0 track disentangling/no ideal grouping (first column, in black); no F0 track disentangling/ideal grouping; F0 track disentangling/ideal grouping; F0 track disentangling/no ideal grouping; MIDI-informed separation (final column, in white).

stage that provides further refinement of this information[58, Ch. 3]. In the case of Every's system, this stage does not fulfil this need to an adequate degree. The automatic, unsupervised alternative presented in this thesis, shows a relatively more consistent behaviour.

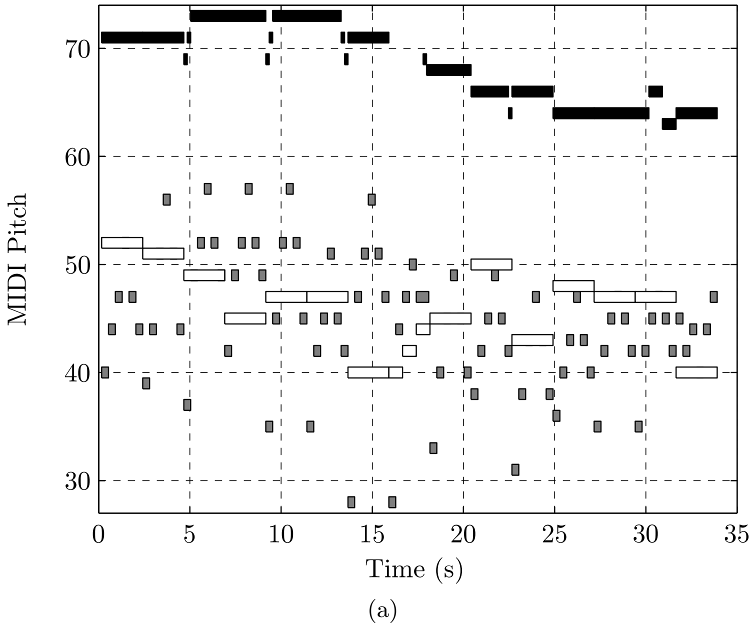
This previous comparisons showed how F0 track disentangling as a composite process has an effect on the separation performance. A way to examine the degree to which the swapping and correction of F0s contribute separately to the performance is to include a case where frame-wise ideal grouping is included after the source extraction stage: a process of sorting the extracted frames in their respected sources according to how well they match with the original source frames. A separation example is carried out, next, that employs ideal grouping.

This time a larger number of sources is used, for providing a different challenge to the system because of the increased mixture complexity; the mixture contains 7 anechoically recorded violin notes mixed with equal RMS energies. With ascending order from $j = 1$ to 7, the notes are F5, A♭5, A5, B5, D♭6, E6 and G♭6. The separation performance for this mixture in various situations is shown in Fig. 5.18 and the related audio can be listened to on the web at [150]. First of all, it can be seen that the automatic version of the separation system exhibits similar, or in most cases, better performance than the MIDI-informed version (*i.e.*, Every's system). This is shown more starkly for the last two notes where, while the MIDI-informed version fails to detect them completely, the automatic system detects and extracts them adequately. Secondly, the improvement of 'no F0 track disentangling/ideal grouping' compared to 'no F0 track disentangling/no ideal grouping', along with the performance similarity between 'F0 track disentangling/ideal grouping' and 'F0 track disentangling/no ideal grouping' shows that the F0 swapping part of the F0 disentangling process vastly contributes to a good separation performance. Thirdly, the improvement of the 'F0 track disentangling/ideal grouping' case compared to 'no F0 track disentangling/ideal grouping' shows that the F0 correction part of the disentangling stage also contributes to an increase in SDR.

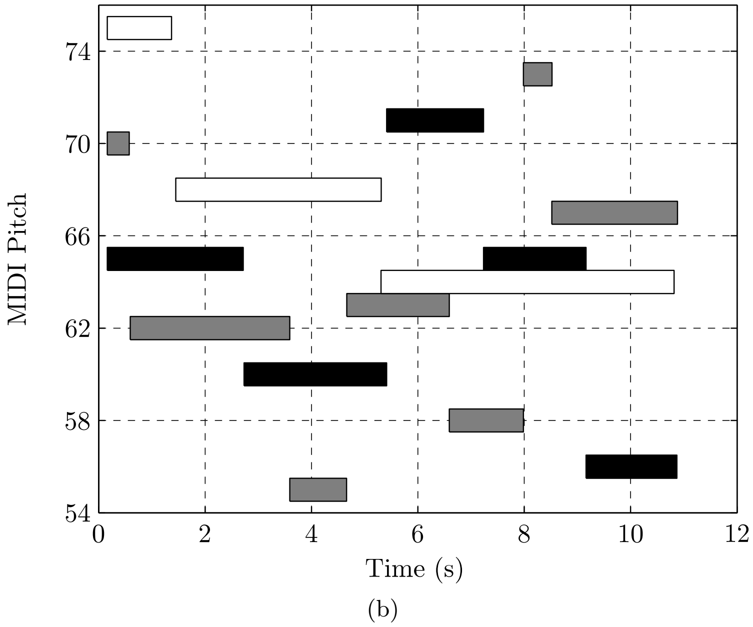
Case 2-type mixtures

There are 6 mixtures used here that fall into the category of Case 2 (as defined in §5.2): 3 versions of `melody_mix1` (Fig. 5.19a) and 3 versions of `melody_mix2` (Fig. 5.19b). The difference between mixture versions has to do with the inter-source energy ratio. The sources in `melody_mix1` are made out of MIDI-triggered anechoically recorded samples, whereas `melody_mix2` contains synthesised audio instead. In this way, this group of melody mixtures can cover a variety of cases: a combination of different pitches, note lengths, types of sources (different instruments, and real/synthetic) and different energies. The instruments used in `melody_mix1` are E♭ clarinet (s_1), piano (s_2) and French horn (s_3), while for `melody_mix2`, the instruments are alto saxophone (s_1), flute (s_2) and tenor saxophone (s_3). Finally, it should be pointed out that the difference in RMS energy (as measured in dB) between source j and source $j+1$ in a mixture will be referred to as:

$$\Delta E \;\equiv\; \Delta E(s_j, s_{j+1}) = 10 \log_{10} \frac{\|s_j\|^2}{\|s_{j+1}\|^2}, \quad j = 1, 2, \ldots, J-1. \tag{5.31}$$



(a)



(b)

Figure 5.19: The 'pianoroll' representations of (a) `melody_mix1` and (b) `melody_mix2`. Different colours represent different sources: black, grey and white coloured bars correspond to s_1, s_2 and s_3, respectively.

It is acknowledged that the polyphony is not absolutely constant in `melody_mix1`: in the short interval between 16 s and 18 s it is $O = 2$, whereas for the rest of the mixture it is $O = 3$. Because of this, the particular mixture does not strictly belong to Case 2; however, it will be assumed by the system in this thesis that it does. The first reason for this is that `melody_mix1` is based on an excerpt of *Cavatina*, a quite popular music piece. As such, it would be a helpful example to use it as it stands for the evaluation of the system. The second reason for assuming constant polyphony for this mixture is to see how the system copes with small errors in this kind of prior information.

The F0 track disentangling process operates at a level which can be called 'sustained note interval' (*i.e.*, intervals that are expected to be containing only sustained parts of notes). The way to move from the ability to process Case 1-type mixtures to further cases is achieved by also providing the note onset and offset timings as additional prior information, apart from the polyphony. A simple process then segments the time-domain mixture into sustained note intervals in which all the subsequent processing will take place individually. Additionally, a process is included for the purpose of further improving F0 estimates: for a single note that exists within more than one interval, there is the chance that more than one different F0 value has been estimated. When this situation arises, the F0 value corresponding to the longest interval is then selected as the most reliable one to represent the whole note. The decision for this is based on the assumption that the longer an interval is, the higher the success of the disentangling stage will be in providing reliable F0 estimates for a particular note.

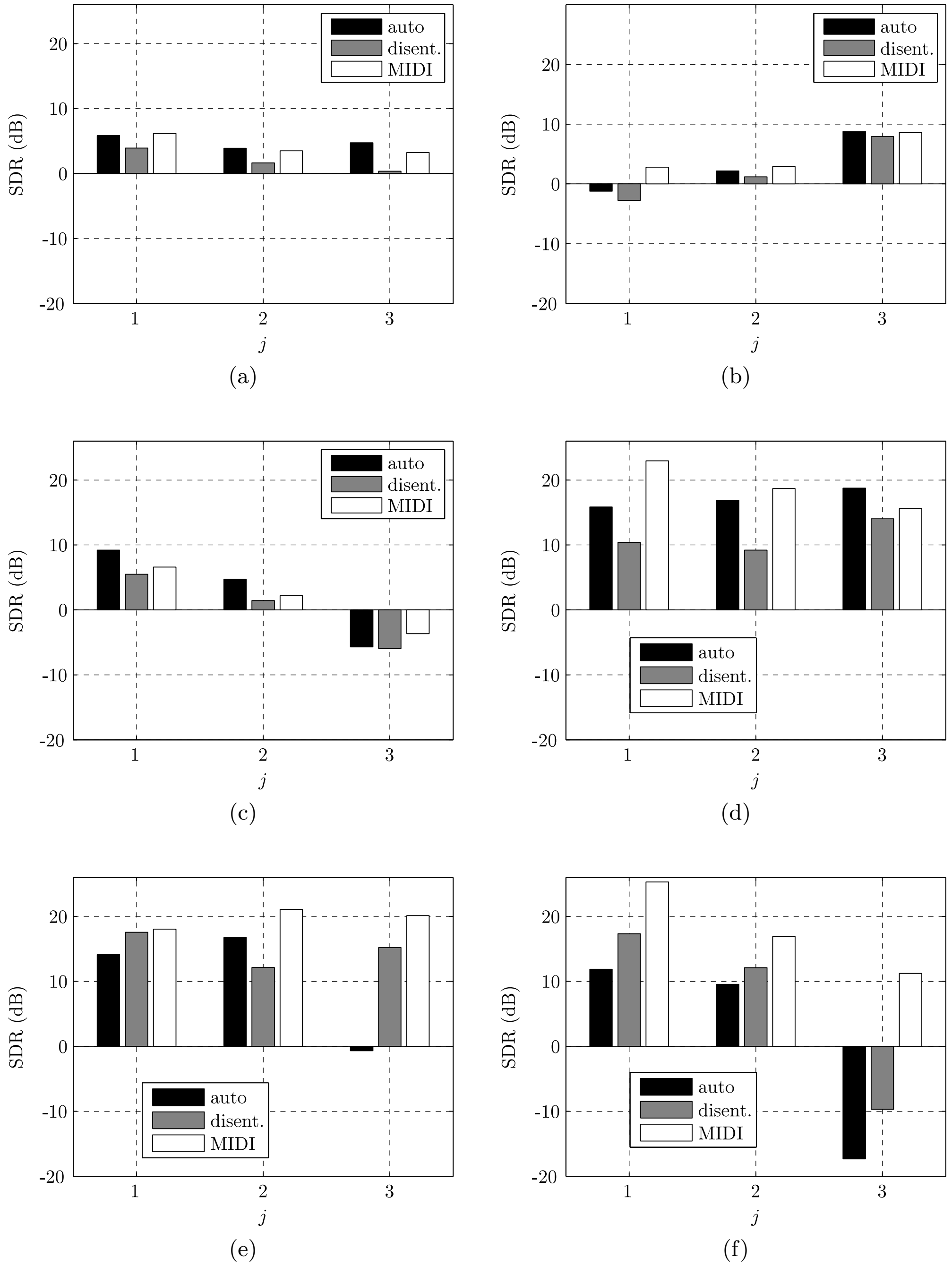
Ideal frame-by-frame grouping will be used in all cases involving melody mixtures. This is because the current system does not provide a method for clustering the extracted structures into sources. What will be examined primarily in the separation evaluations for these mixtures is the ability for the system to successfully identify and extract all existing source structures in them. The use of ideal grouping means, however, that only the F0 correcting process will be assessed within the F0 disentangling stage, and not the F0 swapping one.

Figs. 5.20 and 5.21 present performance evaluation comparisons for the 6 melody mixtures in terms of separation and multiF0 estimation, respectively. The multiF0 estimates were compared this time with ground-truth estimates provided by the YIN algorithm [41]. This allows for a more meaningful and independent basis for comparison. The audio results referring to these comparisons can be listened to on the web at [150].

First of all, it can be seen that, on the whole, the performance in multiF0 estimation reflects the relative SDR levels – something that shows the importance of a reliable multiF0 estimation stage in producing high separation performance. It is also clear by observing the results for different ΔE that a lower RMS energy for a certain source (compared to the other sources) leads to deterioration of both its separation and multiF0 estimation performances.

Further, it can be seen that lower performance is exhibited on average for `melody_mix1` compared to `melody_mix2`. This is to be expected as `melody_mix1` is a much more complex mixture, containing a multitude of short piano notes, the attacks of which add a considerable noise element to the processed segment. In this type of complexity, the F0 track disentangling stage does not provide improvement; in fact, there are cases where it contributes to slight performance deteriorations. A first factor contributing to this is the use of highly-overlapped long windows, which can lower performance when fast percussive instruments (like the piano, here) are included.

Secondly, because of the short note durations, the intervals in which the disentangling stage operates are very short: specifically for `melody_mix2`, their length ranged from just 3 to 54 frames. In comparison to the 200 frames of the mix in Fig. 5.18 and the 70 to 100 frames of the mixes in 5.17 the intervals of `melody_mix2` are much shorter. This combination of very short intervals with long window frames not only compromises the accuracy of the system in correcting the estimated F0s, but also leads to more errors. A way forward from this can be to use shorter windows – provided that the detection and identification of source struc-



(a) (b) (c) (d) (e) (f)

Figure 5.20: Separation performance of the system presented in this chapter ('auto', 'disent.') compared to its previous version by Every ('MIDI') in terms of the SDR. 'auto' and 'disent.' refer to an automatic multiF0 estimation front-end, without and with F0 track disentangling, respectively. The comparisons are presented individually for every source $j \in [1, 3]$. (a): `melody_mix1`, $\Delta E = 0$ dB; (b): `melody_mix1`, $\Delta E = -5$ dB; (c): `melody_mix1`, $\Delta E = 5$ dB; (d): `melody_mix2`, $\Delta E = 0$ dB; (e): `melody_mix2`, $\Delta E = 6$ dB; (f): `melody_mix2`, $\Delta E = 10$ dB.

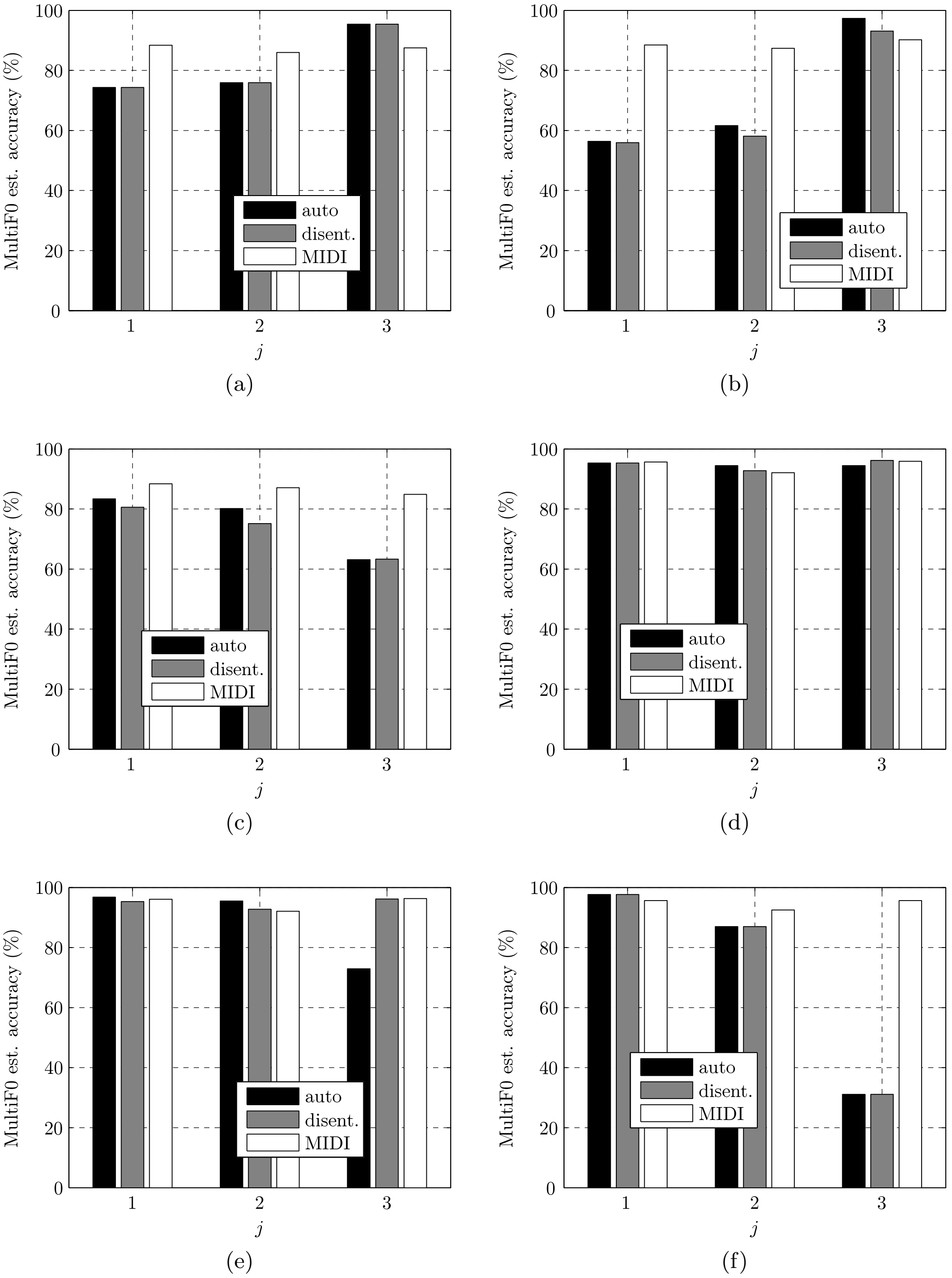


Figure 5.21: Performance of the system presented in this chapter ('auto', 'disent.') compared to its previous version by Every ('MIDI') in terms of its MultiF0 estimation accuracy. 'auto' and 'disent.' refer to an automatic multiF0 estimation front-end, without and with F0 track disentangling, respectively. The comparisons are presented individually for every source $j \in [1, 3]$. (a): `melody_mix1`, $\Delta E = 0$ dB; (b): `melody_mix1`, $\Delta E = -5$ dB; (c): `melody_mix1`, $\Delta E = 5$ dB; (d): `melody_mix2`, $\Delta E = 0$ dB; (e): `melody_mix2`, $\Delta E = 6$ dB; (f): `melody_mix2`, $\Delta E = 10$ dB.

tures can be adequately robust even with shorter windows, the F0 disentangling could prove to be a valuable addition.

Finally, for the case of melody_mix1 the proposed system shows comparable performance to Every's MIDI version, in contrast to melody_mix2 where Every's system is superior. This is unsurprising, since the MIDI-informed version is expected to perform very well with steady, synthesised sounds. However, when real sounds are involved in a much more complex context (such as in melody_mix1), the MIDI-informed version does not show significant superiority against the proposed system that uses the automatic front-end. Additionally, although the automatic system operates with a slightly erroneous prior information, it still shows quite a comparable performance to Every's system. This illustrates that although the basic automatic approach and the previous MIDI-informed approach both have their limitations, their overall performance on real mixtures is similar. This not only allows the removal of a human operator from the front-end of the process, but also, with possible further enhancements arising from the use of the residual signal (discussed in Ch. 6), the automatic approach offers increased potential for separation of real mixtures.

5.9 Summary

This chapter presented a study of the proposed separation approach for single-channel musical mixtures. As a starting point, the new approach is based on an existing non-blind method that makes use of a MIDI front-end for identifying source structures. Because one of the main limitations of non-blind methods is the high degree of user intervention (which itself may require a significant degree of expertise and/or specialist knowledge from that user), in the new method the MIDI front-end is removed and replaced by an automatic multiF0 estimator. This estimator has been chosen because of its sufficiently low-error performance in a variety of polyphonies [98]. A further modification to the previous system was introduced at the source parameter estimation stage; the new system includes the ability to estimate the amplitudes of F0 peaks that might overlap with others during the total duration of a note, through linear extrapolation from the adjacent higher harmonics.

Possible F0 estimation errors arising from the new, automatic front-end to the system were dealt with through the use of a F0 track disentangling stage. This stage is designed to sort the F0 estimates into F0 tracks, with each track corresponding to an individual source. In the process, it is also intended to refine, and improve upon, the initial F0 estimate for each source. Assuming that the note timings are known (or that the mixture consists of synchronous notes and of the same duration) and that the multiF0 estimator gives correct estimates for the majority of frames for the duration of the note, it was shown that most of the erroneous estimates were improved, at least for the case of mixtures containing synchronous single-note sources. Although there was evidence that the corrected estimates could potentially be improved further, they were still within a half-semitone of their reference MIDI pitch.

When applied to mixtures of melodies (Case 2 mixtures) the F0 track disentangling stage appeared, however, to cause a slight lowering of performance in most cases. This was attributed to the combination of the long, highly overlapping, frames used in this system, the fast non-harmonic spectral content around note transitions, and the relatively short processing intervals. For interweaving melodies, the processing is carried out separately within sustained note intervals which are the result of time-segmenting the mixture. The intervals in the particular examples used here were much shorter compared to those for the Case 1 mixtures, which were essentially equal to whole note lengths, with an average duration of 2 s. Shorter intervals mean that less information is available to work with and if, at the same time, this information has already been compromised, as can happen when transitory note energy appears within with long frames, this is not a favourable situation for the F0 track disentangling stage.

Furthermore, in order to choose the most suitable performance measure, a comparison was carried out between two quite popular measures, the SDR and SRR; this was made in both a theoretical and practical framework, and it confirmed the superiority of the SDR. However, this does not mean that it is the most appropriate metric for the evaluation of this system. Although its development is beyond the scope of this thesis, an alternative metric which might reveal more consistent performance could, for example, take into account only the parts of the mixture that are known to be consistent with the system's assumptions.

The results of the comparison between the old system and new systems indicate that the use of a good-quality automatic multiF0 estimation stage (coupled with an F0 track correction and refinement process) can lead to a highly comparable (if not better) separation performance, than that obtained using a MIDI-based front-end. In fact, especially for the case of single-note sources, the proposed approach showed a somewhat more consistent behaviour, revealing some of the problems of a MIDI-based front-end. In particular, the automatic system is not susceptible to user error, in the form of misleading or inaccurate pitch information. That is not to say that the automatic approach would not benefit further from additional user-input however; even approximate additional information would be helpful in situations where the melodic structure is too complicated for reliable F0 estimation to be carried out. The key point, however, is that if user intervention is not an option for the application at hand, the proposed automatic front-end provides a viable approach.

There is room for performance improvement and achieving less reliance of the system to prior information. The next chapter will look at further extending the capability and flexibility of the automatic approach through exploiting the residual channel as an extra source of information and, simultaneously, introducing a feedback loop which passes system-derived information back to inform the separation process, hence moving towards a system that offers both separation and understanding (i.e., a SAU system) via an iterative framework.

CHAPTER 6

Residual-based system extension and improvement

One of the primary goals of this thesis is to propose a SAU system: a system where the processes of mixture separation and understanding support each other in an iterative framework, with the purpose of providing possible solutions and outputs for a range of post-processing applications, while relying rather less on user intervention than earlier approaches. The previous chapter described how an automatic one-pass version of such a system can be realised. This chapter describes the introduction of a feedback loop between separation and understanding, and in this sense establishes the structure of a system based on iterative improvement. A basic component of the feedback loop is the use of the residual channel.

After an introduction describing how to interpret the residual in new ways, two specific tasks are presented where it could be of particular use: multiF0 estimation correction (and consequently improvement of separation) and note onset detection. System evaluation results, as well as comparisons with alternative existing methods are presented and discussed for both cases and for a variety of mixtures. The results indicate the strong potential of the residual as a source of information for improving a system of this kind. Finally, a brief discussion is given of an additional possible use of the residual when handling stereo mixtures.

6.1 The residual channel

The residual, after the first separation pass can be simply defined as the subtraction of the synthesised signal from the original mixture in the time domain:

$$x_{\mathsf{res}}(n) \;=\; x(n) \;-\; \sum_{j=1}^{J} \hat{s}_j(n). \tag{6.1}$$

Very little work has looked at the residual signal resulting from a separation process as a means of useful information. For example, Every [59] used the residual for separating overlapping broadband noise content expected to be found there (assuming that all the pitched part of the original mixture has been extracted), while others have used it in somewhat more trivial ways: for example, re-distributing its content to the extracted sources based on a measure that indicates the likelihood of a source to be the only active source at a certain TF point, or considering it as one of the sources [52]. These are indeed viable ways to consider the residual for the purpose of improving the performance of a separation system. Additional ways, however, also exist. The rest of this section will re-establish the idea of the residual as a means of providing various types of information about the mixture – in a sense considering it as a more 'active' channel in the separation process.

6.1.1 Re-interpreting the content of the residual

We will see that the residual is a channel that can provide useful information which can improve the performance of the system, in terms of both separation and understanding processes – thus realising a SAU separation system. Especially for an unsupervised semi-blind separation system (such as the one presented here) this information can be crucial for constraining the solution set. Indeed, in a considerable number of cases, its role is not just to provide the information for the refinement of some largely satisfying results, but actually preventing the system from almost completely failing in certain tasks. It could be said that it enhances a number of existing signal features, while enabling the existence of others. These features can potentially provide clues for the quality of the multiF0 estimation and the adequacy of the underlying models and assumptions, enhance processes such as the automatic onset detection of the extracted sources, or enable novel methods for the extension of this system to stereo mixtures.

In order to make effective use of the residual channel, it is required to have a knowledge about the kind of content one would expect in there. The shift from a non-blind to a semi-blind approach – as it is described in this thesis – demands a reconsideration of how to interpret the content of the residual channel: its content is not 'unwanted' any more. The first step in order to do this will be to focus on the extraction process – the way in which the residual is actually created. Because of the particular extraction method used here, any identified (harmonic or near-harmonic) source content is expected to be reliably extracted from the mixture, leaving no remnants of its main-lobe part at the residual. This has the result that any harmonic or near-harmonic energy detected in the residual will most probably not be the result of erroneous extraction, but of the identification process. Having that in mind, the following cases can be distinguished regarding the content of x_{res}, mentioned in order of importance for the present approach:

- Harmonic content as a result of erroneous multiF0 and/or parameter estimation of the harmonics. (The last one occurs, *e.g.*, when overlapping occurs for the whole duration of the note.)
- Harmonic content that is in a sense the γ-th version of the original mixture, containing $J - \gamma + 1$ sources (see the residual-using algorithms in §6.2).
- Nonharmonic content of a fast, impulsive nature (onset transients of harmonic notes, and percussive nonharmonic sources).
- Nonharmonic content of structured broadband-noise nature. Examples of this content is structured noise that signifies, *e.g.*, breathiness (in the flute), or the scratch of the bow (bowed string instruments, such as the cello). It has to be noted here, though, that the fact that the existence of this particular content in the residual depends a lot on the recording conditions: *e.g.*, the further the microphone is placed from the source of the sound, the

'purer' the recorded sound will be, in the sense that the broadband noise content will not be picked up.

- Harmonic content that belongs to a source that does not fit the existing source models; this means that its existence as part of the residual channel is to be expected.
- Unassigned harmonic content (content that was not intended to be in the mix, and as a result does not fit the definition of a source; for example, the sound of a squeaking door closing during the live recording of a musical piece).

The degree of importance attached to the above cases is related to the size of their effect on the performance of the system as a SAU one and the opportunity for improvement of this performance. In particular, errors in the multiF0 estimation and disentangling stage can have the most severe effect on the performance; this is because errors of this sort will lead to a source being partially or wholly undetected, hence not extracted from the mixture, in one or several time frames.

It is important to understand that, in this work, the residual is a channel with a 'shifting identity'; because it is used in an iterative fashion, it is in a sense a 'live channel' whose functionality changes according to the current processing goal and what it contains. For example, as it will be seen in §6.2, the residual can be re-inserted at the input of the system as a different version of the original mixture (2nd, 3rd, $\ldots, \gamma$-th version of the original mixture). These different versions can act as supplementary information for the mixture or a separated source, depending on the iteration step and processing goal of the system.

6.1.2 Exploiting the residual channel as source of information

When observing the original mixture signal in order to identify the sources in it, a structure-recognising process has to be applied, as §4.3.1 explains. In the case where the residual is the observed signal, one might search for structured energy in similar ways as it is done with the original mixture. Since consideration of the residual in the way it is proposed in this thesis is largely unexplored, we will look at simple energy-based identification and feature extraction in the time domain. There are three basic situations that we can come across and make a decision regarding the energy of the residual:

- It is below a certain threshold (no significant residual energy exists for exploiting it further);
- It is above a certain threshold and should be a part of one of the extracted sources;
- It is above a certain threshold and the energy is supposed to belong to the residual (for example, energy associated with note attacks).

The last two cases can make possible the further use of the residual. Two examples of this use is multiF0 error correction and note onset detection, and will be studied next.

6.2 MultiF0 error correction

In this section a method is presented that uses the residual channel as a means for performing multiF0 error correction.

6.2.1 Description of the algorithm

This method operates on mixtures containing synchronous single-note sources (Case 1-type mixtures) only with the knowledge of O_r, or on melody mixtures after they have been time-segmented into sustained note intervals using the additional knowledge of note onset/offset timings.

The assumption used here is that F0 estimation errors after the first pass lead to the existence of harmonic or near-harmonic content in the residual that is associated with those erroneously estimated F0s. According to this, two simple conditions have to be satisfied at the same time in order for the residual to be used in this way:

1. 'something is missing' in one or more of the extracted source channels;

2. there is significant harmonic or near-harmonic energy in the residual.

The residual and the extracted source channels can be observed and searched at different levels, in the same way as the original mixture signal is searched for individual source elements. Thus, the search can be frame-based, note-based, or it can involve the whole signal at once. The search can also take place in different domains: it can be a time-domain, frequency-domain, or TF-domain process. Here, examination of the whole time-domain signal at once is chosen because of its simplicity. In particular, the RMS total energy is used as a feature for deciding whether content that should be at one of the extracted source channels ended up in the residual. According to this, the first condition above is satisfied when:

$$U \equiv \{j \in [1, J] : \hat{s}_j^{\mathsf{rms}} < \kappa_{\mathsf{s}}\} \neq \emptyset, \tag{6.2}$$

In other words, the first condition is satisfied when there is at least one extracted source channel with RMS energy that is lower than a specified threshold κ_{s}. The second condition is satisfied when:

$$x_{\mathsf{res}}^{\mathsf{rms}} > \kappa_{\mathsf{r}}, \tag{6.3}$$

i.e., the RMS energy of the residual has to be higher than a threshold κ_{r} in order to consider the possibility for it to carry un-extracted harmonic or near-harmonic source content.

Before continuing with the description of the proposed algorithm, the additional parameter of the *residual polyphony*, $\mathfrak{O}$, is introduced here. As its name implies, $\mathfrak{O}$ is the polyphony value associated with the residual. Because of this, it is not a quantity that can be defined beforehand (*e.g.*, as prior information); $\mathfrak{O}$ is a product of the iterative procedure that employs the residual, and it changes in parallel with it. Thus, apart from generally being a function of r (as O can be in a general mixture case), it is also a function of another index, the *iteration index* γ. $\mathfrak{O}_{\gamma r}$ takes values in the following interval:

$$0 \leqslant \mathfrak{O}_{\gamma r} \leqslant O_r, \quad \forall \gamma, r. \tag{6.4}$$

Like O, $\mathfrak{O}$ is employed as an input to the multiF0 estimator. The difference is that the value of $\mathfrak{O}$ is not determined by the user, but by the algorithm presented here.

Two alternative versions of the algorithm, the 'multi-single' and the 'iterative multi-single' one are illustrated in Figs. 6.1 and 6.2, respectively. Both algorithms share the same philosophy: in each pass, provided that the above two conditions are satisfied, the residual is fed back to the input of the system along with the updated $\mathfrak{O}$. $\mathfrak{O}$ however is updated differently in the two cases. In the case of the 'multi-single' algorithm, $\mathfrak{O}$ is set to 1 and follows a one-by-one extraction process of all the previously erroneously estimated sources; the implication is that the multiF0 estimation has failed for one or more sources so it is replaced by an iterative single F0 estimation, leading to the extraction of the most salient source each time. In the case of 'iterative multi-single' algorithm there is an additional option: $\mathfrak{O}$ is set to 1 only when the multiF0 estimation has failed for all the sources, otherwise $\mathfrak{O}$ is set to $|U|$. This adaptability of the 'iterative multi-single' version of the algorithm can potentially render it faster than the 'multi-single' one (*i.e.*, it can lead to fewer iterations); this is the reason it is offered as an alternative.

The present method for multiF0 error correction relies on the assumptions that

- the factor mostly responsible for a source to be largely undetectable (and, thus, largely not extracted from the mixture) is the failure of the multiF0 estimation stage; and

```
1 Input $O$ and $x$.
2 Set $\mathfrak{O} \leftarrow O$ and $\iota \leftarrow O$.
3 Extract signals $\{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_{\mathfrak{O}}, x_{\mathsf{res}}\}$ from mix $x$. If $\mathfrak{O} = 1$ go to step 3;
  otherwise, go to step 4.
4 If $\iota = 1$, stop; otherwise set $\iota \leftarrow \iota - 1$ and go to step 5.
5 Calculate $\{\hat{s}_{\mathsf{o}}^{\mathsf{rms}}\}_{\mathsf{o}=1}^{\mathfrak{O}}$ and $x_{\mathsf{res}}^{\mathsf{rms}}$. If both conditions 6.2 and 6.3 are satisfied,
  set $\mathfrak{O} \leftarrow 1$ and $\iota \leftarrow |U|$; otherwise stop.
6 Set $x \leftarrow x_{\mathsf{res}}$ and iterate to step 2.
```

Figure 6.1: The 'multi-single' algorithm that uses the residual for multiF0 error correction.

```
1 Input $O$ and $x$.
2 Set $\mathfrak{O} \leftarrow O$ and $\iota \leftarrow O$.
3 Extract signals $\{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_{\mathfrak{O}}, x_{\mathsf{res}}\}$ from mix $x$. If $\mathfrak{O} = 1$ go to step 3;
  otherwise, go to step 4.
4 If $\iota = 1$, stop; otherwise set $\iota \leftarrow \iota - 1$ and go to step 6.
5 Calculate $\{\hat{s}_{\mathsf{o}}^{\mathsf{rms}}\}_{\mathsf{o}=1}^{\mathfrak{O}}$ and $x_{\mathsf{res}}^{\mathsf{rms}}$. If both conditions 6.2 and 6.3 are satisfied,
  continue to next step; otherwise stop.
6 If $|U| < \mathfrak{O}$, set $\mathfrak{O} \leftarrow |U|$ and $\iota \leftarrow 1$; otherwise set $\mathfrak{O} \leftarrow 1$ and $\iota \leftarrow |U|$
  (*transition to one-by-one extraction*).
7 Set $x \leftarrow x_{\mathsf{res}}$ and iterate to step 2.
```

Figure 6.2: The 'iterative multi-single' algorithm that uses the residual for multiF0 error correction.

- the error rate of the multiF0 estimator decreases with the decrease of $\mathfrak{D}$.[1]

The first assumption has been confirmed in the analysis of various results above (such as the results of Fig. 5.16). The second assumption will be examined to some degree and confirmed through the analysis of the experimental results in the next section.

6.2.2 Initial experimental results

The effectiveness of the algorithm described above will be explored here through a set of experiments. One of the ways to produce challenging situations for both the multiF0 estimation and the extraction stage is to experiment with energy ratios. This can be effectively carried out using ΔE.

The following results involve the comparison of the separation performance of several 2-note mixes, with and without employing the residual in different energy ratio scenarios. In this case ΔE is equivalent to $\Delta E(s_1, s_2)$. It is also noted that, since $J = 2$, the two versions of the algorithm shown in Figs. 6.1 and 6.2 are equivalent. What is examined here, instead, is the viability of the general philosophy underlying both of them and some of the implications associated with the use of the conditions stated in Eqs. 6.2 and 6.3.

The values of $\kappa_\mathsf{r} = -60\,\text{dB}$ and $\kappa_\mathsf{s} = -34\,\text{dB}$ were set after considerable experimentation with a large number of mixtures, made up of two simultaneous notes of equal length, whereas the value of $\kappa_\mathsf{s} = -41\,\text{dB}$ was set for melody mixtures. Figs. 6.3-6.5 show the SDR values for separating the sources in 7 different mixtures of Case 1. The first observation, looking at the left-hand plots, is that if a source is below a particular energy level compared to the interference, the multiF0 estimator simply fails to detect it. (It was found that, for this set of mixtures, the SDR values for a certain source that are below -30 dB correspond to a failure of the multiF0 estimator to detect that source.) This critical value appears to be

[1] It is reminded here that $\mathfrak{D}$ represents the number of sources as the input information to the system. This does not mean that it always reflects the true number of sources existing in the mixture.

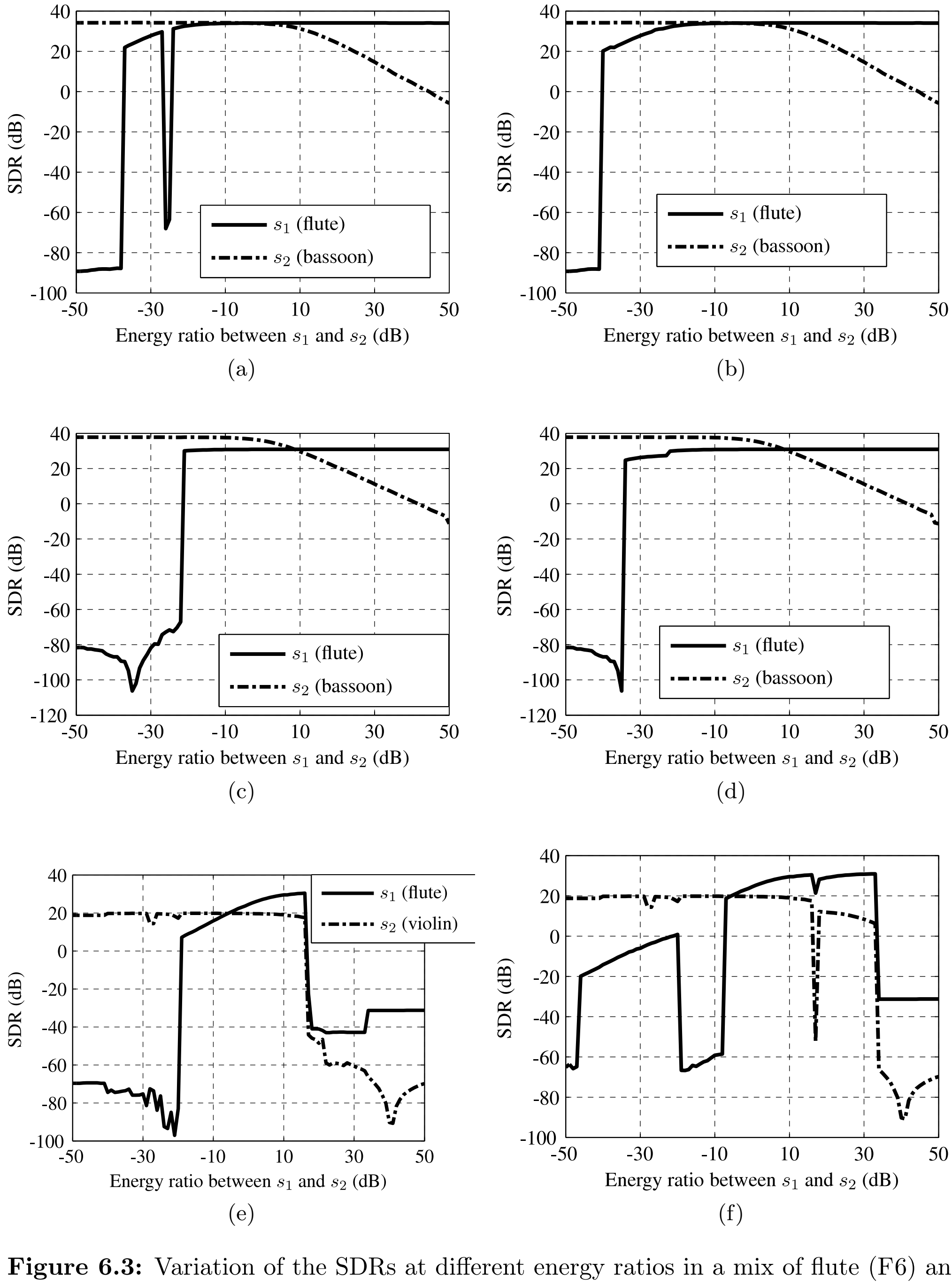
when ΔE is around -10 dB (s_1 is 10 dB quieter compared to s_2) and 10 dB (s_2 is 10 dB quieter compared to s_1) for all the cases, apart from the bassoon sounds in Figs. 6.3a and 6.3c. The reason for not failing for these sounds is that their pitch is a lot lower than the pitch of the other source. For these cases, this has the effect that their lower two partials (the fundamental and its first harmonic) appear with absolutely no interference in their vicinity, rendering them enough to provide a strong detection cue, even if they are very faint. This is, in other words, a particularly favourable situation for the multiF0 estimator.

If we now compare the left-hand plots with the right-hand ones it can be seen that in most cases the proposed method for employing the residual extends significantly the range of ΔE over which the two sources can be reliably detected. In other words, the robustness of the multiF0 estimator and the separation system is improved.

Having said that, there are situations where undesired behaviour occurs. This is due to two factors: the robustness of the conditions for deciding whether successful separation has taken place, and the absence of a sophisticated model for the partial amplitudes. The first factor can have two effects:

- False-positive errors. These errors occur when one of the sources is identified as unresolved, while in reality both of the sources have been correctly separated. This leads to feeding back the residual, forcing the multiF0 estimator to apply the harmonicity assumption on the nonharmonic content. An arbitrary harmonic structure is then extracted from the residual and replaces the (falsely identified as unresolved) source. Examples of these errors can be observed in Figs. 6.3f, 6.4b and 6.4d when ΔE is roughly between -22 dB and -6 dB.

- False-negative errors. These errors occur when an extracted signal is falsely identified as an actual resolved source signal. Examples of this type of error can be observed, *e.g.*, in the extraction of the flute in Fig. 6.3d (when ΔE is between -50 dB and -41 dB), or in the extraction of the saxophone in Fig. 6.4f (when ΔE is between 18 dB and 50 dB). What is mainly responsible



(a)

(b)

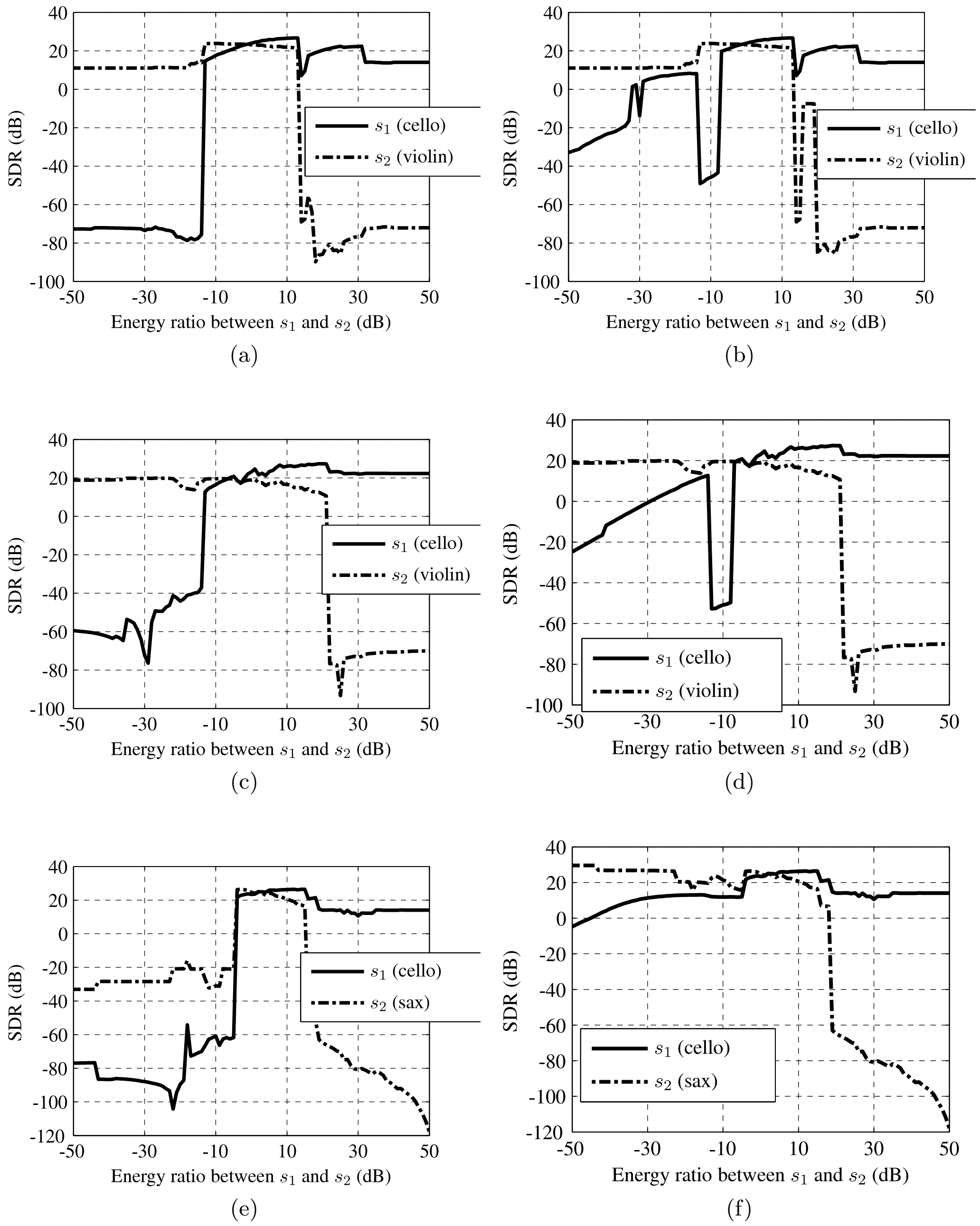
(c)

(d)

(e)

(f)

Figure 6.3: Variation of the SDRs at different energy ratios in a mix of flute (F6) and bassoon (G4) (a) when the residual is not used and (b) when the residual is used; flute (D6) and bassoon (A4) (c) when the residual is not used and (d) when the residual is used; flute (G♭4) and bassoon (D♭4) (e) when the residual is not used and (f) when the residual is used. The energy ratio between s_1 and s_2 is $\Delta E(s_1, s_2) = 10 \log_{10} \frac{\|s_1\|^2}{\|s_2\|^2}$.



(a) (b)

(c) (d)

(e) (f)

Figure 6.4: Variation of the SDRs at different energy ratios in a mix of cello (B3) and violin (B♭4) (a) when the residual is not used and (b) when the residual is used; cello (B♭3) and violin (D♭4) (c) when the residual is not used; and (d) when the residual is used; cello (B3) and soprano sax (A3) (e) when the residual is not used and (f) when the residual is used. The energy ratio between s_1 and s_2 is $\Delta E(s_1, s_2 = 10 \log_{10} \frac{\|s_1\|^2}{\|s_2\|^2})$.

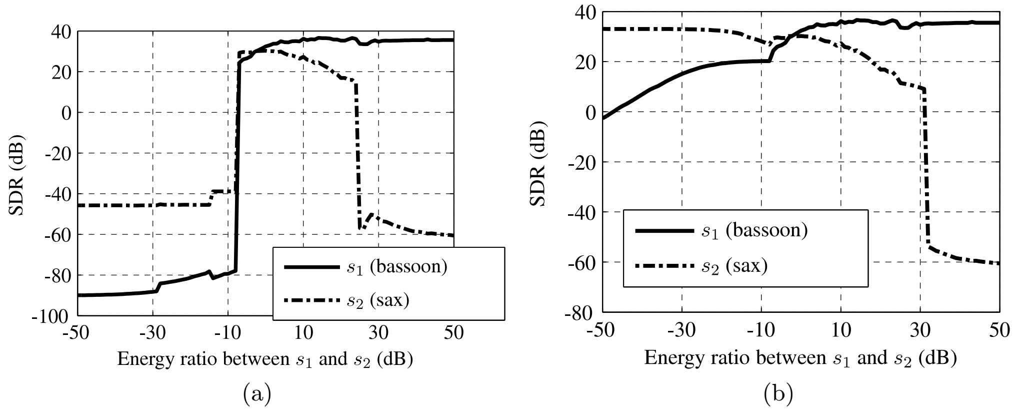


Figure 6.5: Variation of the SDRs at different energy ratios in a mix of bassoon (F3) and soprano sax (A♭3) (a) when the residual is not used and (b) when the residual is used. The energy ratio between s_1 and s_2 is $\Delta E(s_1, s_2) = 10 \log_{10} \frac{\|s_1\|^2}{\|s_2\|^2}$.

for this type of error is the high energy difference between the nonharmonic content of the louder source and the other signal. This has the effect of the loud signal masking heavily the other, much quieter signal, with the result the second signal is undetectable.

A solution to this problems should focus on devising more intelligent conditions for deciding about the content and further use of the residual.

The second factor for degradation of performance is not related to multiF0 estimation error. In this situation the sources have been correctly detected. However, when their energy is starting to be comparable to the nonharmonic broadband energy floor of the other source, artefacts associated with poor spectral amplitude estimation become more noticeable. In fact, the amplitudes of the higher frequency partials are below the noise content of the previously extracted dominant source. Since no knowledge of timbral structure is provided and no other source is considered to be there during the second pass ($\mathfrak{O}$ is 1), the interference is leaked on the extracted quieter source. The results of the bassoon sounds on the left-hand side of the Figs. 6.3b and 6.3d or the cello sounds on the right-hand side of the Figs. 6.4d and 6.4f are examples of this type of degradation. A solution to this problem for this mixture case would be to provide more advanced spectral source models as prior information.

6.3 Performance comparison

In this section, the residual-based extension of the system is applied to melody mixtures, and compared not only to its one-pass version, but also some alternative approaches. The comparison is carried out in terms of its separation and multiF0 estimation performance on the set of 6 mixtures comprising variations of `melody_mix1` and `melody_mix2` (the same mixtures used in §5.8.2).

The alternative systems chosen for comparing the separation performance are by Duan and Pardo [50] (in which the MIDI front-end has been replaced by a multiF0 estimation and tracking system) and Li *et al.* [107]. Both of them were chosen because they are recent and belong to the same category with the proposed system: they perform semi-blind unsupervised separation and they operate using frame-wise F0 estimates. The researchers were provided with the mixtures and sent back the extracted source signals.

For the case of [50], the extracted signals were ideally grouped on a frame-by-frame basis and this has also been done for the proposed system. This step was taken since they employ an F0 tracking process, while the proposed system does not. Hence, using ideally grouped signals enables a consistent comparison between the systems concentrating on their relative ability for identification and extraction of source structures. As with the proposed system, [50] employ the prior assumption of constant polyphony.

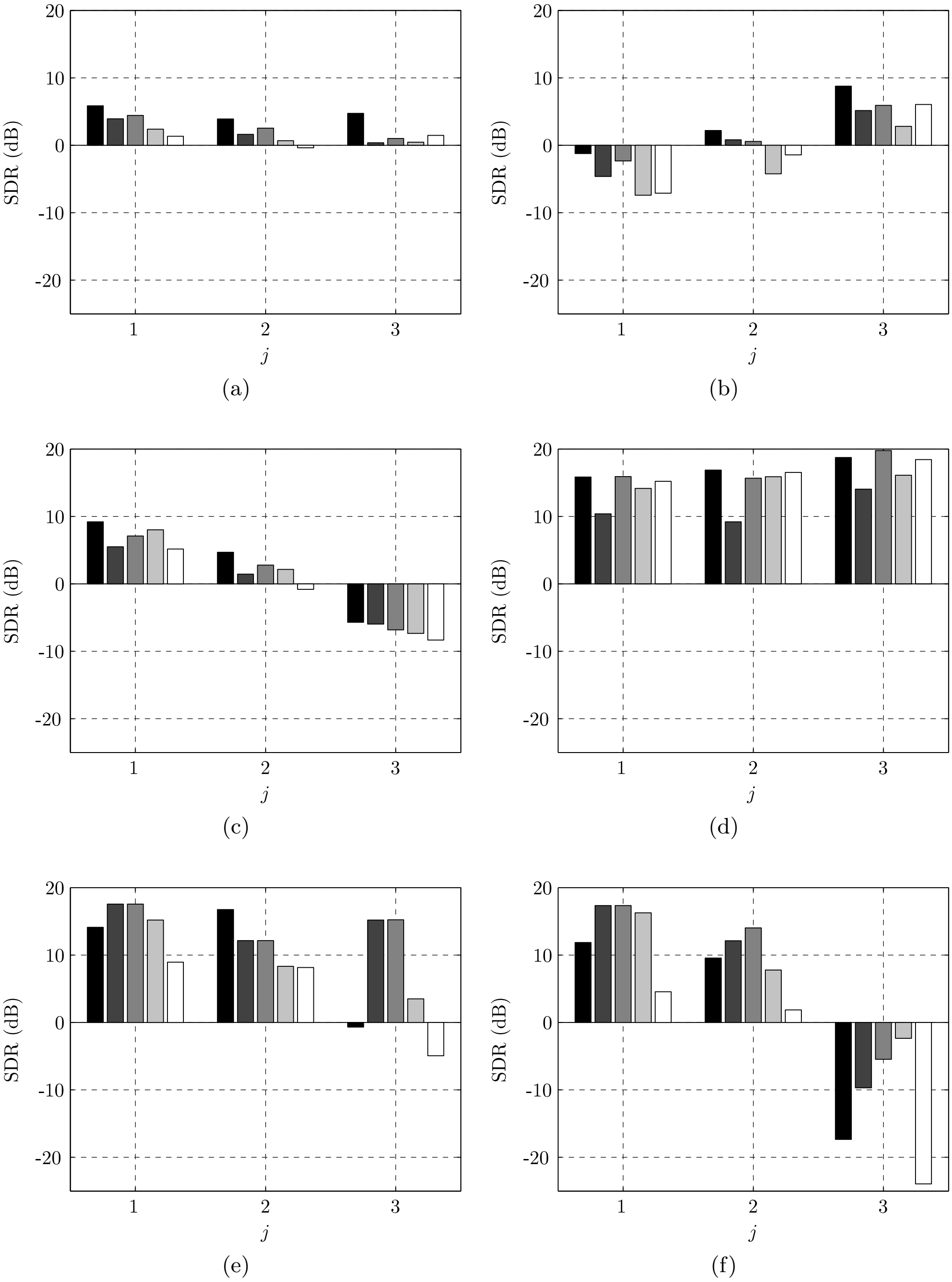
For the case of [107], the authors were provided with additional information in the form of the frame-wise F0 values as they came out of the proposed system's multiF0 estimation stage (before F0 track disentangling), but ideally tracked: the F0s were compared to the ground-truths in order to create F0 tracks that are as close to the original source tracks as possible. For this reason the extracted signals did not need to be ideally grouped in the same way as for [50]. Hence, since the system in [107] is using the same F0 values as the proposed system as its starting point, in this case what is being compared is the relative performance of the systems in terms of F0 refinement/correction, and source identification and extraction.

The multiF0 estimation comparison was carried out using two multiF0 estimation systems which have shown high performances in terms of accuracy at the Fundamental Frequency Estimation & Tracking task of the MIREX 2010 evaluation campaigns. The systems are by Yeh and Roebel [181, 182] and Duan *et al.* [49, 51]. The authors were provided with the mixtures and sent back the frame-wise F0 estimates in steps of 1024 samples. While [51] accepts the frame-wise polyphony as prior additional input, [182] does not; Instead, it automatically infers the polyphony. This has the effect of producing a variable number of F0s for each frame and, since only the accuracy of the F0 estimation is being evaluated here, the ground-truth is compared (in a frame-wise fashion) with all of the estimated F0 values until matches are found.

Figs. 6.6 and 6.7 show the source separation and multiF0 estimation results, respectively. Additionally, the related audio files can be listened to on the web at [150].

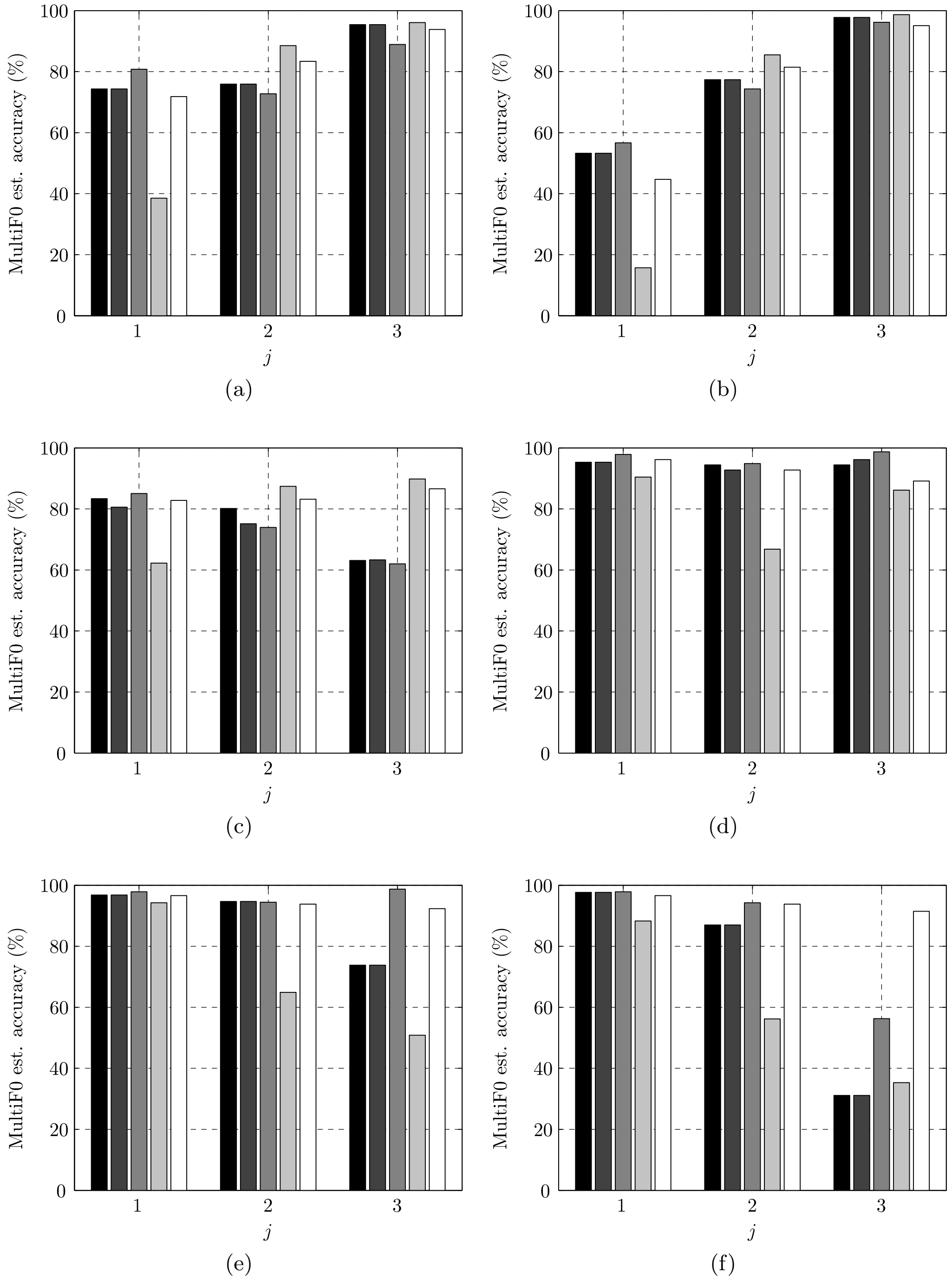
It can be observed that the use of the residual (column three in these graphs) increases both the SDR and multiF0 estimation accuracy in most of the cases compared to the performance of the system after the F0 track disentangling stage. This indicates that the residual feedback contributes in counterbalancing the errors caused by that stage (mostly in `melody_mix1`) and in some cases leads to overall improvement (certain cases in `melody_mix2`).

Looking specifically at the cases of overall improvement, the tenor saxophone (s_3) part of `melody_mix2` consists of three notes, two of which are the longest ones in the mixture (Fig. 5.19b). This is an easier situation for the proposed system, as it has more frames to operate on. On the other hand, `melody_mix1` is a more complex mixture containing real sounds with fast percussive elements, such as the short piano notes, making it a more challenging case for the system, as can be seen from its performance both in SDR and multiF0 estimation accuracy.



(a) (b) (c) (d) (e) (f)

Figure 6.6: Performance of the residual-based system compared to [50] and [107] in terms of its separation performance. Different cases of the proposed system are compared with two other methods: proposed system without F0 track disentangling (first column, in black); proposed system with F0 track disentangling; proposed system with residual loop; Duan and Pardo [50]; Li *et al.* [107] (final column, in white). The comparisons are presented individually for every source $j \in [1, 3]$. (a): `melody_mix1`, $\Delta E = 0$ dB; (b): `melody_mix1`, $\Delta E = -5$ dB; (c): `melody_mix1`, $\Delta E = 5$ dB; (d): `melody_mix2`, $\Delta E = 0$ dB; (e): `melody_mix2`, $\Delta E = 6$ dB; (f): `melody_mix2`, $\Delta E = 10$ dB.



(a) (b) (c) (d) (e) (f)

Figure 6.7: Performance of the residual-based system compared to [51] and [182] in terms of its multiF0 estimation accuracy. Different cases of the proposed system are compared with two other methods: proposed system without F0 track disentangling (first column, in black); proposed system with F0 track disentangling; proposed system with residual loop; Duan *et al.* [51]; Yeh *et al.* [182] (final column, in white). The comparisons are presented individually for every source $j \in [1, 3]$. (a): `melody_mix1`, $\Delta E = 0$ dB; (b): `melody_mix1`, $\Delta E = -5$ dB; (c): `melody_mix1`, $\Delta E = 5$ dB; (d): `melody_mix2`, $\Delta E = 0$ dB; (e): `melody_mix2`, $\Delta E = 6$ dB; (f): `melody_mix2`, $\Delta E = 10$ dB.

It is important to note that the proposed system shows consistently competitive and, in most of the cases, better performance compared to the other systems, both in terms of separation and multiF0 estimation accuracy. Particularly in comparison to the system by Li *et al.* (white bars in Fig. 6.6), in the vast majority of the cases the proposed system performs better. Since the same F0 values have been employed as a starting point, this difference in separation performance indicates that the proposed system employs a more effective process for F0 correction and identification/extraction of source components (although it is not clear here how each of these parts contributes to a better performance). One further comment, with regard to the comparison with the system by Yeh and Roebel [182], is that because their method provides a lot more than three F0 estimates to be compared with the ground-truths, the chances for it to have a better accuracy and appear less sensitive to low ΔE are increased. For all the rest of the cases shown in Fig. 6.7, three F0 estimates are consistently provided per frame.

Finally, one important difference between the proposed system and the others is that it makes use of the note onset/offset timings as additional prior information when F0 track disentangling and the residual feedback are used. As a further step, this information might be provided in an automatic manner using an iterative residual-based philosophy such as the one described in §6.4.

A further note on the use of the SDR

It should be emphasised that there is no reliable way to assess separation quality, and that the SDR is still just one possible statistic that can be used to provide some insights. Although the SDR is used within this work to as a tool for assessing the relative performance of several approaches when applied to the same mixture, it may be misleading when used to compare results obtained from different signals.

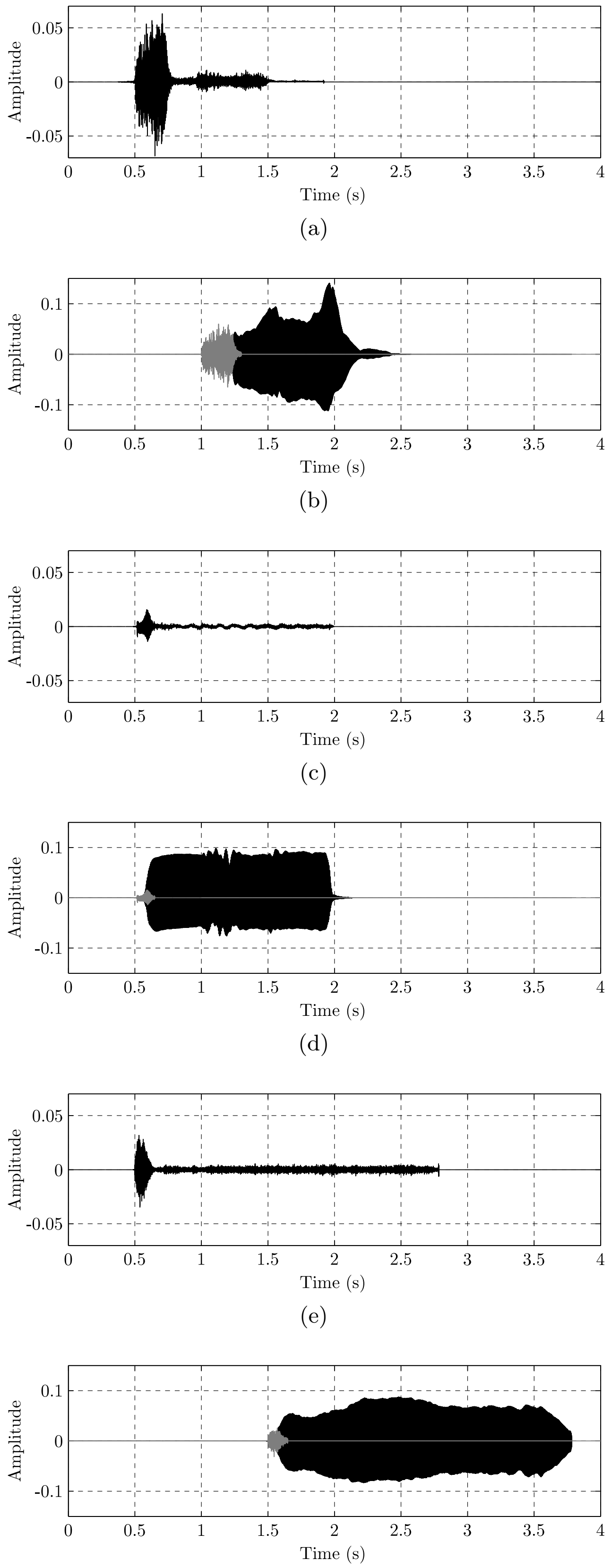
This is due to the fact that each signal potentially contains different melodies with different instruments and different rhythmic structures. In short, one signal may be intrinsically more “difficult” than another in terms of its complexities – such as the number, rate, strength and duration of note attacks. The SDR measure, however, merely provides a statistic integrated over the entire signal length – there is no normalisation process which takes into account what proportion of each signal is going to contribute to reducing the overall SDR.

One possible approach to obtaining a statistic which better represents the "average quality of separation, where separation was reasonably possible" might be to only consider contributions to the SDR from frames that are identified as not containing a note attack. This, and the development of other separation performance measures, are beyond the scope of this thesis, but are considered to be a priority area for further work.

6.4 Enhanced onset detection

A different way to make use of the residual is to exploit the fact that it is the channel where the unmodelled content of the mixture ends up naturally. If harmonicity or near-harmonicity is used as the primary part of the source model and the mixture consists of discrete musical note events (preferably no voice), the signal content that ends up in the residual channel can provide access to valuable timing information. Fig. 6.8 shows what is contained in x_{res} after passing three isolated notes through the separation system, and how this content is compared with its original version. Among the observations that can be made about these, there are three particular ones to point out:

- The remaining energy in x_{res} is generally very small compared to the energy of the original signal, except from well-defined time segments associated with nonlinear sound generation processes taking place during the beginning of the note. These segments are largely related to the nonharmonic attack portion of the note.
- The attacks of the three instruments exhibit different behaviour in terms of their amplitudes, durations and envelopes.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 6.8: The residual channel after extraction of the harmonic part of 3 signals: (a) cello F4, (c) soprano saxophone A3, and (e) violin E4. The same signals also appear in the context of the original notes in (b), (d) and (f), respectively. (Note the scale difference in the amplitude axis.)

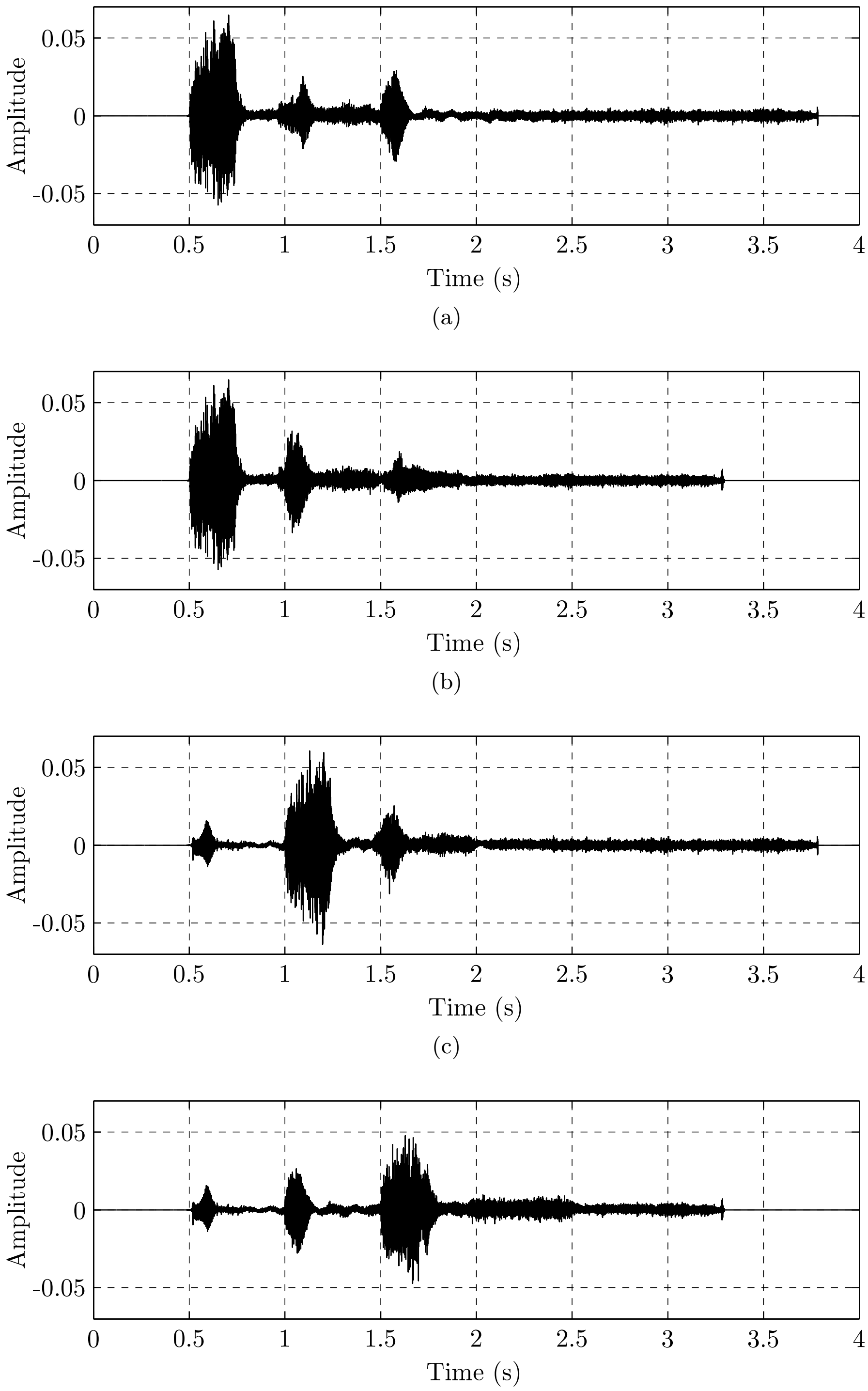
- In every case there is poor correlation between the duration of the separated attack temporal envelope and the parameters estimated using the commonly used Attack-Decay-Sustain-Release (ADSR) approximation.

Note onset detection – which is used in automatic musical beat/metre analysis, as well as other MIR-oriented applications – involves the often significant problem of detecting multiple attacks in the presence of other sounds. Indeed, for such purposes, the fast nonharmonic elements of the musical mixture are actually the *ideal* signals to begin with, because they contain all that is needed for obtaining note-onset information (in the same way that signals containing only harmonic elements are the ideal signals to begin with when performing multiF0 estimation). Through the availability and the particular use of the residual, the system presented here addresses this problem by delivering considerably pronounced attacks compared to the original mixture signal.

Figs. 6.9 and 6.10 show the residual signal after separation of three non-synchronous notes of equal RMS energy from a mixture, while Figs. 6.11 and 6.12 show the residuals in comparison with their respective original mixtures. (The notes are the same ones whose residuals appear isolated in Fig. 6.8.) All the different combinations are tried in terms of order of appearance for the notes, in order to observe how the context might affect the extraction[2] of the residual each time and, as a consequence, how the accuracy in note onset detection might be affected.

In Figs. 6.9 and 6.10 it can be seen that for this set of mixtures the broadband noise content of the cello note is responsible for the highest level of interference on the attacks of the other notes when it precedes them; in those cases, the attacks of the saxophone and the violin appear less accentuated. Having said this, though, and as Figs. 6.11 and 6.12 can reveal, the residual channel can still provide on the whole substantially improved time-domain information compared to the original mixtures.

[2] Note, here, that the retrieval of the residual is referred to as 'extraction'. The use of this term, again, highlights that the residual is something desirable: we *want* to identify and extract these structures.



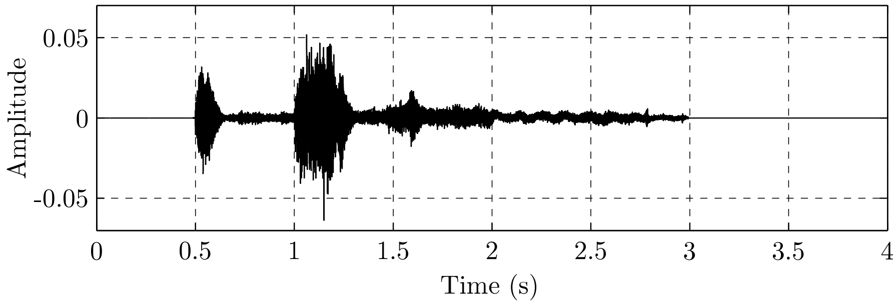
(a)

(b)

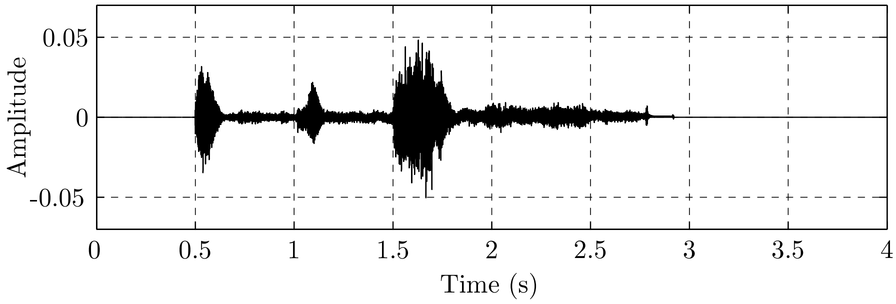
(c)

(d)

Figure 6.9: The residual after source separation from a mixture containing three musical notes played in 0.5 s intervals. The notes are cello F4, soprano sax A3 and violin E4, and they appear in four different orders in the mixture: (a) cello-sax-violin, (b) cello-violin-sax, (c) sax-cello-violin and (d) sax-violin-cello. The identification of the observable note attacks can be used for estimating the note onset timings at 0.5 s intervals.

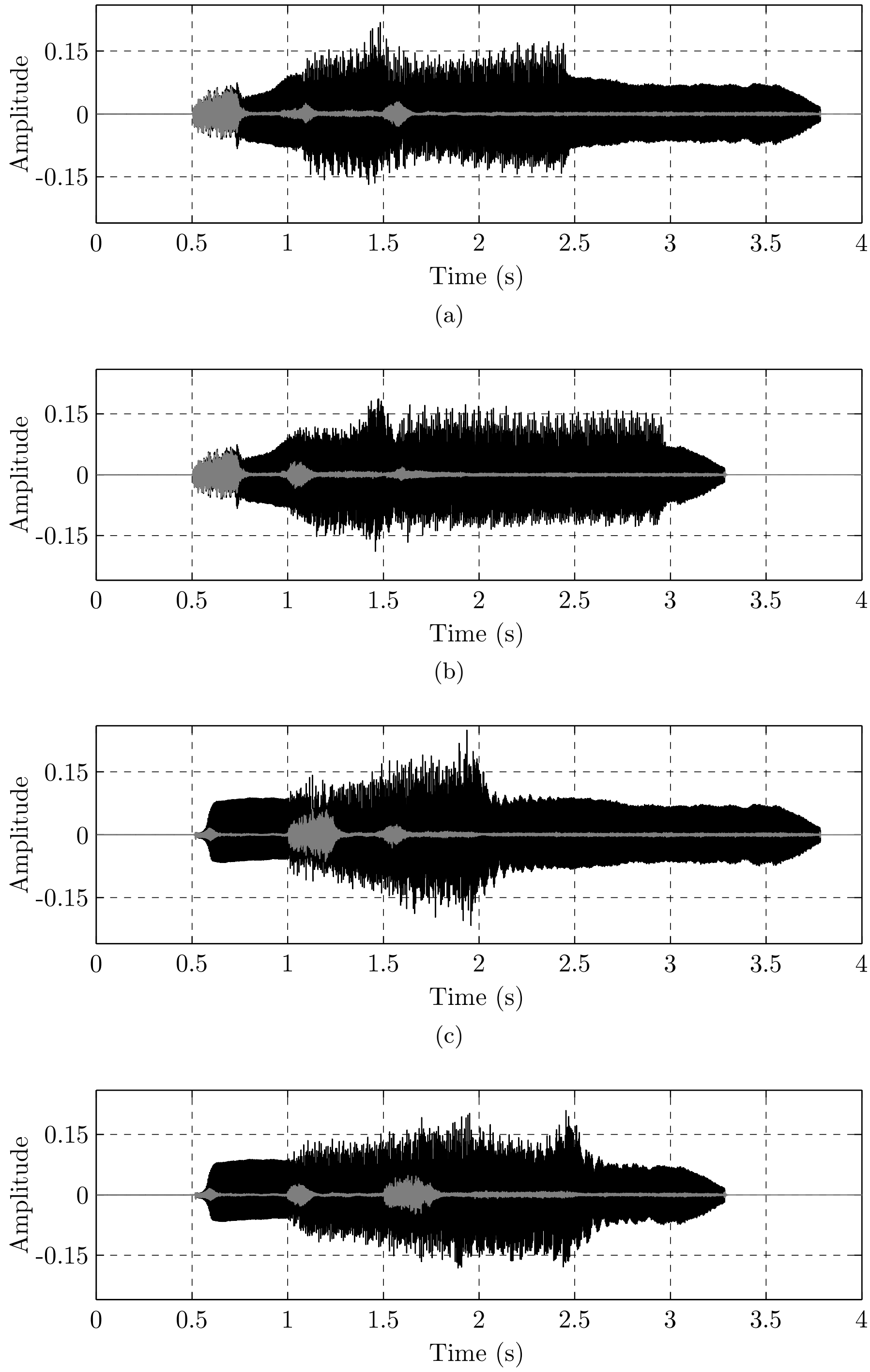


(a)



(b)

Figure 6.10: The residual after source separation from a mixture containing three musical notes played in 0.5 s intervals. The notes are cello F4, soprano sax A3 and violin E4, and they appear in two different orders in the mixture: (a) violin-cello-sax and (b) violin-sax-cello. The identification of the observable note attacks can be used for estimating the note onset timings at 0.5 s intervals.



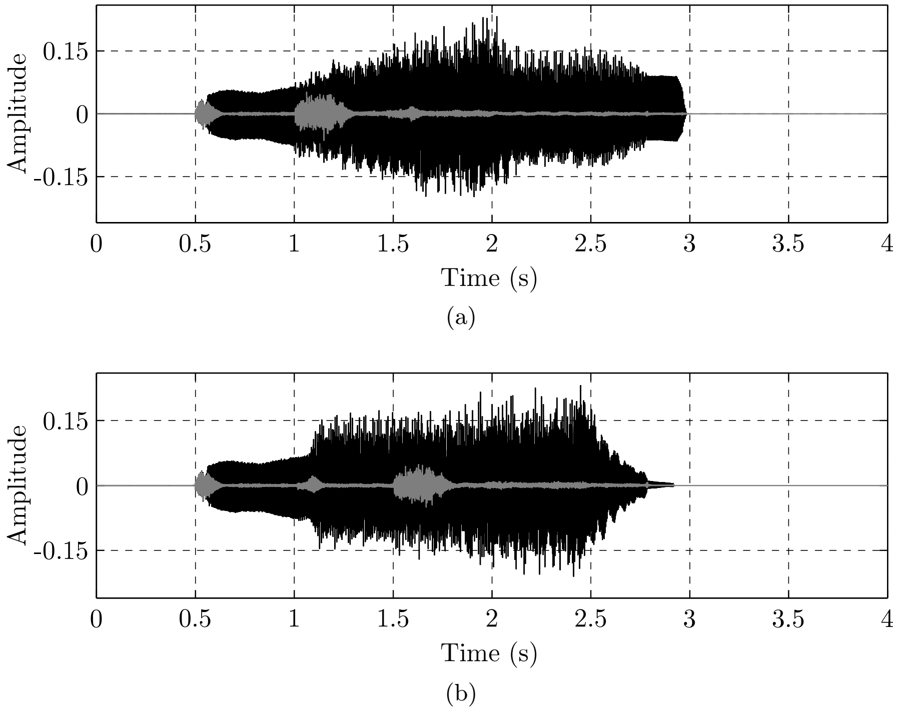
(a)

(b)

(c)

(d)

Figure 6.11: Comparison between the original mixture containing three musical notes played in 0.5 s intervals (shown in black) and the residual signal (shown in grey) after extracting the sources. The notes are cello F4, soprano sax A3 and violin E4 and they appear in four different orders in the mixture: (a) cello-sax-violin, (b) cello-violin-sax, (c) sax-cello-violin and (d) sax-violin-cello. Note the amplitude scale difference between these figures and Fig. 6.9.



(a)

(b)

Figure 6.12: Comparison between the original mixture containing three different musical notes played in 0.5 s intervals (shown in black) and the residual signal (shown in grey) after extracting the sources. The notes are cello F4, soprano sax A3 and violin E4 and they appear in two different orders in the mixture: (a) violin-cello-sax and (b) violin-sax-cello. Note the amplitude scale difference between these figures and Fig. 6.10.

6.4.1 Onset detection method

Onset detection algorithms (*i.e.*, algorithms for locating the instances when note events begin) [8, 32] often define a type of an Onset Detection Function (ODF) and then peak picking has to be applied on that function, in order to determine the timings of the onsets (*e.g.*, [158, 186, 42, 12, 61, 15]). However, to the knowledge of the author, there has not been a method that operates with the residual. Also, transient/steady-state extraction methods do exist and can be used for onset detection [38], but they have been mostly applied to monophonic signals. An onset detection method that operates entirely on the residual (resulting from extracting three sources from a mixture) is presented next.

The majority of the current onset detection methods would not be particularly suitable for operating on the residual, since the input audio is expected to have a quite different overall structure compared to the residual. However, the residual has a clear temporal structure, which is closer to an ODF rather than a time-domain musical signal or mixture. One can easily estimate where the onsets are located just by observing the evolution of the signal in time. Thus, the task is to make this process automatic.

It is worth noting that the proposed method bears similarities with parts of the onset detection method carried out by Every [58, p. 53]. That method, however, works on the original mixture signal using a complex-domain ODF; the philosophy of the method presented here is different: an ODF-like function is derived from the time-domain residual signal.

As a first step, it would make sense to derive a simplified version of the residual, one that preserves the shape of the energy peaks associated with the note attacks, while smoothing out other spurious peaks. To do that, the amplitude envelope of x_{res} is calculated first. The use of the Hilbert transform for acquiring the envelope proved to be adequate. This transformation is followed by taking the absolute value, low-pass filtering and normalisation to unity, resulting in the envelope function $\mathfrak{E}$. The locations of the attacks are identified using a thresholding operation assisted by a weighted median-filtered version of the envelope (which provides the base threshold); this is followed by a gradient-based onset position identification process.

For computationally simplifying the median-filtering operation, $\mathfrak{E}$ is downsampled by a factor of 100. For a filter with an even order $\mathfrak{H}$, the weighted median-filtered envelope can be estimated as:

$$\mathfrak{E}^{\mathsf{med}}(n) \;=\; \mathfrak{v} + \mathfrak{r}\;\mathsf{median}\left(\mathfrak{E}\Big(n-\frac{\mathfrak{H}}{2}\Big), \mathfrak{E}\Big(n-\frac{\mathfrak{H}}{2}+1\Big), \ldots, \mathfrak{E}\Big(n+\frac{\mathfrak{H}}{2}-1\Big)\right). \quad (6.5)$$

Here, the values of $\mathfrak{H} = 150$, $\mathfrak{v} = 0$ and $\mathfrak{r} = 1.2$ were selected for mixtures containing single notes (Case 3), whereas for cases of melody mixtures the value of $\mathfrak{r} = 2.0$ deemed sufficient. The algorithm then seeks the following set of samples $\mathfrak{M}$:

$$\mathfrak{M} \;=\; \{\mathfrak{m} : \mathfrak{E}(\mathfrak{m}) \geqslant \mathfrak{E}^{\mathsf{med}}(\mathfrak{m}) \;\wedge\; \mathfrak{E}(\mathfrak{m}-1) < \mathfrak{E}^{\mathsf{med}}(\mathfrak{m}-1)\} \quad (6.6)$$

Each note attack should correspond exclusively to a different value of $\mathfrak{m}$; this means that $|\mathfrak{M}|$ should be equal to the number of note events. If this is true, the actual onset location can be identified for each $\mathfrak{m} \in \mathfrak{M}$. This is performed by first finding the position $\mathfrak{n}$ in an interval $[\mathfrak{m} - \mathfrak{d}, \mathfrak{m}]$ so that

$$\mathfrak{n} \;=\; \mathsf{arg}\min_{\mathfrak{p}\in[\mathfrak{m}-\mathfrak{d},\mathfrak{m}]} \mathfrak{E}(\mathfrak{p}) \quad (6.7)$$

and then, with the help of the gradient of the envelope, the onset location corresponding to $\mathfrak{m}$ can be derived as:

$$\mathfrak{p}^{(\mathfrak{m})}_{\mathsf{ons}} \;=\; \mathsf{arg}\max_{\mathfrak{p}\in[\mathfrak{n},\mathfrak{m}]} (\mathsf{grad}\,\mathfrak{E}(\mathfrak{p}) < \mathfrak{z}), \quad (6.8)$$

where the constants $\mathfrak{z} = 0.001$ and $\mathfrak{d} = 40$ were chosen after preliminary experimentation. Furthermore, an additional minimum constant threshold $\mathfrak{a}$ is used for $\mathfrak{E}$, where $\mathfrak{a} = 0.005$. Fig. 6.13 depicts the relationship between $\mathfrak{E}$ and $\mathfrak{E}^{\mathsf{med}}$, along with the identified onset locations for the cello-violin-sax mixture of Fig. 6.9b. Finally, the set of values $\{\mathfrak{p}^{(\mathfrak{m})}_{\mathsf{ons}}\}_{\mathfrak{m}\in\mathfrak{M}}$ is multiplied with the previously used downsampling factor in order to reflect the timings in the original time-domain signal.

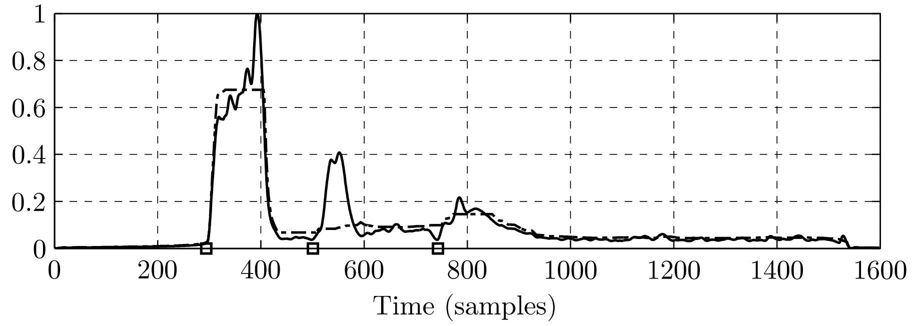


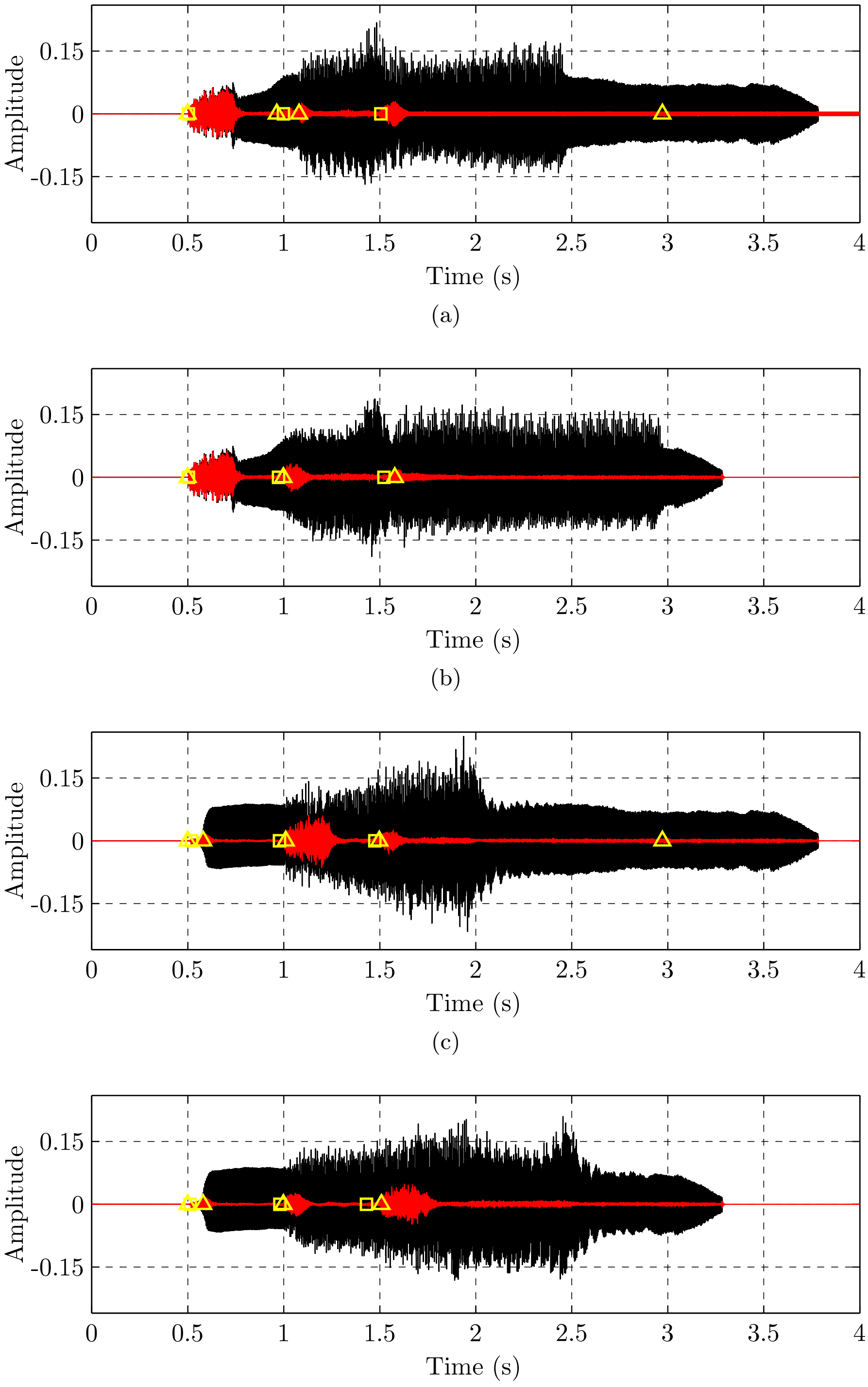
Figure 6.13: The downsampled envelope $\mathfrak{E}$ (solid line) of the cello-violin-sax mixture of Fig. 6.9b, along with its median-filtered version $\mathfrak{E}^{\mathsf{med}}$(dotted line). The estimated onset positions are indicated by the squares.

6.4.2 Initial experimental results

The results of the proposed onset detection algorithm were compared to the results obtained using the *Aubio Onset Detector* (AOD) plugin[3] within the *Sonic Visualiser* software [28]. More information about the particular onset detection method can be found in [19, 20]. The 3-note mixtures of Figs. 6.11 and 6.12 were used for the comparison, which means that the ground-truth onset timings are 0.5 s, 1.0 s and 1.5 s. The AOD was initially applied multiple times on each mixture in order for an 'optimal' value (*i.e.*, the value that would lead to the best overall results) to be set for the ODF peak picker threshold. This value was then fixed for all the mixtures for the purpose of the actual comparison. Lastly, the choice of the complex-domain ODF was made for the AOD.

Figs. 6.14 and 6.15 show the onset detection results for the two methods. It can be seen that the proposed method located all the onset locations correctly, while it did not produce any false positives. On the other hand, although the AOD had success in detecting the majority of the true onsets, it did also produce a number of false positives. It is worth noting that that the majority of the false positives occurs in monophonic audio segments. For example, regarding the mixtures of Figs. 6.15a and 6.15b, the AOD produces a false positive during an isolated violin segment (around 0.7 s). This is not the case, however, when the violin is masked by another instrument. The same applies for the false positives at around 3 s in

[3] This plugin belongs to the *aubio* library – a collection of audio annotation tools [18].



(a)

(b)

(c)

(d)

Figure 6.14: Comparison in onset detection accuracy between the proposed method (squares) and the AOD (triangles) for the mixtures of Fig. 6.11: (a) cello-sax-violin, (b) cello-violin-sax, (c) sax-cello-violin and (d) sax-violin-cello. The ground-truth onset timings are located at 0.5 s, 1.0 s and 1.5 s.

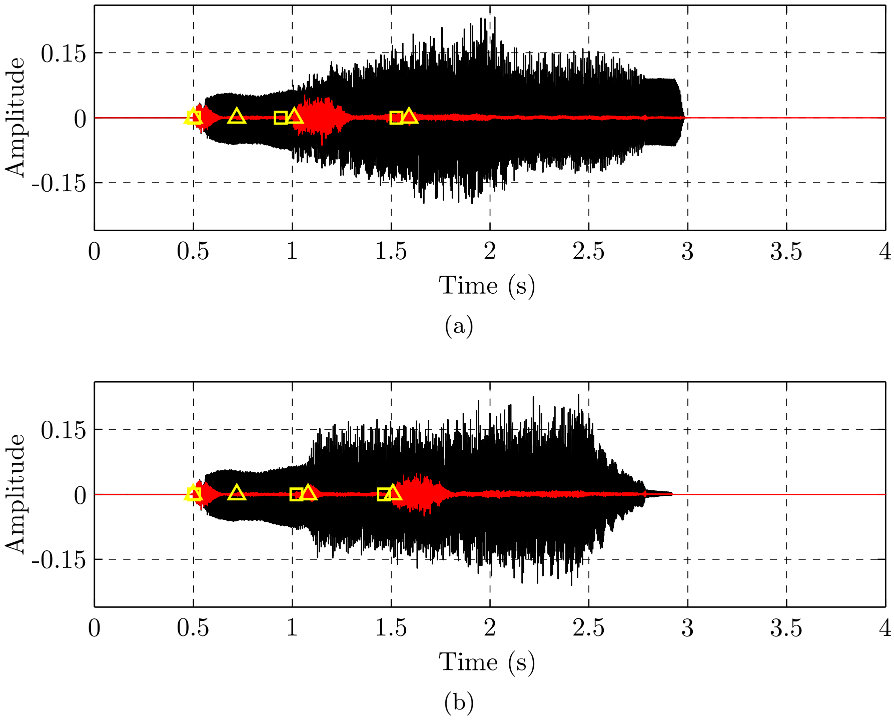


Figure 6.15: Comparison in onset detection accuracy between the proposed method (squares) and the AOD (triangles) for the mixtures of Fig. 6.12: (a) violin-cello-sax and (b) violin-sax-cello. The ground-truth onset timings are located at 0.5 s, 1.0 s and 1.5 s.

Figs. 6.14a and 6.14c. This is an indication that the presence of a note within a mixture can play a role in reducing false positives. Furthermore, it was found that increasing the peak picker threshold would eliminate the false positives, but this would be at the expense of the true positives. Overall, the proposed method was more robust compared to the AOD for this set of mixtures. It could be argued that, since this method is working with the residual, it is quite unlikely to produce any false positives for the non-attack segments of isolated sounds. The assumption here is that the harmonic part has been extracted correctly during the previous stages and the residual contains observable energy associated with all the note attacks.

6.4.3 Evaluation on melodies

In this section the proposed onset detection algorithm is evaluated with regard to its performance on much more complex and realistic music audio, compared to the 3-note examples of the previous subsection. The corpus of 6 mixtures that comprises different variations of `melody_mix1` and `melody_mix2` (Fig. 5.19) is chosen for this reason. It is expected that they will present a variety of challenges for onset detection: the different inter-source energy ratios (as defined in Eq. 5.31) can offer a variety of types of "interference" between the sources, in the sense that the same attacks will occur in different contexts, and will be masked to varying degrees.

As described above, the proposed algorithm operates on the residual channel after the harmonic or near-harmonic content has been extracted. It is important that this extraction is successful so that the residual is of a form which will allow the onset detection algorithm to operate most effectively. However, as shown previously in this thesis, cases can exist where the separation system is not able to extract all the expected content because of errors in detecting the existing source structures and inadequate corrective use of the residual loop. For this reason, and for the purpose of carrying out a more valuable analysis of the results, the method proposed here will be applied to a pre-constructed ideal form of residuals.

Every original source signal corresponding to each one of the mixtures is passed through the separation system (which in this case simply acts as a harmonic/non-harmonic energy separator), generating in this way its own individual residual. The individual residuals are then summed together to form the composite residual signals that correspond to the initial mixtures. These composite residuals can be called the *ideal residuals*, simply because they are expected to contain only non-harmonic energy corresponding to the original sources in the mix.

The evaluation of performance will be carried out using the *F-measure*, as it can provide an overall view of how well the system can do, and it has already been used at the Audio Onset Detection task of the MIREX evaluation campaign [116]:

$$\text{F-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \tag{6.9}$$

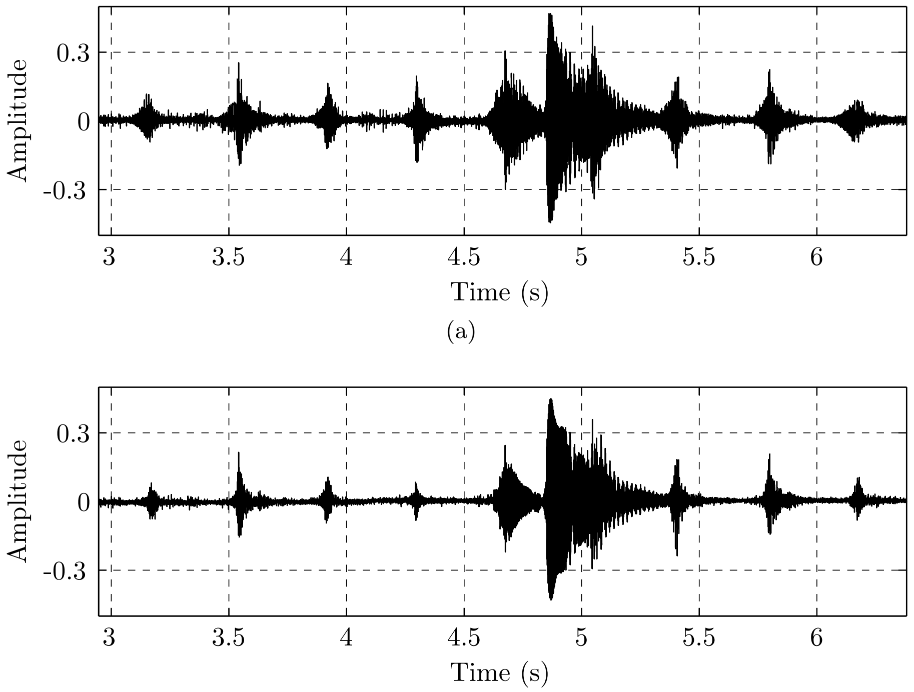
where

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6.10}$$

and “TP” are the number of true positives (correctly detected onsets); “FP” the number of the false positive onsets; and “FN” the number of false negatives (missed onsets). The algorithm's performance will also be compared with results obtained with the method described by Böck *et al.* [15], a recent onset detection system which performed best in terms of the average F-measure against other methods at the Audio Onset Detection task of the MIREX 2010 campaign. The authors were provided with the original mixture audio and sent back the onset timings obtained with their system.

Different frame lengths used for the extraction of the harmonic part of the sources, and hence the production of the residual, can potentially have an influence on the performance of the proposed onset detection algorithm. As can be seen from Fig. 6.16, the 4096-sample version generally produces sharper energy rises compared to the 8192-sample version, which is a result of the reduced amount of overlapping between frames and the increased resolution in the time domain. This can potentially lead to a higher accuracy in determining the location of the onsets, and a reduction of the number of missed onsets especially for onsets that are very close together.

The hand-corrected MIDI onset timings associated with the mixtures were used as a reference, and missed onsets were defined if no matches were found within 50 ms. Fig. 6.17 shows the comparison between the performance of the proposed algorithm and the system in [15] in terms of the F-measure. (The “precision” and “recall” measures are not shown, as they follow the same pattern.) As a first observation, `melody_mix1` displays lower performance in all cases, as might be expected, since it is a much more complicated mixture. Secondly, the system in [15] (white bars) performs better, and more consistently, in all cases. Thirdly, we can see that there is a clear performance improvement for the proposed algorithm when the shorter frames are used. This is in agreement with the expectations above regarding the relationship between smaller frames, fast energy changes at



(a)

(b)

Figure 6.16: Two versions of an excerpt from the ideal residual corresponding to `melody_mix1`, corresponding to different frame lengths used for the production of the residual: (a) 8192 samples, and (b) 4096 samples. The sources in this example are mixed with equal RMS energies ($\Delta E = 0$dB).

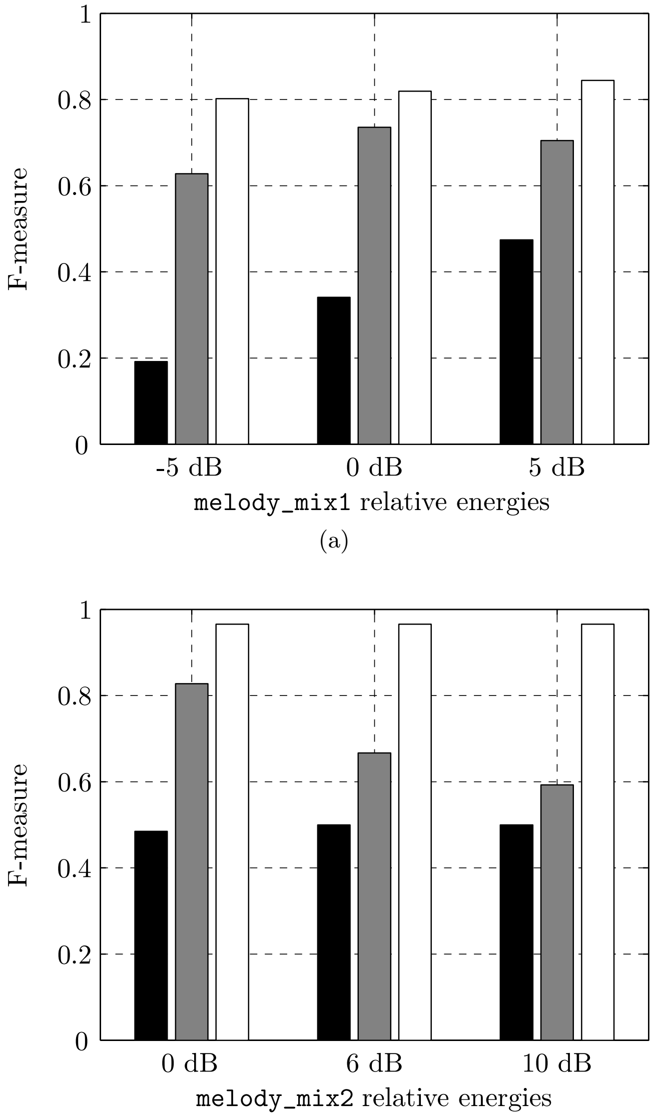


Figure 6.17: Evaluation of the proposed onset detection algorithm using the F-measure, in comparison with the system by Böck *et al.* [15] (white bars) on (a) `melody_mix1` and (b) `melody_mix2`, for different relative source energies. Two alternatives are presented for the proposed algorithm according to the frame length used for the production of the residual: 8192 samples (black bars) and 4096 samples (gray bars).

the residual and an increased detectability of onsets. Unfortunately, there is an inevitable compromise here, in that decreasing the frame length further will result in a loss of accuracy in the frequency domain, and hence degrading the quality of the overall separation, including the residual channel.

The lowering of the performance of the proposed algorithm in `melody_mix2` can be attributed mainly to the flute (s_2 in Fig. 5.19b): when ΔE is 6 dB or 10 dB, it means that the flute is 6 dB or 10 dB quieter than the alto saxophone. At the same time it was observed that the flute notes used here contained relatively weak attacks, compared to the alto and tenor saxophone notes of equal energy (the other sources in the mix). Because, also, of the large number of flute events, the inability to detect it led to a larger drop in the F-measure. The performance in `melody_mix1` appears more stable. This is partly because of the fact that the largest part of the onset data comes from the piano notes (s_2 in Fig. 5.19a), which have strong, detectable attacks. This means that the majority of the errors are caused by the other sources (with weaker attacks than the piano) which, because they contribute less to the onset data, result in a smaller performance drop than in `melody_mix2`.

The development of an onset detection algorithm was not one of the primary aims of this work but, overall, the results show that even a very simple detector applied to the residual signal can perform fairly well, even when compared to a highly developed system such as that from Böck *et al.* [15]. More importantly, because it operates on a by-product of a separation process (the residual channel), it can potentially be integrated more effectively with the separation, correction and extraction processes presented in this thesis, towards a SAU system. Further work is required to establish the extent to which improvements can be achieved if a more sophisticated onset detector can be developed which is optimised to operate on the residual signal rather than the full audio signal.

6.5 Extension to stereo mixtures

The residual as a means of supplementary information can also be exploited in combination with the other output channels within a scenario that involves a stereo, rather than just a single-channel mixture. Instead of estimating the gain and phase differences associated with the stereo set-up before the separation takes place (and, thus, often assuming strict disjointness criteria for the sources), this information can be estimated *after* source separation has taken place. By applying the current proposed system on each of the stereo channels individually, the onset locations (estimated using the algorithm of §6.4.1) can be used along with the envelope characteristics of the extracted sources to estimate the stereo-related information. This information can be used (for example) as a feature for subsequent automatic clustering of the extracted notes, or inferring the location of the sources in the acoustic space. See Ch. 7 for further comments on the potential of this approach.

6.6 Summary

This chapter concentrated on exploring the potential of the residual signal as a means of further extending and enhancing the existing one-pass automatic system presented in Ch. 5. Having in mind that a SAU system is an additional goal of this thesis, the residual offers a way to realise this. Two main ways are proposed for its use. The first way is to use it iteratively for the purpose of multiF0 error correction and improvement of source separation. Two algorithms were presented and the results showed that the use of the residual increased the range of the accepted relative energies between two synchronous notes in a mixture so that they are both detectable for differences up to a level of 40 dB.

Experiments on a corpus of realistic melody mixtures were then presented, where the performance of the one-pass approach was compared with the system extension that includes the residual loop and two recent alternative separation methods that employ similar separation philosophies. The extended residual-aided proposed system exhibited consistently better separation performance when compared to the other methods.

With regards to the ability of the iterative residual loop to further improve the estimation of the F0 tracks, evaluation results were shown in terms of the multiF0 estimation accuracy, where the extended system was compared to its one-pass version, as well as two other recent alternative multi|F0 estimation systems. Based on the already strong performance of the multiF0 estimation stage by Klapuri, the proposed way for exploiting the residual information led a to further increase in the accuracy of the F0 estimates, which in most of the cases was comparable or higher than the accuracy achieved with the alternative systems.

As a further step towards developing an understanding mechanism via a separation process (in the spirit of a SAU system), the use of the residual as a way for correctly detecting the onsets of note events was also examined. An algorithm for automatic onset detection based entirely on the residual signal was proposed, and was then compared with an existing algorithm. The proposed method outperformed this algorithm in a group of mixtures with three asynchronously-played notes.

This was followed by additional experiments carried out on melody mixtures, where the onset detection algorithm was applied to ideal forms of residuals, *i.e.*, residuals that contain only non-harmonic energy which has not been filtered out by the separation system. The effect of using shorter processing frames for the production of the residual on the accuracy of onset detection was also considered. It was found that the use of a 4096-sample processing frame leads to higher performance compared to the use of a 8192-sample one, as the energy spikes corresponding to the note attacks are better localised and less ‘blurred’. The proposed algorithm was finally compared to the results obtained by another alternative method that has been assessed in open competition as current state-of-the art. While the residual-based detection approach was not better in this case, its average performance can be characterised as fairly good, considering the simplicity of the algorithm and the sophistication of the competitor. In addition, it provided a complementary indication for the potential of the residual as a means for feature extraction. This potential can also be extended to stereo mixtures, as was briefly mentioned.

Plans for work concerning the further development of the onset detection algorithm, as well as the other parts of the system proposed in this thesis are presented in the next chapter.

CHAPTER 7

Conclusions and further work

Source separation from musical recordings is a hard, multi-disciplinary problem that has been attracting growing interest in recent years. This thesis addresses this problem through the use of an iterative framework that enables the integration of UFS and SFU paradigms towards an SAU approach, *i.e.*, an approach that can deliver both extracted musical sources and audio/music content-related information (such as F0 contours and note onset timing information), depending on the target application. It is a semi-blind unsupervised system falling under the category of CASA-inspired methods.

More specifically, the system is based on work previously reported by Every [58] which, through its particular estimation and extraction stages, produces a residual signal which has the advantage of being relatively free from extraction artefacts: no remnants of the estimated content (at least its main-lobe energy) are to be found there. By taking advantage of this, this thesis extends and formalises the idea of the residual channel as a key concept for the realisation of a SAU-type system.

An additional goal of this work was to design a system that would ideally not need significant user intervention in order for it to operate. The MIDI front-end of the previous system was, therefore, replaced by an automatic multiF0 estimator that provides the basic cue for the identification of the individual source structures.

Since the use of the residual later assumes that the estimation of those structures has been carried out with acceptable accuracy, and the reliability of the F0-track information is crucial to this, a post-processing stage for improving the robustness of the F0 estimates was introduced. Provided that the multiF0 estimator was correct $> 50\%$ of the time, it was shown that the introduction of the F0 correction stage led to an improvement of the multiF0 estimation performance in the mixes containing synchronous single-note sources.

A further modification to the system by Every was also carried out on the parameter estimation stage. Originally, the cases of overlapping harmonics were dealt with by linear amplitude interpolation in the frequency and time domains, using information from adjacent partials and time frames. It was found that the inclusion of the ability of the algorithm to extrapolate was also important. In this way, it was possible to estimate the amplitudes of lower harmonics (including the fundamental) that happened to be masked by other sources throughout the whole duration of a certain note. Although these estimates were not guaranteed to be correct, they were better than totally rejecting those harmonics.

In order to test the performance of the proposed system, a performance measure had to be chosen. Since the SRR and the SDR have been very popular as measures within the source separation community it was deemed important to compare their relative merits. The analysis that was carried out using both a theoretically-based and a practically-based framework indicated that the SDR is superior to SRR because of its insensitivity to fixed-gain distortion, while being more sensitive than SRR to the angle between estimated and extracted signal.

The one-pass version of the proposed system was compared to the MIDI-based system by Every in terms of separation and multiF0 estimation on a variety of mixture cases comprising synchronous single-note sources, as well as interweaving melodies. The findings for the case of single-note sources showed that the alternative of using an automatic front-end was not only possible but that the performance of the proposed system was better in a large number of cases. For the particular examples of interweaving melodies chosen here, the automatic version also performed well, with the F0 track disentangling process sometimes being responsible for a lowering of the performance. This was firstly due to the fact that, for these examples, the sustained note intervals that the disentangling stage operated within were much shorter compared to the ones used for the single-note source mixtures, thus leading to a less reliable F0 swapping and correction process. Secondly, the use of highly overlapping windows led to a high amount of fast non-harmonic energy content (which the system is not designed to work with) around the beginning and the end of notes, which compromised the initial F0 estimates.

Finally, as a further comment on the comparison with Every's MIDI-based version, the new automatic system is more consistent in the sense that the human element of the MIDI front-end can sometimes be unpredictable and, if the pitch refinement/correcting processes are not robust enough, will lead to a deterioration in performance.

The residual channel was explored as a source of information in two ways: providing a means of further correcting multiF0 estimates and, as a consequence, improving the separation performance through an iterative process, and enabling the estimation of note onset locations. For the multiF0 error correction, the iterative use of the residual increased the range of relative energies between the two single-note sources in a mixture whereby these sources can be detected by the multiF0 estimator (and consequently extracted from the mixture) to up to around 40 dB. Experiments were also carried out on a group of melody mixtures and the performance of the system was compared to the performance of two other alternative techniques. The separation results showed that the residual-based system provided an improvement against the one-pass version, and performed consistently better than the other methods. With regards to the multiF0 estimation results, improvement against the previous stages was shown, as well as a better average F0 estimation accuracy compared to two other methods.

The note onset detection process presented here was achieved by identifying the energy remaining in the residual signal that could be associated with the note attacks. The proposed onset detection method appeared to be more robust than the AOD when applied to six different mixtures containing three asynchronous notes starting at 0.5 s time intervals, in the sense that it detected all the true onsets and no false positives. The AOD, on the other hand, did not detect all true onsets and did also produce a number of false positives. With regards to the algorithm's performance in melody mixtures, it was observed that the ability to locate onsets could be improved consistently bys reducing the frame size – the energy bursts corresponding to the note attacks were in this way better defined. Finally, although it did not perform as well and consistently compared to a more sophisticated algorithm, considering the simplicity of the proposed onset detection method, the results were promising and indicated a potential for an intelligent use of the residual as a means of feature extraction via separation.

7.1 Further work

This section outlines the possible future directions that the present system could follow. First of all, a specific modification can be carried out at the mixture preprocessing stage, which may possibly lead to the improvement of source detection. At the moment the spectral peak picking process is independent of the multiF0 estimation stage. An alternative, which might be beneficial for the purpose of enhanced source detection, would be to make use of the knowledge of the estimated F0s for every time-frame as an additional source of information for deciding which spectral peaks to select for further processing. It can be guaranteed, in this way, that any detectable energy component located at and around the predicted locations of the harmonics will not be neglected from further processing. Mixture cases with a very high energy difference between the sources, as well as cases where desired high-frequency content is buried under the noise floor of another source could benefit from this modification. Of course, this assumes a certain degree of belief in the reliability of the F0 estimates. It needs to be tested how strong this belief should be so that it does not affect the results negatively.

The F0-track estimation and correction stage can also be further improved. In particular, for the mixture cases 2 and 4 (see the mixture classifications on p. 112), *i.e.*, mixtures containing melodies, the rule that F0 tracks do not usually cross each other in Western music [82] could be employed. This assumption has been made by other workers as well (*e.g.*, [107]). Additionally, the use of models which have been learned prior to the separation could also enhance the performance of the F0-track estimation and correction. By this we mean a learning process of statistical models for the calculated features, such as the saliences, RMS values or onset timings on sufficiently large databases of isolated and mixed audio samples. For example, Ryynänen and Klapuri [142] have followed this route in the context of AMT.

Furthermore, additional experimentation with a larger corpus of musical mixtures is needed in order to increase the robustness and extend the range of applications regarding the use of the residual channel. This will also allow the exploration of the differences between the algorithms of Figs. 6.1 and 6.2. As part of the further work that involves the residual, it would be crucial to come up with more intelligent criteria for dictating its iterative use. Extensions of the applications of the residual feedback loop as a correcting/refining device can then be examined in more complex situations, such as errors in the parameter estimation stage and inter-source leaking.

Similar considerations hold for the residual-based onset detection method. A number of the parameters have been heuristically set, and they may not be applicable to more complex mixtures, or to a different combination of types of instruments and volumes. A modification to a more data-adaptive approach can be beneficial in this case: for instance, the thresholding parameters can be a function of the frequency of high energy spikes in the envelope. Also, a second set of (slightly delayed) delayed "onsets" is available, associated with each of the harmonic note portions – combining the timings from these with those obtained separately from the residual might be another useful way ahead.

The SDR has been a useful metric for measuring and evaluating separation performance in this thesis. However, it does not offer a way to normalise its estimates according to whether certain parts of the mixture comply or not with specific assumptions of the separation system. For example, the proposed system only deals with harmonic or near-harmonic sounds. A way to realise an overall metric that offers a more valuable analysis of the separation performance would be to weigh differently, or even ignore completely, parts of the audio that do not comply with this assumption.

A variable frame size will need to be considered for the TF representation and subsequent processing of the mixture (paying attention to the degree of compromise for spectral analysis in smaller frame sizes). This can be designed within an iterative framework: a note onset detection process can be followed by an alteration of the frame size around the onset locations and then fed back for a repeated separation. This will potentially lead to a more reliable source separation and production of residual.

The system could additionally benefit from a prior learning process where the behaviour of specific features could be modelled. Any learning process, however, has to be kept sufficiently general, if the goal is to retain the unsupervised nature of the system.

Finally, the various ways to exploit the residual lead naturally to an additional extension of the proposed approach: a system that can be applied to stereo mixtures. Most of the systems performing stereo source separation carry out source estimation by taking advantage of the common inter-channel amplitude and delay differences for every source as a form of grouping cues. In order for this information to be reliable, it has to be assumed that the use of an appropriate TF representation of the mixture will lead to an adequate disjointness between the sources. In fact, a well-known method of this sort is the Degenerate Unmixing Estimation Technique (DUET) algorithm [87, 183] which is based on strict WDO. As discussed in §4.12.1, this assumption is probably not as appropriate in music as it may be in speech signals, for example. Extensions of this algorithm dealing particularly with music mixtures have kept the WDO assumption (*e.g.*, [30]) or tried to move beyond its limitations by introducing methods for resolving the overlapping content (*e.g.*, [179]). An entirely different approach would be to carry out source separation individually for each of the two channels (in other words, treating them as individual mono signals) by using the system presented here, and then deriving the inter-channel information as a post-processing step. A 'dual mono' approach can make use of the residuals and estimated source channels for retrieving enhanced amplitude and onset timing information for each of the sources. If it can be assumed that the sources are at specific static positions in space, this information can be useful for note grouping purposes (*i.e.*, grouping the notes belonging to the same source), as well as calculating the spatial locations of distinct sources (the relative timing of the onsets and, potentially, the harmonic content of each note can provide the inter-channel delays). A preliminary exploration of this possibility (using a rather simpler separation method) was carried out in [139], where it was shown that there is potential in employing the particular 'dual mono' post-separation philosophy.

Acronyms

ADSR	Attack-Decay-Sustain-Release
AHS	Average Harmonic Structure
AOD	Aubio Onset Detector
AM	Amplitude Modulation
AMT	Automatic Music Transcription
AQO	Audio Quality Oriented
ASA	Auditory Scene Analysis
ASS	Audio Source Separation
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
CQ	Constant-Q
DFT	Discrete Fourier Transform
DUET	Degenerate Unmixing Estimation Technique
DWT	Discrete Wavelet Transform
ERB	Equal Rectangular Bandwidth
F0	Fundamental frequency
FIR	Finite Impulse Response
FM	Frequency Modulation

FFT	Fast Fourier Transform
FT	Fourier Transform
GMM	Gaussian Mixture Model
HAS	Human Auditory System
HMM	Hidden Markov Model
HWPS	Harmonically Wrapped Peak Similarity
IBM	Ideal Binary Mask
ICA	Independent Component Analysis
IDFT	Inverse DFT
IFFT	Inverse FFT
ISA	Independent Subspace Analysis
JND	Just-Noticeable Difference
MIDI	Musical Instrument Digital Interface
MIR	Music Information Retrieval
MIREX	Music Information Retrieval Evaluation eXchange
NMF	Non-negative Matrix Factorization
ODF	Onset Detection Function
RMS	root-mean-square
SAR	Signal-to-Artifacts Ratio
SAU	Separation And Understanding
SDR	Signal-to-Distortion Ratio
SFU	Separation For Understanding

SIR	Signal-to-Interference Ratio
SMS	Spectral Modeling Synthesis
SNR	Signal-to-Noise Ratio
SO	Significance Oriented
SRR	Signal-to-Residual Ratio
SiSEC	Signal Separation Evaluation Campaign
STFT	Short-time Fourier Transform
SWU	Separation Without Understanding
TF	time-frequency
UFS	Understanding For Separation
UWS	Understanding Without Separation
WDO	W-Disjoint Orthogonality
WPT	Wavelet Packet Transform

Notation and conventions

$\{a, b, c\}$	Unordered set of elements a, b and c.
(a, b, c)	Ordered set of elements a, b and c.
$[a, b]$	$\{x \in \mathbb{R} : a \leqslant x \leqslant b\}$
$]a, b]$	$\{x \in \mathbb{R} : a < x \leqslant b\}$
$]a, b[$	$\{x \in \mathbb{R} : a < x < b\}$
$[a, b[$	$\{x \in \mathbb{R} : a \leqslant x < b\}$
$\mathbf{A} := (a_{jk})_{J \times K}$	Matrix $\mathbf{A}$ of size $J \times K$ with elements a_{jk}.
$\mathbf{a} := (a_j)_{j=1}^{J} \equiv (a_j)_{J \times 1}$ $= [a_1\ a_2\ \ldots\ a_J]^\mathsf{T}$	Column vector $\mathbf{a}$ of length J, with elements a_j.
$\langle \mathbf{a}, \mathbf{b} \rangle$	Inner product of vectors $\mathbf{a}$ and $\mathbf{b}$.
$\angle(\mathbf{a}, \mathbf{b})$	Angle between vectors $\mathbf{a}$ and $\mathbf{b}$.
$\lVert \mathbf{a} \rVert := \sqrt{\langle \mathbf{a}, \mathbf{a} \rangle}$	ℓ^2-norm of vector $\mathbf{a}$. If $\mathbf{a}$ is a discrete time signal, $\lVert \mathbf{a} \rVert^2$ represents its energy.
$\lVert \mathbf{A} \rVert_\mathsf{F}$	Frobenius norm of $\mathbf{A}$.
$\mathbf{a}^\mathsf{rms}$	RMS energy of $\mathbf{a}$.
$\mathbf{A} \circ \mathbf{B}$	Hadamard (element-wise) product of $\mathbf{A}$ and $\mathbf{B}$ that are of the same size.
$\mathbf{A}^\mathsf{T},\ \mathbf{a}^\mathsf{T}$	Transpose of $\mathbf{A}$ and $\mathbf{a}$.
$\mathbf{A}^\mathsf{H},\ \mathbf{a}^\mathsf{H}$	Hermitian (complex conjugate) transpose of $\mathbf{A}$ and $\mathbf{a}$.
$\mathsf{grad}(\mathbf{a})$	Gradient of $\mathbf{a}$.
$\mathsf{mean}(\mathbf{a})$	Arithmetic mean of the elements in $\mathbf{a}$.

$\delta_{ij} := \begin{cases} 1, & \text{for } i = j \\ 0, & \text{otherwise} \end{cases}$	The Kronecker delta.
$x(t)$	Continuous time signal, $-\infty < t < \infty$.
$x(n) \equiv x_n \equiv x(nT_s)$	Discrete time signal, where $n \in \mathbb{Z}$. Also, only for this kind of signal the notations $\mathbf{x}$ and x are equivalent (where $\mathbf{x}$ is a column vector with elements $\{x_n : n \in \mathbb{Z}\}$).
$\mathcal{X}$, $\boldsymbol{\mathcal{X}}$	TF or frequency-domain representation of a time-domain signal x, and its matrix notation.
$\bar{z}$	Complex conjugate of $z \in \mathbb{C}$.
$\lfloor a \rfloor$	The nearest integer that is smaller than $a \in \mathbb{R}$ (a flooring operation on a).
$[a]$	Rounding to the nearest integer of $a \in \mathbb{R}$.
$\lvert a \rvert$	Absolute value of $a \in \mathbb{C}$, or its cardinality if a is a set.

A note on musical note naming

The method of *scientific pitch notation* [184] is used for the naming of Western musical notes. According to this, the note A4 corresponds to 440 Hz.

Bibliography

[1] SiSEC 2008. http://sisec2008.wiki.irisa.fr/tiki-index.php, Dec. 2009.

[2] MIREX Home. http://www.music-ir.org/mirex/wiki/MIREX_HOME, Dec. 2010.

[3] K. Achan, S. T. Roweis, and B. J. Frey. Probabilistic inference of speech signals from phaseless spectrograms. *Neural Information Processing Systems*, 16:1393–1400, 2003.

[4] E. Alpaydin. *Introduction to Machine Learning.* MIT Press, Cambridge, MA, 2004.

[5] J.-J. Aucouturier and F. Pachet. Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

[6] J. G. Beerends. Audio quality determination based on perceptual measurement techniques. In *Applications of Digital Signal Processing to Audio and Acoustics* [88], Ch. 1, 1998.

[7] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier. Perceptual Evaluation of Speech Quality (PESQ): The new ITU standard for end-to-end speech quality assessment. Part II – Psychoacoustic model. *J. Audio Eng. Soc.*, 50(10):765–778, 2002.

[8] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler. A tutorial on onset detection in music signals. *IEEE Trans. Speech and Audio Processing*, 13(5):1035–1047, 2005.

[9] A. Ben-Shalom, S. Shalev-Shwartz, M. Werman, and S. Dubnov. Optimal filtering of an instrument sound in a mixed recording using harmonic model and score alignment. In *Proc. Int. Computer Music Conf. (ICMC'04)*, pages 715–718, Miami, FL, USA, 2004.

[10] L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *IEEE Trans. Audio, Speech, and Language Processing*, 14(1):191–199, 2006.

[11] L. Benaroya, R. Gribonval, and F. Bimbot. Non negative sparse representation for Wiener based source separation with a single sensor. In *Proc. 2003 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'03)*, pages 613–616, Hong Kong, 2003.

[12] E. Benetos and Y. Stylianou. Auditory spectrum-based pitched instrument onset detection. *IEEE Trans. Audio, Speech, and Language Processing*, 18(8):1968–1977, 2010.

[13] R. Blouet, G. Rapaport, I. Cohen, and C. Févotte. Evaluation of several strategies for single sensor speech/music separation. In *Proc. 2008 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP '08)*, pages 37–40, Las Vegas, NV, USA, 2008.

[14] T. Blumensath and M. Davies. Unsupervised learning of sparse and shift-invariant decompositions of polyphonic music. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, volume 4, Montreal, Canada, 2004.

[15] S. Böck, F. Eyben, and B. Schuller. MIREX 2010 submission: Onset detection with bidirectional long short-term memory neural networks. In *The 6th Music Information Retrieval Evaluation eXchange (MIREX '10)*, 2010.

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

[17] E. O. Brigham. *The Fast Fourier Transform and its Applications.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[18] P. Brossier. aubio, a library for audio labelling. `http://aubio.org/`, Dec. 2009.

[19] P. Brossier, J. Bello, and M. Plumbley. Real-time temporal segmentation of note objects in music signals. In *Proc. 2004 Int. Computer Music Conf. (ICMC 2004)*, Miami, Florida, USA, Nov. 1-6 2004.

[20] P. M. Brossier. *Automatic Annotation of Musical Audio for Interactive Applications*. PhD thesis, Queen Mary, London, U.K., Aug. 2006.

[21] G. J. Brown. *Computational Auditory Scene Analysis: A Representational Approach*. PhD thesis, University of Sheffield, Sheffield, UK, 1992.

[22] J. C. Brown. Calculation of a constant Q spectral transform. *J. Acoust. Soc. Am.*, 89(1):425–434, 1991.

[23] J. J. Burred. *From Sparse Models to Timbre Learning: New Methods for Musical Source Separation*. PhD thesis, Technical University of Berlin, Berlin, Germany, 2009.

[24] J. J. Burred and T. Sikora. Comparison of frequency-warped representations for source separation of stereo mixtures. In *121st AES Convention*, San Fransisco, USA, Oct. 2006.

[25] J. J. Burred and T. Sikora. Monaural source separation from musical mixtures based on time-frequency timbre models. In *International Conference on Music Information Retrieval (ISMIR 2007)*, Vienna, Austria, Sep. 2007.

[26] J.-F. Cardoso. Blind source separation: statistical principles. *Proc. IEEE*, 9(10):2009–2025, 1998.

[27] M. A. Casey and W. Westner. Separation of mixed audio sources by independent subspace analysis. In *Proc. Int. Computer Music Conf. (ICMC'00)*, pages 154–161, Berlin, Germany, Aug. 2000.

[28] Centre for Digital Music, Queen Mary, University of London. Sonic Visualiser. `http://www.sonicvisualiser.org/`, Dec. 2009.

[29] E. C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.*, 25(5):975–979, 1953.

[30] M. Cobos and J. J. López. Stereo audio source separation based on time–frequency masking and multilevel thresholding. *Digital Signal Processing*, 18(6):960–976, 2008.

[31] M. Cobos, J. J. Lopez, A. Gonzalez, and J. Escolano. Stereo to wave-field synthesis music up-mixing: An objective and subjective evaluation. In *3rd Int. Symp. on Communications, Control and Signal Processing (ISCCSP 2008)*, pages 1279–1284, March 2008.

[32] N. Collins. A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions. In *Proc. 118th Audio Engineering Society Convention*, Barcelona, Spain, May 2005.

[33] P. R. Cook, editor. *Music, Cognition, and Computerized Sound.* MIT Press, Cambridge, Massachusetts, 1999.

[34] M. Cooke. *Modelling Auditory Processing and Organisation.* PhD thesis, University of Sheffield, Sheffield, UK, 1991.

[35] M. Cooke, S. Beet, and M. Crawford, editors. *Visual Representations of Speech Signals.* John Wiley & Sons, Inc., New York, NY, USA, 1993.

[36] G. Cornuz, E. Ravelli, P. Leveau, and L. Daudet. Object coding of harmonic sounds using sparse and structured representations. In *Proc. 10th Int. Conf. on Digital Audio Effects (DAFx-07)*, Bordeaux, France, Sept. 10-15 2007.

[37] R. E. Crochiere. A weighted overlap-add method of short-time Fourier analysis/synthesis. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-28(1):99–102, 1980.

[38] L. Daudet. A review on techniques for the extraction of transients in musical signals. *Lecture Notes in Computer Science*, 3902:219, 2006.

[39] M. Davy and S. J. Godsill. Bayesian harmonic models for musical signal analysis. *Bayesian Statistics*, 7, 2003.

[40] A. de Cheveigné. Multiple F0 estimation. In *Computational Auditory Scene Analysis* [176], Ch. 2, 2006.

[41] A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111(4):1917–1930, 2002.

[42] N. Degara, A. Pena, M. E. P. Davies, and M. D. Plumbley. Note onset detection using rhythmic structure. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP' 07)*, pages 5526–5529, 2010.

[43] P. N. Denbigh and J. Zhao. Pitch extraction and separation of overlapping speech. *Speech Communication*, 11:119–125, 1992.

[44] M. Desainte-Catherine and S. Marchand. High precision Fourier analysis of sounds using signal derivatives. *J. Audio Eng. Soc.*, 48(7/8):654–667, 2000.

[45] D. Deutsch. Grouping mechanisms in music. In *The Psychology of Music* [46], Ch. 9, 1998.

[46] D. Deutsch, editor. *The Psychology of Music.* Academic Press, 2nd edition, 1998.

[47] J. S. Downie. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[48] K. Dressler. An auditory streaming approach on melody extraction. MIREX Audio Melody Extraction Contest Abstracts, 2006.

[49] Z. Duan, J. Han, and B. Pardo. Harmonically informed multi-pitch tracking. In *The 6th Music Information Retrieval Evaluation eXchange (MIREX '10)*, 2010.

[50] Z. Duan and B. Pardo. Soundprism: An online system for score-informed source separation of music audio. *IEEE Selected Topics in Signal Processing* (submitted), 2011.

[51] Z. Duan, B. Pardo, and C. Zhang. Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Trans. Audio, Speech, and Language Processing*, 18(8):2121–2133, 2010.

[52] Z. Duan, Y. Zhang, C. Zhang, and Z. Shi. Unsupervised single-channel music source separation by average harmonic structure modeling. *IEEE Trans. Audio, Speech, and Language Processing*, 16(4):766–778, 2008.

[53] J. L. Durrieu, G. Richard, and B. David. Singer melody extraction in polyphonic signals using source separation methods. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'08)*, pages 169–172, 2008.

[54] J. L. Durrieu, G. Richard, and B. David. An iterative approach to monaural musical mixture de-soloing. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'09)*, pages 105–108, Washington DC, USA, 2009.

[55] D. P. Ellis. *Prediction-driven Computational Auditory Scene Analysis.* PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, June 1996.

[56] D. P. W. Ellis and D. Rosenthal. Mid-level representations for computational auditory scene analysis. In *Proc. of the Computational Auditory Scene Analysis Workshop; Int. Joint. Conf. Artificial Intelligence*, Montreal, Quebec, Aug. 1995.

[57] P. A. A. Esquef, V. Valimaki, and M. Karjalainen. Restoration and enhancement of solo guitar recordings based on sound source modeling. *J. Audio Eng. Soc.*, 50(4):227–236, 2002.

[58] M. R. Every. *Separation of Musical Sources and Structure from Single-Channel Polyphonic Recordings.* PhD thesis, Department of Electronics, University of York, U.K., 2006.

[59] M. R. Every and J. E. Szymanski. Separation of overlapping impulsive sounds by bandwise noise interpolation. In *Proc. 8th Int. Conf. on Digital Audio Effects (DAFx'05)*, Madrid, Spain, Sep. 20-22 2005.

[60] M. R. Every and J. E. Szymanski. Separation of synchronous pitched notes by spectral filtering of harmonics. *IEEE Trans. Audio, Speech, and Language Processing*, 14(5):1845–1856, Sept. 2006.

[61] F. Eyben, S. Böck, B. Schuller, and A. Graves. Universal onset detection with bidirectional long short-term memory neural networks. In *Proc. 11th Int. Conf. on Music Information Retrieval (ISMIR'10)*, pages 589–594, Utrecht, Netherlands, 2010.

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

[63] N. H. Fletcher and T. D. Rossing. *The Physics of Musical Instruments.* Springer-Verlag, New York, 2nd edition, 1998.

[64] B. Fox, A. Sabin, B. Pardo, and A. Zopf. Modeling perceptual similarity of audio signals for blind source separation evaluation. In *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation*, London, UK, Sept. 9-12 2007.

[65] S. J. Godsill, P. Rayner, and O. Cappé. Digital audio restoration. In *Applications of Digital Signal Processing to Audio and Acoustics* [88], Ch. 4, 1998.

[66] M. Goodwin. Residual modeling in music analysis-synthesis. In *Proc. 1996 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'96)*, volume 2, pages 1005–1008, Atlanta, GA., May 7-10 1996.

[67] M. M. Goodwin. *Adaptive Signal Models: Theory, Algorithms, and Audio Applications.* Kluwer Academic Publishers, 1998.

[68] M. Goto. Music scene description. In *Signal Processing Methods for Music Transcription* [96], Ch. 11, 2006.

[69] D. W. Griffin and J. S. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 32(2):236–242, 1984.

[70] F. J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proc. of the IEEE*, 66(1):51–83, 1978.

[71] W. M. Hartmann. Pitch, periodicity, and auditory organization. *J. Acoust. Soc. Am.*, 100(6):3491–3502, 1996.

[72] S. Haykin and Z. Chen. The cocktail party problem. *Neural Computation*, 17(9):1875–1902, 2005.

[73] M. Helén and T. Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *13th European Signal Processing Conference*, Antalaya, 2005.

[74] L. M. F. Helmholtz. *On the sensations of tone as a physiological basis for the theory of music* (A. J. Ellis Trans.). Dover, New York, 2nd English edition, 1954 (Original work published 1877).

[75] P. Herrera-Boyer, A. Klapuri, and M. Davy. Automatic classification of pitched musical instrument sounds. In *Signal Processing Methods for Music Transcription* [96], Ch. 6, 2006.

[76] P. Herrera-Boyer, G. Peeters, and S. Dubnov. Automatic classification of musical instrument sounds. *J. New Music Research*, 32(1):3–21, 2003.

[77] J. R. Hopgood and P. J. W. Rayner. Single channel nonstationary stochastic signal separation using linear time-varying filters. *IEEE Trans. Signal Processing*, 51(7):1739–1752, 2003.

[78] C. L. Hsu, J. S. R. Jang, and T. L. Tsai. Separation of singing voice from music accompaniment with unvoiced sounds reconstruction for monaural recordings. In *125th AES Convention*, San Fransisco, USA, Oct. 2-5 2008.

[79] G. Hu and D. L. Wang. Speech segregation based on pitch tracking and amplitude modulation. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA'01)*, pages 79–82, 2001.

[80] G. Hu and D. L. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Trans. Neural Networks*, 15(5):1135–1150, 2004.

[81] D. Huron. Voice denumerability in polyphonic music of homogeneous timbres. *Music Perception*, 6(4):361–382, 1989.

[82] D. Huron. The avoidance of part-crossing in polyphonic music: perceptual evidence and musical practice. *Music Perception*, 9(1):93–104, 1991.

[83] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13:411–430, 2000.

[84] ISMIR. The International Society for Music Information Retrieval – Conferences, Publications and Related Activities. `http://www.ismir.net/`, March 2009.

[85] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. Integration and adaptation of harmonic and inharmonic models for separating polyphonic musical signals. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, Honolulu, HI, USA, 2007.

[86] G. J. Jang and T. W. Lee. A maximum likelihood approach to single-channel source separation. *Journal of Machine Learning Research*, 4:1365–1392, 2003.

[87] A. Jourjine, S. Rickard, and O. Yılmaz. Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'00)*, volume 5, pages 2985–2988, Istanbul, Turkey, June 5-9 2000.

[88] M. Kahrs and K. Brandenburg, editors. *Applications of Digital Signal Processing to Audio and Acoustics.* Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[89] K. Kashino. Auditory scene analysis in music signals. In *Signal Processing Methods for Music Transcription* [96], Chapter 10.

[90] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka. Application of Bayesian probability network to music scene analysis. In *Working Notes of Int. Joint Conferences on Artificial Intelligence, Workshop of Computational Auditory Scene Analysis (IJCAI-CASA)*, pages 52–59, Aug. 1995.

[91] M. Kim and S. Choi. Monaural music source separation: Nonnegativity, sparseness, and shift-invariance. *Lecture Notes in Computer Science*, 3889:617, 2006.

[92] T. Kinoshita, S. Sakai, and H. Tanaka. Musical sound source identification based on frequency component adaptation. In *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI) Workshop on Computational Auditory Scene Analysis*, pages 18–24, Stockholm, August 1999.

[93] P. Kisilev and Y. Zeevi. A multiscale framework for blind separation of linearly mixed signals. *J. Machine Learning Research*, 4(2003):1339–1364, 2003.

[94] A. Klapuri. Introduction to music transcription. In *Signal Processing Methods for Music Transcription* [96], Ch. 1.

[95] A. Klapuri. Multipitch analysis of polyphonic music and speech signals using an auditory model. *IEEE Trans. Audio, Speech, and Language Processing*, 16(2):255 – 266, Feb. 2008.

[96] A. Klapuri and M. Davy, editors. *Signal Processing Methods for Music Transcription.* Springer, New York, 2006.

[97] A. P. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Trans. Speech and Audio Processing*, 11(6):804–816, 2003.

[98] A. P. Klapuri. Multiple fundamental frequency estimation by summing harmonic amplitudes. In *7th Int. Conf. on Music Information Retrieval (ISMIR-06)*, Victoria, Canada, Oct. 2006.

[99] B. Kollmeier and R. Koch. Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. *J. Acoust. Soc. Am.*, 95(3):1593–1602, 1994.

[100] M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis. Normalized cuts for predominant melodic source separation. *IEEE Trans. Audio, Speech, and Language Processing*, 16(2):278–290, Feb. 2008.

[101] M. Lagrange, L. G. Martins, and G. Tzanetakis. Semi-automatic mono to stereo up-mixing using sound source formation. In *122nd AES Convention*, Vienna, Austria, May 5-8 2007.

[102] M. Lagrange, L. G. Martins, and G. Tzanetakis. A computationally efficient scheme for dominant harmonic source separation. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'08)*, pages 165–168, 2008.

[103] B. P. Lathi. *Signal Processing and Linear Systems*. Oxford University Press, USA, 2000.

[104] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[105] Y. Li and D. L. Wang. Separation of singing voice from music accompaniment for monaural recordings. *IEEE Trans. Audio, Speech, and Language Processing*, 15(4):1475, 2007.

[106] Y. Li and D. L. Wang. Musical sound separation based on binary time-frequency masking. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009, 2009.

[107] Y. Li, J. Woodruff, and D. L. Wang. Monaural musical sound separation based on pitch and common amplitude modulation. *IEEE Trans. Audio, Speech, and Language Processing*, 17(7):1361–1371, 2009.

[108] R. F. Lyon. A computational model of filtering, detection, and compression in the cochlea. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'82)*, pages 1282–1285, Paris, May 1982.

[109] R. C. Maher. Evaluation of a method for separating digitized duet signals. *J. Audio Eng. Soc.*, 38(12):956–979, 1990.

[110] S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, 2nd edition, 1999.

[111] D. Marr. *Vision.* Freeman, New York, 1982.

[112] L. G. Martins. *A Computational Framework for Sound Segregation in Music Signals.* PhD thesis, School of Engineering of the University of Porto (FEUP), Portugal, Feb. 2008.

[113] P. Masri, A. Bateman, and N. Canagarajah. The importance of the time-frequency representation for sound/music analysis-resynthesis. *Organised Sound*, 2(3):207–214, 1997.

[114] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 34(4):744–754, 1986.

[115] D. K. Mellinger. *Event Formation and Separation in Musical Sound.* PhD thesis, Stanford University, CA, USA, 1991.

[116] MIREX. 2010:Audio Onset Detection. `http://www.music-ir.org/mirex/wiki/2010:Audio_Onset_Detection`, June 2010.

[117] M. K. I. Molla and K. Hirose. Single-mixture audio source separation by subspace decomposition of hilbert spectrum. *IEEE Trans. Audio, Speech, and Language Processing*, 15(3):893–900, March 2007.

[118] B. C. J. Moore. *An introduction to the psychology of hearing.* Academic Press, 5th edition, 2003.

[119] B. C. J. Moore and B. R. Glasberg. A revision of Zwicker's loudness model. *Acta Acustica united with Acustica*, 82(2):335–345, 1996.

[120] T. Nakatani, M. Goto, and H. G. Okuno. Localization by harmonic structure and its application to harmonic sound stream segregation. In *Proc. 1996*

IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'96), volume 2, pages 653–656, 1996.

[121] J. J. Nattiez. *Music and Discourse* (C. Abbate Trans.). Princeton University Press, Princeton, New Jersey, 1990 (Original work published 1987).

[122] S. H. Nawab, T. F. Quatieri, and J. S. Lim. Signal reconstruction from short-time Fourier transform magnitude. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 31(4):986–998, 1983.

[123] A. Nuttal. Some windows with very good sidelobe behavior. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 29(1):84–91, February 1981.

[124] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval. Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs. *IEEE Trans. Audio, Speech, and Language Processing*, 15(5):1564–1578, 2007.

[125] T. W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *J. Acoust. Soc. Am.*, 60(4):911–918, 1976.

[126] M. S. Pedersen. *Source Separation for Hearing Aid Applications.* PhD thesis, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Kgs. Lyngby, Denmark, 2006.

[127] G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Technical report, CUIDADO I.S.T. Project Report, 2004.

[128] J. R. Pierce. Introduction to pitch perception. In *Music, Cognition, and Computerized Sound* [33], Chapter 5.

[129] J. R. Pierce. The nature of musical sound. In *The Psychology of Music* [46], chapter 1.

[130] R. Plomp. *Aspects of Tone Sensation: A Psychophysical Study.* Academic Press, London, 1976.

[131] M. D. Plumbley, S. A. Abdallah, T. Blumensath, M. G. Jafari, A. Nesbit, E. Vincent, and B. Wang. Musical audio analysis using sparse representations. In *COMPSTAT 2006 Proc. in Computational Statistics*, pages 104–117, Rome, Italy, 2006. Physical-Verlag.

[132] M. R. Portnoff. Time-frequency representation of digital signals and systems based on short-time Fourier analysis. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 28(1):55–69, 1980.

[133] J. G. Proakis and D. G. Manolakis. *Digital Signal Processing: Principles, Algorithms, and Applications*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 3rd edition, 1996.

[134] S. Qian. *Introduction to Time-Frequency and Wavelet Transforms*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2002.

[135] D. M. Randel, editor. *The Harvard Dictionary of Music*. Belknap Press of Harvard University Press, Cambridge, MA, 4th edition, 2003.

[136] C. Raphael. A classifier-based approach to score-guided source separation of musical audio. *Computer Music Journal*, 32(1):51–59, 2008.

[137] J. C. Risset and D. L. Wessel. Exploration of timbre by analysis and synthesis. In *The Psychology of Music* [46], Chapter 5.

[138] X. Rodet. Musical sound signal analysis/synthesis: Sinusoidal+residual and elementary waveform models. In *Proc. IEEE Time-Frequency and Time-Scale Workshop (TFTS'97)*, Coventry, UK, Aug. 27-29 1997.

[139] M. Rodríguez. Stereo musical source separation. Master's thesis, Department of Electronics, University of York, Aug. 2009.

[140] T. D. Rossing. *The Science of Sound*. Addison-Wesley, 1990.

[141] S. T. Roweis. One microphone source separation. *Advances in Neural Information Processing Systems (NIPS 13)*, pages 793–799, 2001.

[142] M. Ryynänen and A. P. Klapuri. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32(3):72–86, 2008.

[143] M. Ryynänen, T. Virtanen, J. Paulus, and A. Klapuri. Accompaniment separation and karaoke application based on automatic melody transcription. In *IEEE Int. Conf. on Multimedia and Expo*, pages 1417–1420, Hannover, Germany, 2008.

[144] P. Schaeffer. *La musique concréte.* Presses Universitaires de France, Paris, France, 1967.

[145] E. D. Scheirer. *Music-Listening Systems.* PhD thesis, MIT Media Laboratory, Apr. 2000.

[146] J. F. Schouten. The perception of subjective tones. *Proc. Kon. Acad. Wetensch.(Neth.)*, 41:1086–1094, 1938.

[147] X. Serra. *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition.* PhD thesis, Stanford University, 1989.

[148] X. Serra and J. Smith III. Spectral Modelling Synthesis: A Sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24, 1990.

[149] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[150] G. Siamantas. Giorgos Siamantas - Thesis audio results. `http://www-users.york.ac.uk/~jes1/SiamantasPhD/Thesis_examples/index.html`, April 2011.

[151] M. Slaney and R. F. Lyon. On the importance of time – a temporal representation of sound. In *Visual Representations of Speech Signals* [35], pp.95–116.

[152] M. Slaney and R. F. Lyon. A perceptual pitch detector. In *Proc. 1990 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'90)*, pages 357–360, Albuquerque, NM, Apr. 1990.

[153] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2003*, pages 177–180, 2003.

[154] S. Srinivasan, J. Samuelsson, and W. B. Kleijn. Codebook-based Bayesian speech enhancement. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'05)*, pages 1077–1080, Philadelphia, PA, USA, Mar. 2005.

[155] S. Srinivasan, J. Samuelsson, and W. B. Kleijn. Codebook driven short-term predictor parameter estimation for speech enhancement. *IEEE Trans. Audio, Speech, and Language Processing*, 14(1):163–176, 2006.

[156] S. H. Srinivasan and M. Kankanhalli. Harmonicity and dynamics based audio separation. In *Proc. 2003 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP '03)*, pages 640–643, Hong Kong, 2003.

[157] E. Terhardt. Pitch, consonance, and harmony. *J. Acoust. Soc. Am.*, 55:1061–1069, 1974.

[158] B. Thoshkahna and K. R. Ramakrishnan. A psychoacoustics based sound onset detection algorithm for polyphonic audio. In *Proc. 9th Int. Conf. on Signal Processing (ICSP 2008)*, Beijing, China, 2008.

[159] University of Iowa Electronic Music Studios. Musical instrument samples. `http://theremin.music.uiowa.edu/MIS.html`, Dec. 2009.

[160] B. L. Vercoe, W. G. Gardner, and E. D. Scheirer. Structured audio: creation, transmission, and rendering of parametric sound representations. *Proc. IEEE*, 86(5):922–940, 1998.

[161] E. Vincent, S. Araki, and P. Bofill. The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation. In

Proc. of the 8th Int. Conf. on Independent Component Analysis and Source Separation (ICA '09), pages 734–741, 2009.

[162] E. Vincent, C. Févotte, R. Gribonval, L. Benaroya, A. Röbel, X. Rodet, F. Bimbot, and É. le Carpentier. A tentative typology of audio source separation tasks. In *Proc. Int. Symp. ICA and BSS (ICA 2003)*, pages 715–720, Nara, Japan, 2003.

[163] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Trans. on Audio, Speech and Language Processing*, 14(4):1462–1469, 2006.

[164] E. Vincent, R. Gribonval, and M. D. Plumbley. Oracle estimators for the benchmarking of source separation algorithms. *Signal Processing*, 87(8):1933–1950, 2007.

[165] E. Vincent and M. D. Plumbley. Single-channel mixture decomposition using bayesian harmonic models. In *Proc. of the 6th Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA 2006)*, number LNCS 3889, pages 722–730, Charleston, SC, USA, 5-8 March 2006. Springer-Verlag, Berlin.

[166] E. Vincent and M. D. Plumbley. Low bit-rate object coding of musical audio using bayesian harmonic models. *IEEE Trans. Audio, Speech, and Language Processing*, 15(4):1273–1282, 2007.

[167] T. Virtanen. Algorithm for the separation of harmonic sounds with time-frequency smoothness constraint. In *Proc. Int. Conf. on Digital Audio Effects (DAFx).(2003)*, pages 35–40, London, UK, 2003.

[168] T. Virtanen. Separation of sound sources by convolutive sparse coding. In *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA 2004)*, Jeju, Korea, October 2004.

[169] T. Virtanen. *Sound Source Separation in Monaural Music Signals*. PhD thesis, Tampere University of Technology, Tampere, Finland, 2006.

[170] T. Virtanen. Unsupervised learning methods for source separation. In *Signal Processing Methods for Music Transcription* [96], Chapter 9, 2006.

[171] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Trans. Audio, Speech, and Language Processing*, 15(3):1066–1074, 2007.

[172] T. Virtanen and A. Klapuri. Separation of harmonic sound sources using sinusoidal modeling. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'00)*, pages 765–769, Istanbul, Turkey, June 2000.

[173] T. Virtanen and A. Klapuri. Separation of harmonic sounds using multipitch analysis and iterative parameter estimation. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'01)*, New Paltz, NY., 2001.

[174] T. Virtanen, A. Mesaros, and M. Ryynänen. Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music. In *Workshop on Statistical and Perceptual Audition (SAPA2008)*, Brisbane, Australia, 21 Sept. 2008.

[175] B. Wang and M. D. Plumbley. Musical audio stream separation by non-negative matrix factorization. In *Proc. Digital Music Research Network Summer Conf. 2005*, Glasgow, UK, 23-24 July 2005.

[176] D. L. Wang and G. J. Brown, editors. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. IEEE Press/Wiley-Interscience, USA, 2006.

[177] D. L. Wang and G. J. Brown. Fundamentals of computational auditory scene analysis. In *Computational Auditory Scene Analysis* [176], Ch. 1, 2006.

[178] N. Wiener. *Extrapolation, interpolation, and smoothing of stationary time series*. The MIT Press, Cambridge, MA, USA, 1964.

[179] J. Woodruff and B. Pardo. Using pitch, amplitude modulation, and spatial cues for separation of harmonic instruments from stereo music recordings. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007.

[180] J. Woodruff, B. Pardo, and R. Dannenberg. Remixing stereo music with score-informed source separation. In *Proc. ISMIR*, pages 314–319, Victoria, Canada, 8-12 Oct. 2006.

[181] C. Yeh and A. Roebel. Multiple-F0 estimation for MIREX 2010. In *The 6th Music Information Retrieval Evaluation eXchange (MIREX '10)*, 2010.

[182] C. Yeh, A. Roebel, and X. Rodet. Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals. *IEEE Trans. Audio, Speech, and Language Processing*, 18(6):1116–1126, 2010.

[183] O. Yılmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. Signal Processing*, 52(7):1830–1847, 2004.

[184] R. W. Young. Terminology for logarithmic frequency units. *J. Acoust. Soc. Am.*, 11(1):134–139, 1939.

[185] Y. Zhang, C. Zhang, and S. Wang. Clustering in knowledge embedded space. In *Proc. ECML*, pages 480–491, 2003.

[186] R. Zhou, M. Mattavelli, and G. Zoia. Music onset detection based on resonator time frequency image. *IEEE Trans. Audio, Speech, and Language Processing*, 16(8):1685–1695, 2008.

[187] E. Zwicker. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *J. Acoust. Soc. Am.*, 33(2):248, 1961.