ISCTE IUL
Instituto Universitário de Lisboa

Department of Information Science and Technology

Onset Detection in Music Signals

Carlos Manuel Tadeia Rosão

A Dissertation presented in partial fulfilment of the Requirements for the Degree of Master in Open Source Software (Software de Código Aberto)

Supervisor:
Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Assistant Professor,
ISCTE-Instituto Universitário de Lisboa

Co-supervisor:
Doctor David Manuel Martins de Matos, Assistant Professor,
Instituto Superior Técnico, Universidade Técnica de Lisboa

April 2012

ISCTE IUL
Instituto Universitário de Lisboa

Department of Information Science and Technology

Onset Detection in Music Signals

Carlos Manuel Tadeia Rosão

A Dissertation presented in partial fulfilment of the Requirements for the Degree of Master in Open Source Software (Software de Código Aberto)

Supervisor:
Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Assistant Professor,
ISCTE-Instituto Universitário de Lisboa

Co-supervisor:
Doctor David Manuel Martins de Matos, Assistant Professor,
Instituto Superior Técnico, Universidade Técnica de Lisboa

April 2012

Lombada


ISCTE IUL
Instituto Universitário de Lisboa
Onset Detection in Music Signals, Carlos Rosão
April, 2012

Resumo

A Detecção de *Onsets*, ou seja, a tarefa que procura encontrar o momento de início de notas musicais num sinal de áudio, tem sido uma área de investigação activa, uma vez que a Detecção de *Onsets* é comummente utilizada como primeiro passo em tarefas de alto-nível de processamento musical.

Tendo em conta a necessidade de saber que método de Detecção de *Onsets* é mais adequado a cada tarefa de alto-nível, nesta tese foram seguidas duas abordagens que visam, acima de tudo, obter uma informação mais completa sobre cada método de Detecção de *Onsets*.

A primeira abordagem consiste numa comparação em profundidade do comportamento dos métodos de Detecção de *Onsets* que usam características espectrais do sinal. Os resultados obtidos mostram que o comportamento dos diferentes métodos varia significativamente entre as funções de detecção usadas, entre os tipos de *Onset*, e ainda de acordo com a técnica de interpretação do instrumento.

Na segunda abordagem avalia-se a influência do passo final de Selecção de Picos nos resultados globais de Detecção de *Onsets*. Os resultados obtidos mostram que o passo de Selecção de Picos influencia profundamente os resultados – negativa e positivamente –, e que esta influência difere significativamente de acordo com o tipo de *Onset* e com o método de Detecção de *Onsets* usado.

Abstract

Onset Detection, that is, the quest for finding the starting moment of musical notes in an audio signal, is an active research subject since note onset detection is commonly used as a first step in high-level music processing tasks.

Driven by the need to know which Onset Detection method can suit better each high-level music processing task, two approaches are followed in this thesis in order to obtain a more complete information about the different onset detection methods.

The first consists in a full comparison of the performance of Onset Detection Methods that use Spectral Features. Our results in two distinct datasets show that the behaviour of onset detection varies clearly between onset types and between detection functions, as well as between instrument interpretation style.

The other approach assesses the influence of the final Peak Selection step in the global results of Onset Detection. Our results show that the Peak Selection step used deeply influences both positively and negatively the results obtained, and that its influence differs significantly according to the onset classes and to the onset detection functions.

Palavras Chave
Keywords

Palavras Chave

Detecção de *Onsets*
Segmentação de Notas Musicais
Captação de Informação Musical
Transcrição Musical Automática
Detecção de Novidade
Processamento de Sinal

Keywords

Onset Detection
Note Segmentation
Music Information Retrieval
Automatic Music Transcription
Novelty Detection
Signal Processing

Acknowledgements

I would like to thank to my supervisors Ricardo Ribeiro and David Martins de Matos for all the helpful discussions and fruitful advices.

Another thank goes to Professors Carlos Costa and Manuela Aparício for all their advices and guidance during the completion of the master course.

A great thank you goes also to my colleagues in the Master in Open Source Software, who pointed some interesting questions concerning the subject of my thesis.

I would like also to thank my friends for being there for me.

Finally, I am most grateful to my family, especially my mother, grand-mother and brother for all the invaluable support and for sharing with me the taste for learning.

Lisboa, Abril 30, 2012

Carlos Manuel Tadeia Rosão

"Music is a moral law. It gives soul to the universe, wings to the mind, flight to the imagination, and charm and gaiety to life and to everything"

Plato

Contents

List of Figures

List of Tables

1 Introduction

1.1 Context & Motivation

It is thought that music appeared between 60.000 and 30.000 years ago (Wallin et al., 2001) when humans started creating several art forms such as painting and jewellery. On the other hand, there is no general agreement in the precise form it appeared. Some suggest that its origin is based on natural occurring sounds and rhythms while others suggest its origin is related to hunting purposes (Wallin et al., 2001).

The sound, and specifically music, have been present in the human life since music was discovered, thus it is no surprise that humans have an innate capacity to understand some of its elements (tone, rhythm, dynamic, texture).

Music has been the subject of study and research for at least more than two millennia. Pythagoras, 2.500 years ago, is usually credited with the first studies in proportion and harmony (Abraham, 1979).

The study of music has continued throughout the centuries, although in the past 50 years, with the fast advances in computers and digitalization, new means were created to aid this study (Casey et al., 2008), which led to faster and in-depth advancements. More recently, when music downloaded over the internet outsold music in CD format, the music industry turned its way even more to music in digitalized format which created an exponential growth in research on this area (Casey et al., 2008).

The work presented with this thesis is part of the broad area of *Music Information Retrieval* (MIR). MIR is a recent and emerging research area devoted to fulfil the listeners' music information needs (Orio, 2006). That is, it aims at finding and retrieving relevant information to humans from the musical signal. This information is not necessarily a mathematical property of the music, such as pitch or *tempo*, it can also be a "psychological property" of

the music, such as musical style or mood (Orio, 2006).

In this thesis, we are mainly concerned with Onset Detection, a very specific subject within the MIR broad area. In a very general way, Onset Detection deals with finding the starting moments of notes in an musical audio signal (Dixon, 2006)[1].

Onset Detection has innumerable applications. It can be used as the first step for segmentation (Duxbury, Bello, Davies, & Sandler, 2003), beat tracking (Davies & Plumbley, 2007) and query by humming (Ding et al., 2011). It can also be used as the basis for the retrieval of many high level musical features (Eyben et al., 2010), such as Chord Estimation, Harmonic Description or Music Genre Classification (Davies & Plumbley, 2007), which allows one to make content-based querying and retrieval (Casey et al., 2008).

Onset Detection can even be applied to Biology, as was done by the pioneering work of Barthet et al. (2010). In this work, onset detection methods were used to analyse the calcium activity in zebrafish (tropical freshwater fish) living embryos.

1.2 Objectives

The main purpose of the work described in this thesis is to explore the most common methods of Onset Detection, trying to understand how their results vary according to the musical type, and which factors influence most their performance.

Two different but complimentary research directions are followed: a comparison of six different Onset Detection Methods and a study in how the Peak Selection step of Onset Detection influences its results.

Both these approaches aim at providing the most complete information about the different Onset Detection methods, that is, how the results of different Onset Detection methods differ amongst themselves, according to the musical type, and which Peak Selection Method suits best each Onset Detection method.

This information can be of great help to someone who wants to use Onset Detection for a particular application – query by humming or segmentation, to name just two examples – and does not know which method suits best his needs.

[1] We will define Onset more precisely in Chapter 2.

1.3 Thesis Contributions

The outcome of this work consists mainly of five contributions:

state of the art - A vast introduction and explanation of the most common Onset Detection Methods that make use of many types of tools: from Time-domain analysis, to Machine-learning approach, passing through the Spectral Domain analysis.

comparison of onset detection methods - A full comparison of the results of six distinct Onset Detection Methods that use Spectral Features.

influence of violin interpretation style on onset detection - A comparison of the influence of violin interpretation style in the results of Onset Detection Methods.

analysis of the peak selection influence on onset detection - A deep study of the influence of five of the most common peak selection techniques in the results of Onset Detection.

open source framework for audio analysis - In order to explore and test the results of the different Onset Detection methods, an audio-framework was build with the Java programming language and made freely available [2].

1.4 Related Publications

Alongside the development of the thesis, three papers were produced:

- Rosão, C. & Ribeiro, R. (2011). Trends in Onset Detection. In *Proceedings of the 2011 Workshop on Open Source and Design of Communication*. ACM.

- Rosão, C., Ribeiro, R., & de Matos, D. Martins. (2012). Comparing Onset Detection Methods Based on Spectral Features. In *Proceedings of the 2012 Workshop on Open Source and Design of Communication*. ACM.

[2] https://github.com/Shemahmforash/Audio-Framework

- Rosão, C., Ribeiro, R., & de Matos, D. Martins. (2012). Influence of Peak Selection Methods in Onset Detection. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012).*

1.5 Outline of the Thesis

This document is organized as follows:

Chapter 1 introduces our aims and gives an overview of the worked developed in the thesis and its contributions.

Chapter 2 gives a general overview of the Onset Detection area. It starts with a musical introduction that helps defining clearly an onset and the different onset classes. Next, the complete process that leads from the audio signal to the onsets is explained. The most common onset detection techniques are grouped by Time-domain, Spectral-domain, Probabilistic, Pitch-based, and Data-driven. Each of the groups is explained mathematically and its strengths and weaknesses introduced. The chapter ends with an overview of the most common peak selection methods.

Chapter 3 explains, evaluates and discusses the results of the different experiments made. It starts by defining the datasets used as well as the evaluation methods that are used to analyse the results. Next, it explains the experiments and discusses the results in three disctinct parts: Comparison of Onset Detection Methods, Influence of Interpretation Style and Influence of Peak Selection Methods.

Chapter 4 concludes this thesis: final conclusions are presented, the contributions of the work developed are revisited, and possible directions for future work are enumerated.

2 Onset Detection

The history of onset detection using computers can be traced back to the 1980's when several algorithms, such as the one presented by Schloss (1985), were proposed to find the beats in music signals composed solely of percussion.

However, only in the 1990's Goto & Muraoka (1994) proposed the first algorithm aimed at finding the rhythm of a general audio signal.

Since the 1990's, as the available computational power increased, the interest in music transcription has grown rapidly and several different algorithms to find onsets in music signals were developed (Klapuri & Davy, 2006).

Before introducing the most common Onset Detection Techniques, one must define precisely what an onset is. In order to do so, a short musical introduction is required.

2.1 Definitions

2.1.1 Musical Introduction

In a general way, music is composed by sounds generated simultaneously by several musical instruments of different kinds (Klapuri & Davy, 2006). Thus, one can consider the notes played by these musical instruments as the basic unit for a musical signal.

Notes can be categorized as harmonic or percussive. Harmonic are the ones which humans would categorize as musical notes, because they have a well defined pitch and harmonically related partials (Klapuri & Davy, 2006). On the other hand, percussive notes, do not have a defined pitch, being analogous to noise clouds (Klapuri & Davy, 2006).

2.1.2 Onset

There are several equivalent ways of defining an onset. One can define it as the start of a musical note (not restricting notes to those having a clearly defined pitch) (Dixon, 2006), the starting moment of an acoustic event (Eyben et al., 2010), or as a single instant chosen to mark the temporally extended transient (Bello et al., 2005). The transient can be understood as a short-time interval in which a significant energy change occurs in the signal (Bello et al., 2005; Klapuri & Davy, 2006).

For the ideal case of a single musical note, one can see a clear definition of these concepts in the schema present in Fig. 2.1.

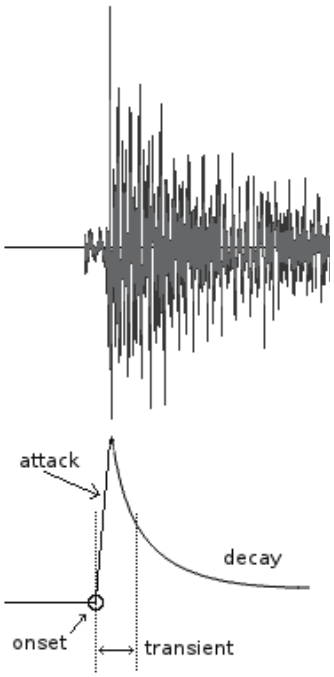

Figure 2.1: Attack, onset and transient in a single note (Bello et al., 2005).

At first sight, an onset seems like a very well-defined concept, although the definition becomes blurred when dealing with polyphonies (see Fig. 2.2) where chords are played "synchronously" (Dixon, 2006) – the start of the notes that compose a chord might spread by tenths of milliseconds – or when we have instruments with long attack times (clarinet for instance) which produce extended transients (see Fig. 2.3), or even when we have certain

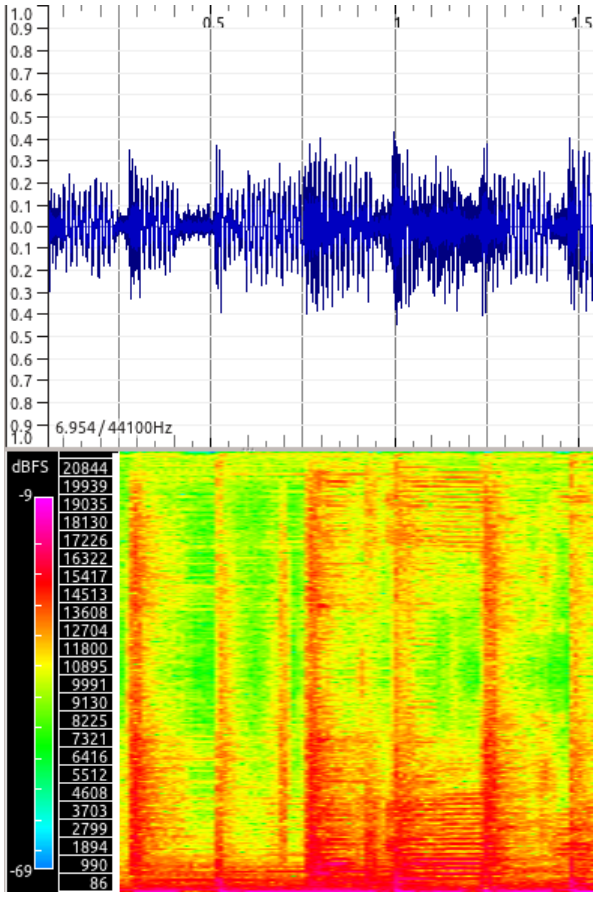

Figure 2.2: Polyphonies – (a)synchronous onsets (signal on top and spectrogram on bottom)

performance techniques such as *glissando*, *tremolo* or *vibrato* (see Fig. 2.4).

To overcome this ambiguities in defining precisely an onset, usually the definition of onset is suited to particular applications (Dixon, 2006).

2.1.3 Onset Classes

As we have seen in Section 2.1.2, we can consider an onset as the starting moment of a musical note, in this way it is possible to categorize onsets the same way we have categorized notes in Section 2.1.1.

Hence, we can distinguish the following onset classes:

- Non-pitched Percussive (NPP);
- Pitched Percussive (PP);

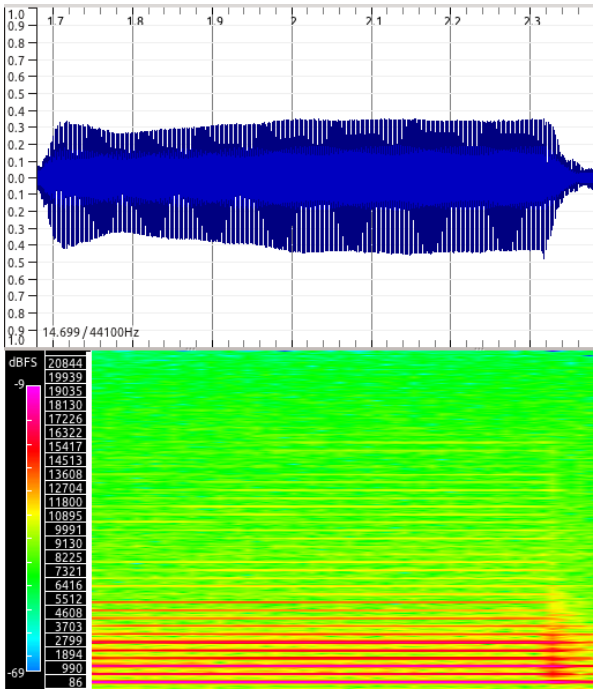

Figure 2.3: An onset produced by a clarinet, as an example of extended transient (signal on top and spectrogram on bottom)

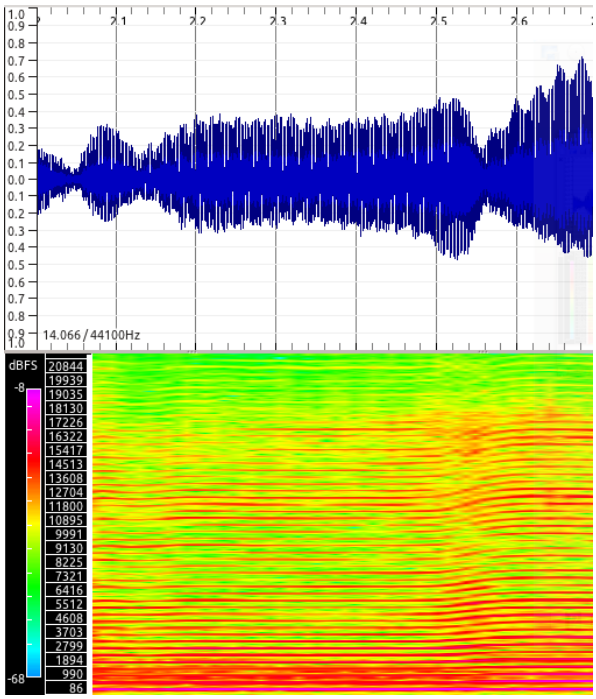

Figure 2.4: The *glissando* as an example of an ambiguous onset (signal on top and spectrogram on bottom)

- Pitched non-percussive (PNP); and,
- Complex Mixtures (Mix).

One can think of Complex Mixture as any polyphonic music where several instruments are played together, something that happens, for instance, in a rock or pop song. The NPP onsets are the ones typically produced by percussion instruments such as drums or cymbals, while the PP onsets are those that have a percussive characteristic but, nonetheless, still maintain a well defined pitch; these onsets appear, for instance, when a piano is playing. Finally, the PNP onsets are those that do not have percussive characteristics and have a very well defined pitch; this category contains onsets from instruments such as bowed strings or wind instruments.

2.2 Onset Detection Methods

Most of the existing Onset Detection algorithms follow a general pattern, depicted in Fig. 2.5, which comprises the following steps (Bello et al., 2005; Eyben et al., 2010; Holzapfel et al., 2010):

1. Preprocessing of the raw audio signal in order to improve the performance of later stages;
2. Computation of a detection function[1], i.e., a function whose peaks should be simultaneous, within a tolerance margin, with onset times (Dixon, 2006); and,
3. Application of a peak-picking algorithm to the detection function in order to select the appropriate peaks.

2.2.1 Preprocessing

Preprocessing is an optional step that transforms the original signal aiming to emphasize its most important properties to onset detection (Bello et al., 2005; Eyben et al., 2010).

[1]In the literature, other terms are also used to denote detection function, for instance Ellis (2007) uses onset strength signal (OSS) and Foote (2007) uses audio novelty. The term audio novelty comes from the novelty function, very common to the literature in machine learning and communication theory.

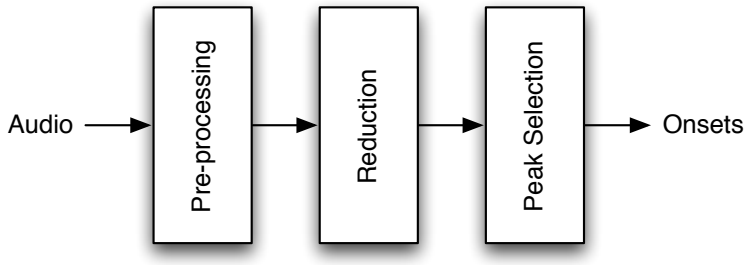

Figure 2.5: Traditional onset detection work-flow (Eyben et al., 2010).

The most common pre-processing methods consist in separating the signal into multiple frequency bands and transient/steady-state separation (Bello et al., 2005). As an example, the algorithms proposed by Goto & Muraoka (1996) and by Scheirer (1998) both use the separation of the original signal in multiple frequency bands. Despite being commonly related to music modelling (Bello et al., 2005), the process of transient/steady-state separation can also be used as a pre-processing method for onset detection; Duxbury et al. (2001) used this possibility in the development of an onset detection method.

There is yet another very common pre-processing method called Adaptive Whitening. This method was proposes by Stowell & Plumbley (2007) and consists in normalising the magnitude of each bin according to a recent maximum, aiming to mitigate the influence of the strong dynamics of most musical audio signals (Stowell & Plumbley, 2007).

2.2.2 Detection Function

A detection function is aimed at detecting changes in the properties of an audio signal (Dixon, 2006), by simplifying it (lowering the sample rate, for instance), but maintaining the important information. Thus, a detection function is the result of a process, sometimes called Reduction, that transforms the original signal into a more simplified function which easily expresses the transients (Bello et al., 2005).

During the years, many detection functions have been proposed. In spite of analysing every approach separately, one can group them in several groups according to their properties. Thus, we have (Eyben et al., 2010): time domain reduction functions, spectral domain reduction functions, probabilistic reduction functions, pitch-based onset detection techniques and data-

driven reduction functions.

2.2.2.1 Time domain reduction functions

Typically, an onset corresponds to an increase in the signal's amplitude (Bello et al., 2005). Early methods, such as the one proposed by Schloss (1985), picked this concept and analysed the amplitude envelope of the signal, obtaining satisfactory results on music with intense percussive transients. This envelop can be obtained by:

$$E_0(n) = \sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1} |x(n+m)|w(m) \tag{2.1}$$

where x is the signal function and $w(m)$ is an N-point rectangular window.

There is an improvement of this method in which the local energy is followed instead of the amplitude:

$$E(n) = \sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1} [x(n+m)]^2 w(m) \tag{2.2}$$

It is possible to further refine this method by analysing the time derivative of the energy and not the energy itself. In this way bursts of energy are converted to peaks in the derivative function (Bello et al., 2005). There is yet another possible refinement which is based on experimental clues that loudness is perceived logarithmically by the human hear (Moore et al., 1997). Accordingly to Klapuri (1999), this strategy causes a better elimination of false onsets. Thus, this method uses the following equation, for the relative difference function W, to mimic the human ear perception of loudness (Klapuri, 1999):

$$W = \frac{d}{dt}(log(E(t))) = \frac{\frac{dE}{dt}}{E} \tag{2.3}$$

2.2.2.2 Spectral domain reduction functions

Instead of analysing the time domain of a signal, recent systems employ reduction functions that work in the spectral domain (Eyben et al., 2010). In Fig. 2.6 one can see this two representations for the same excerpt of piano music.

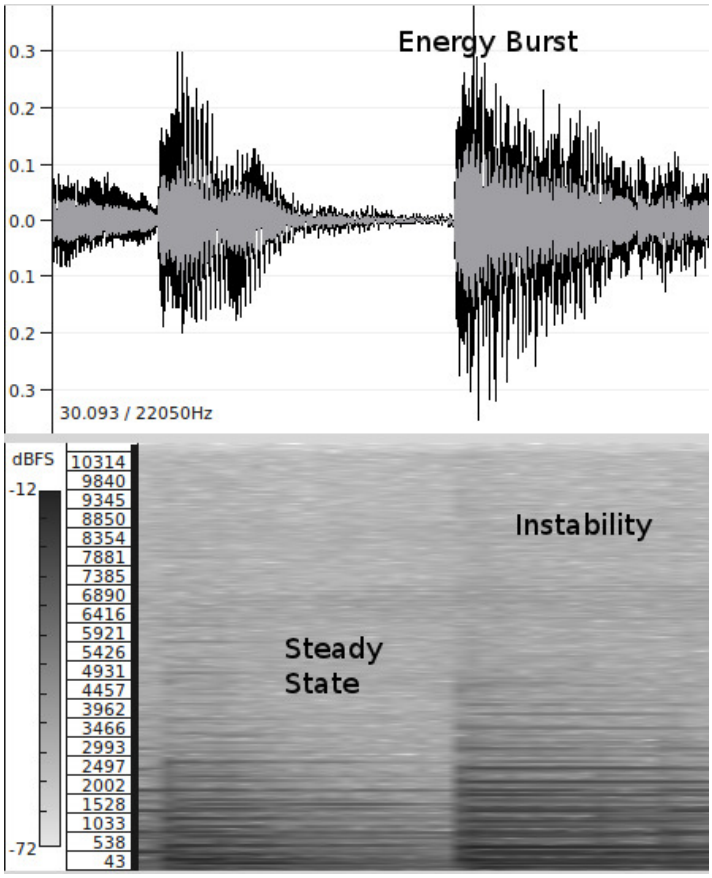

Figure 2.6: Time vs Amplitude (top) and Time vs Spectral Magnitude (bottom) representation of piano notes.

In the following paragraphs we will briefly explain several onset detection methods belonging to the group of spectral domain reduction functions. Note that the methods listed in this subsection are all based on a STFT of the signal, that for a general signal $x(n)$ and the n^{th} bin of the k^{th} frequency can be defined as:

$$X_k(n) = \sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1} [x(nh+m)]^2 w(m) e^{-\frac{-2i\pi mk}{N}} \tag{2.4}$$

where $w(m)$ represents an N-point rectangular window and h the hop-size, i.e., the time-shift between adjacent windows.

From the STFT, one can define an energy envelope function by summing the power of

frequency components in the spectrogram (Klapuri & Davy, 2006):

$$E(n) = \frac{1}{N} \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |X_k(n)|^2 \tag{2.5}$$

2.2.2.2.1 High frequency content By observing that an energy increase in one or more frequency bands can be a simple indicator of an onset (Dixon, 2006), one can notice that an onset has a more intense energy in the bands in which the interference with other simultaneous components is smaller (Dixon, 2006), a situation which typically occurs in the high-frequencies region (Rodet & Jaillet, 2001; Bello et al., 2005). This fact can be exploited by weighting each STFT bin with a factor proportional to its frequency. Hence, by summing all weighted bins, one obtains a function called HFC or $\widetilde{E}$ – see Fig. 2.7 –, that can be used as detection function:

$$HFC(n) = \frac{1}{N} \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} W_k |X_k(n)|^2 \tag{2.6}$$

where W_k represents the frequency dependent weighting.

The authors Masri (1996) and Masri & Bateman (1996) proposed a linear weighting of the frequencies with $W_k = |k|$. Although this method works well for percussive onsets, it shows weaknesses for other onset types (Bello et al., 2005).

Later, Rodet & Jaillet (2001), proposed a similar method and obtained good results for high-frequencies, thus showing consistency with Masri's idea.

2.2.2.2.2 Spectral Difference There is a more general approach based in spectral changes of the signal, and is related to the formulation of the detection function as a "distance" between successive STFT, treating them as points in an N-dimensional space (Bello et al., 2005). According to the distance function used, the detection function receives different names:

1. If using the L1-norm as a distance function, then, the detection function is called SF (Dixon, 2006; Masri, 1996) – see Fig. 2.8;

2. If using other distance function, for instance the L2-norm or the Kullback–Leibler di-

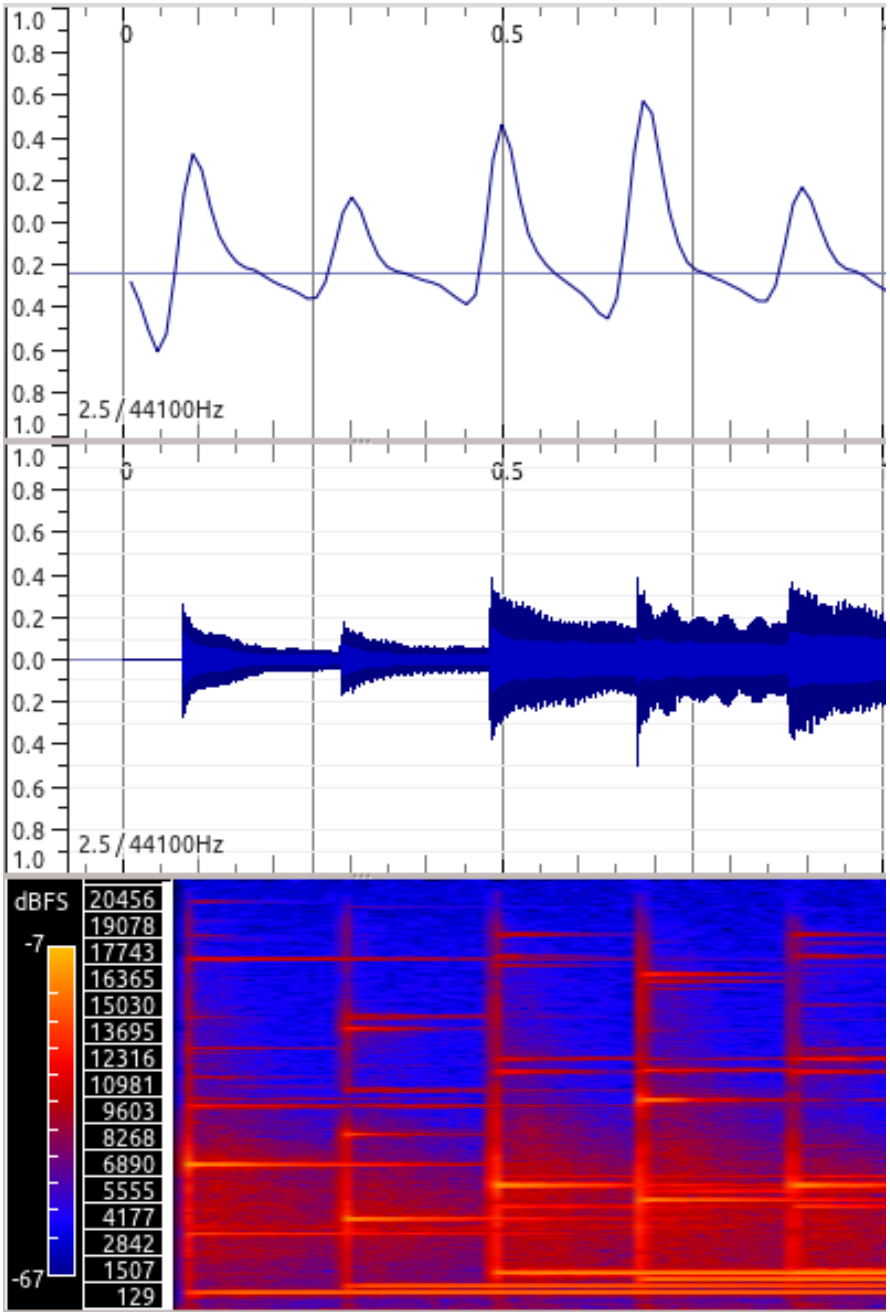

Figure 2.7: High Frequency Content (top), Signal (middle), and Spectrogram (bottom) for 1s of a PP song from the Bello Dataset.

vergence, then, the detection function is called SD (Duxbury et al. (2002) developed an algorithm that uses this distance function).

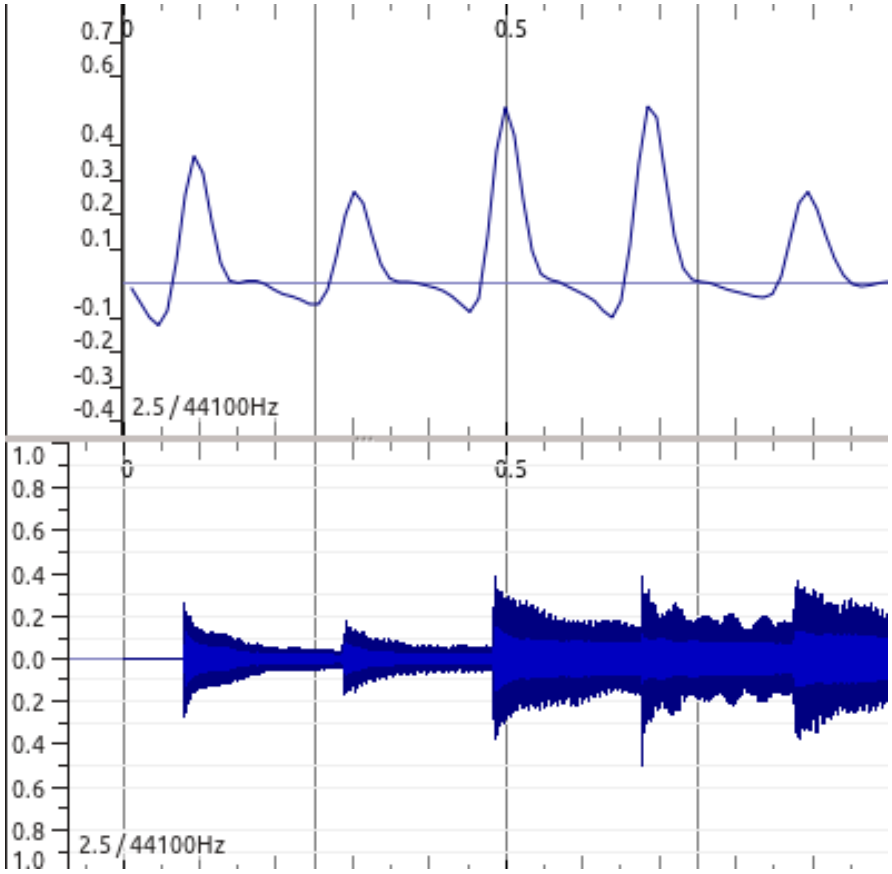

Figure 2.8: Spectral Flux (top) and Signal (bottom) for 1s of a PP song from the Bello Dataset.

Dixon (2006) showed that the results with the L1-norm outperform those with the L2-norm. Nonetheless, in any of the cases, the function measures the change in magnitude in each frequency bin (Dixon, 2006) and is calculated by computing the difference of two consecutive short-time spectra bin by bin (Eyben et al., 2010) (using any of the distance functions referred above to calculate the difference). With the L1-norm it becomes:

$$SF(n) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} H(|X_k(n)| - |X_k(n-1)|) \tag{2.7}$$

and with the L2-norm it takes the form:

$$SD_{L2}(n) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} \{H(|X_k(n)| - |X_k(n-1)|)\}^2 \tag{2.8}$$

where $H(x) = \frac{x+|x|}{2}$ is called the half-wave rectifier function (Dixon, 2006) and has the purpose of eliminating negative differences. In this way, it ignores offsets and sticks to onsets.

There is also a related form of Spectral Difference function introduced by Foote (2007), which uses the correlation between STFT to be calculated. However, according to Bello et al. (2005), this approach is only effective when small width windows are used in its calculus.

The methods using Spectral Difference or Spectral Flux give very good results in finding NPP onsets (Bello et al., 2005). Furthermore, these methods have proved to be among the best overall performers so far (Eyben et al., 2010), so it is not surprising to see many studies using them (Holzapfel et al., 2010), like, for instance, the works of Hainsworth & Macleod (2003), Collins (2005a) and Dixon (2006).

2.2.2.2.3 Phase deviation The methods we have seen so far use the magnitude of the spectrum as their source of information, however, in recent years, several studies make use of the phase spectra. This type of analysis is also important, because much of the temporal structure of a signal is contained in the phase spectrum (Bello et al., 2005).

Let $\phi_k(n)$ be the phase of the transformed signal $X_k(n)$, i.e., respects the relation:

$$X_k(n) = |X_k(n)|e^{i\phi_k(n)} \tag{2.9}$$

where ϕ is mapped onto the range $]-\pi, \pi]$.

It is unlikely that the frequency components of the new sound are in phase with the previous sound, so irregularities in the phase of several frequency bins can be used to indicate the presence of onsets (Dixon, 2006). According to Eyben et al. (2010), the change of the phase in a STFT frequency bin is a rough estimate of its instantaneous frequency, and can be used as indicator of an onset.

Mathematically, a variation is expressed as the derivative (i.e., the difference for discrete

signals), so, one can write the instantaneous frequency as:

$$\phi'_k(n) = \phi_k(n) - \phi_k(n-1) \tag{2.10}$$

And the variation of the instantaneous frequency as:

$$\phi''_k(n) = \phi'_k(n) - \phi'_k(n-1) \tag{2.11}$$

In 2003, Bello & Sandler (2003), proposed an algorithm that analyses the distribution of phase deviations across all frequencies. Later in the same year, Duxbury, Bello, Davies, & Sandler (2003) proposed a simpler measure of the spread of the previous distribution, denoted by PD:

$$PD(n) = \frac{1}{N} \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |\phi''_k(n)| \tag{2.12}$$

A few years later, Dixon (2006) proposed a refinement of the previous method, so it can account for "noise introduced by components with no significant energy" (Bello et al., 2005). The PD considers all frequencies equally, so Dixon (2006) proposed weighting the frequency bins by their magnitude, in order to obtain a new onset detection function which he named WPD:

$$WPD(n) = \frac{1}{N} \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |X_k(n)\phi''_k(n)| \tag{2.13}$$

where $X_k(n)$ represents the magnitude.

A further option is to normalize the previous equation, obtaining a NWPD function:

$$NWPD(n) = \frac{\sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |X_k(n)\phi''_k(n)|}{\sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |X_k(n)|} \tag{2.14}$$

These methods based on phase deviation tend to give better results on PNP onsets than the methods that use the spectral magnitude (Bello et al., 2005), though, in NPP onsets the results are not as good as the ones obtained with the other type of methods.

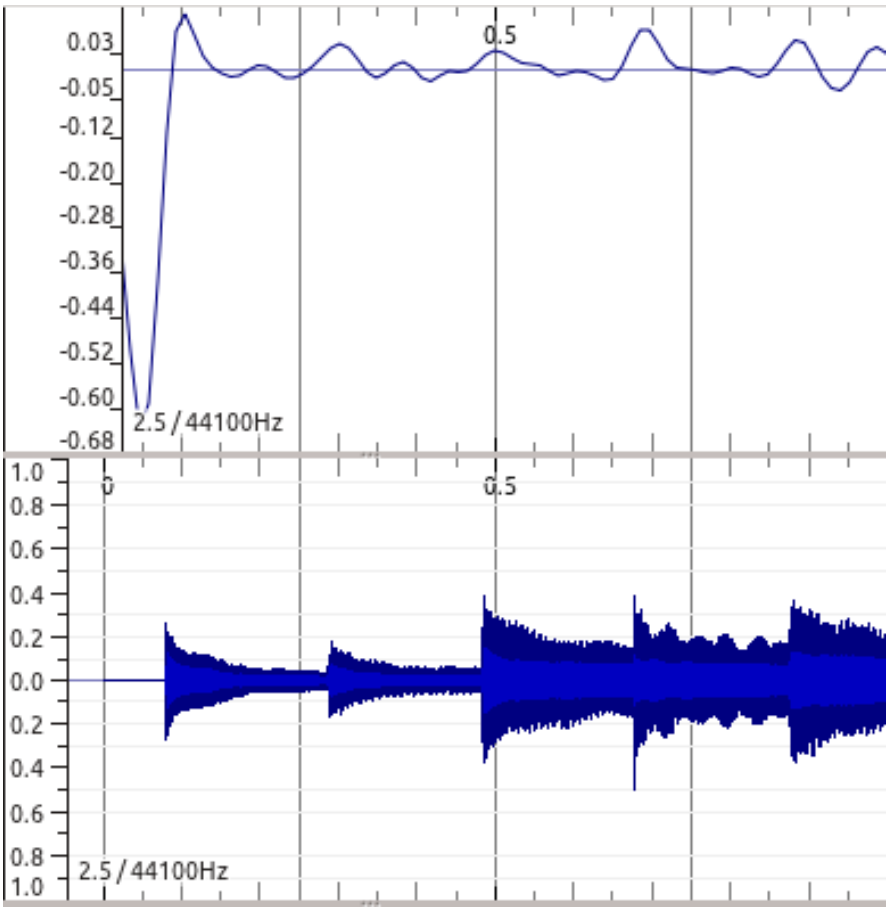

Figure 2.9: Phase Deviation (top) and Signal (bottom) for 1s of a PP song from the Bello Dataset.

2.2.2.2.4 Complex domain Like in the previously defined WPD and NWPD functions, one can combine both energy and phase information for the production of a CD function (Duxbury, Bello, Davies, & Sandler, 2003).

In this way, amplitude and energy are used together to check for irregularities in the steady state (Dixon, 2006). A method that uses this concept was first proposed by Duxbury, Bello, Davies, Sandler, & Others (2003), and later extended by Bello et al. (2004).

A few years later, Dixon (2006) made an equivalent, but simpler formulation, of Bello's approach. He defined the CD as:

$$CD(n) = \frac{1}{N} \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |X_k(n) - \tilde{X}_k(n)| \tag{2.15}$$

where X_k is the amplitude and phase of the current bin, based on the previous two bins, and $\tilde{X}_k$ is the target value, i.e., the value predicted by the previous frames. The prediction assumes constant amplitude and rate of phase change:

$$\tilde{X}_k(n) = |X_k(n-1)|e^{i(\phi_k(n-1)+\phi'_k(n-1))} \tag{2.16}$$

All in all, this method sums the magnitude of the complex differences between the actual values for each frequency bin and the estimated ones, and uses the result as a detection function.

In Fig. 2.10 we see a representation of this OSS.

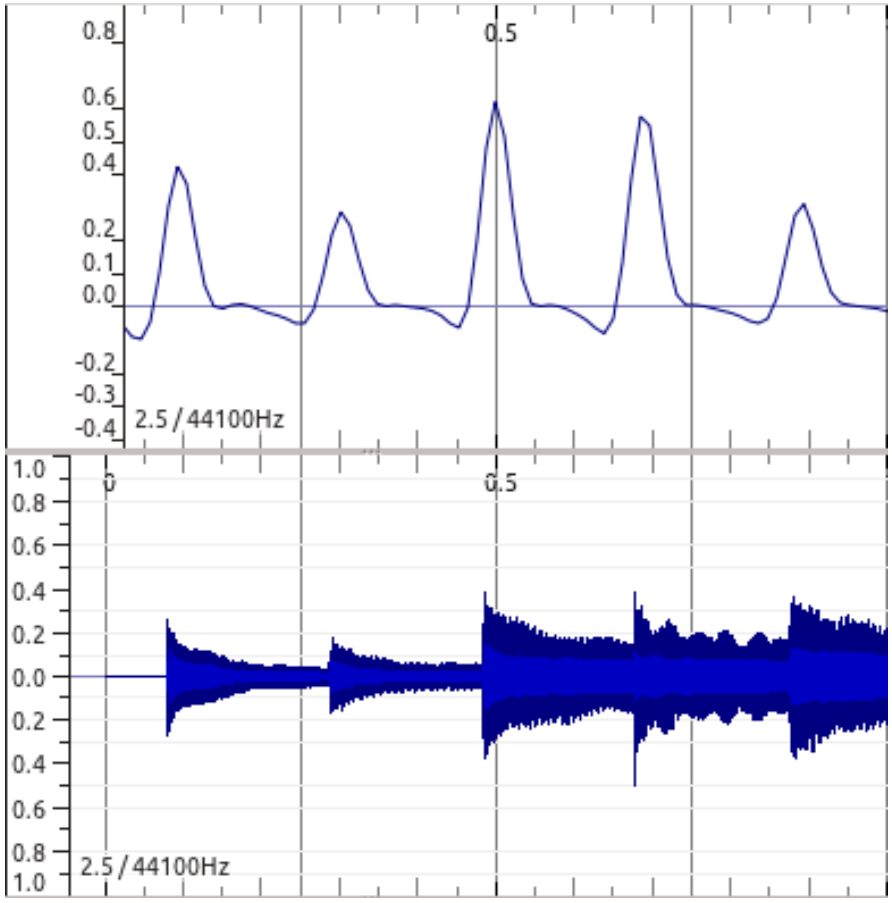

Figure 2.10: Complex Domain (top) and Signal (bottom) for 1s of a PP song from the Bello Dataset.

Dixon (2006) noted that this method does not distinguish well between onsets and offsets,

so in 2006 he proposed a RCD function to surpass this problem. This method uses a half-wave rectification (similar to the SF method) in order to preserve only the positive variations of energy in the spectral bins. Hence, the RCD is defined as:

$$RCD(n) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} RCD_k(n) \tag{2.17}$$

where

$$RCD_k(n) = \begin{cases} |X_k(n) - \tilde{X_k}(n)| & \text{if } |X_k(n)| > |X_k(n-1)| \\ \\ 0 & \text{otherwise} \end{cases} \tag{2.18}$$

Although being a little harder to implement, these kind of methods show very good results in the detection of both NPP and PNP onsets (Bello et al., 2005).

2.2.2.3 Probabilistic reduction functions

An alternative to the previous models is to base the description of the signals on probabilistic/statistical methods, i.e., assuming that the signal can be described by some probabilistic model and to look for abrupt difference between the model and the "real signal". Obviously, the success of this approach is dependent on the degree of closeness between the model and the "real signal". This type of similarity can be quantified using measures of likelihood or Bayesian model selection criteria (Bello et al., 2005).

Every statistical method of this kind follows one of two main strategies:

1. Approaches Based on "Surprise Signals" — the strategy here is to look for a sudden change (surprise signal) of a particular global model that is assumed to describe the signal (Bello et al., 2005);

2. Model-Based Change Point Detection Methods — in this case, the goal is to observe which of the two previously defined probability models the system follows (Eyben et al., 2010).

The negative log-probability, which is used by Abdallah & Plumbley (2003), for instance,

is an example of a model based on "surprise signals" and shows very good results in music with non-percussive onsets (Bello et al., 2005). On the other hand, the sequential probability ratio test, applied, for instance by Basseville & Nikiforov (1993), is a good example of the utility of the second strategy.

The statistical onset detection methods have the peculiarity of providing very good results in the discovery of non-percussive onsets (Bello et al., 2005).

2.2.2.4 Pitch-based onset detection techniques

Pitch is a perceived sound property closely related to frequency that allows one to organize sounds in a frequency-related scale, i.e., with notions of "higher" and "lower" in the common way associated with melody (Klapuri & Davy, 2006). This means that pitch is not a physical property of sound, it is a subjective property, commonly studied in the area of psychoachoustics.

The physical concept closer to pitch is the fundamental frequency, f_0, though, for polyphonic music signals, there are some dificulties in finding the value of this property (Klapuri & Davy, 2006). Pitch-based onset detection methods use the concept of pitch contour, i.e., the variation in time of the fundamental frequency, and assume that discontinuities and abrupt changes in this function indicate the presence of onsets (Eyben et al., 2010). Hence, pitch-based onset detectors are based on finding these kind of discontinuities and abrupt changes in the pitch contour (Collins, 2005b).

Several algorithms were proposed, for instance, the works of Collins (2005b) or Zhou & Reiss (2007), that use this kind of strategy, and according to the comparative study performed by Bello et al. (2005), they show better results in PNP onsets than the algorithms based in the magnitude of the spectrum (Collins, 2005b).

2.2.2.5 Data-driven reduction functions

One of the difficulties that the previous methods have shown is that they perform well only for certain types of music: performance diminishes when range of music signals increases (Eyben et al., 2010).

In order to try to overcome this limitation, data-driven reduction functions have been proposed.

These functions tend to depend on machine-learning algorithms, such as neural networks and some of them have great performance in a large spectrum of music types. For instance, Lacoste & Eck (2005) is based on a feed forward neural network and achieved the best performance in the Mirex 2005 audio onset detection evaluation (MIREX, 2005), and Eyben et al. (2010) uses a bidirectional Long Short-Term Memory recurrent neural network and was the algorithm which performed best in the Mirex 2010 audio onset detection evaluation (MIREX, 2010).

Statistical methods seem to have the best overall performance. However, computational requirements and the need of a training corpus may difficult their application (Bello et al., 2005).

2.2.3 Peak Selection

A detection function created with any of the above methods shows well localized maxima, generally with some variability due to noise (Bello et al., 2005). In order to pick the onsets from the detection function, several steps are used that typically fit in the following categories: Post-processing, Thresholding and Peak-picking.

2.2.3.1 Post-processing

Post-processing deals with simplifying the subsequent processes of thresholding and peak picking by increasing the uniformity in the detection function (Bello et al., 2005). This process of increasing the uniformity of the detection function typically makes use of normalization methods and filters.

The normalization typically works in one of two ways (Dixon, 2006; Holzapfel et al., 2010):

- Subtract the average value of the function from each value, so that the average will be zero and then divide by the maximum value so that the function will be in the interval [-1,1] (later we will call this normalization method by norm).

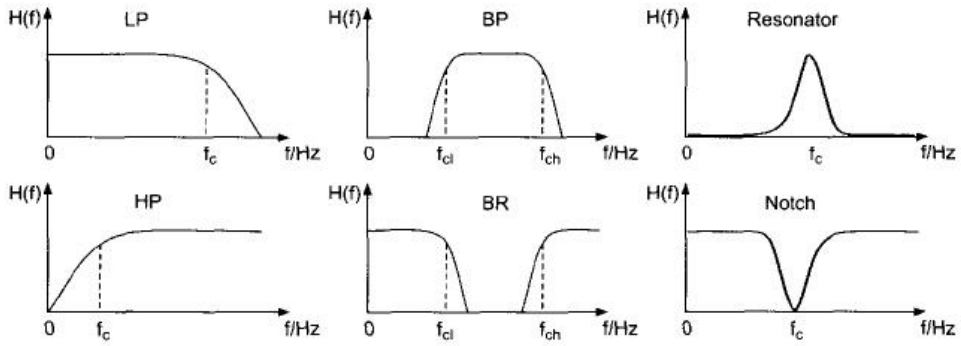

Figure 2.11: Filter classification (Zölzer et al., 2002).

- Subtract the average value of the function from each value and then divide by the maximum absolute deviation, so that the average will be 0 and the standard deviation 1 (later we will call this normalization method by stdev).

The filters used are typically low pass filters (Bello et al., 2005; Holzapfel et al., 2010; Dixon, 2006), which, in general, select low frequencies up to the cut-off frequency (f_c) and attenuate frequencies higher than f_c (Zölzer et al., 2002) and can be defined as:

$$y_i = \alpha x_i + (1 - \alpha) y_{i-1} \tag{2.19}$$

With the smoothing factor α defined as:

$$\alpha = \frac{\Delta T}{RC + \Delta T} \tag{2.20}$$

where ΔT is the inverse of the sampling frequency and the cut-off frequency is:

$$f_c = \frac{1}{2\pi RC} \tag{2.21}$$

2.2.3.2 Thresholding

Even after post-processing there will be certain peaks that do not correspond to onsets. So, it is common to define a threshold that separates event-related and non-event-related peaks (Bello et al., 2005).

The first approach is to define a constant threshold (Klapuri et al., 2006) (see Fig. 2.12), δ, and, in this case, onsets would be peaks where the detection function, d, is bigger than the threshold:

$$d(n) \geq \delta \tag{2.22}$$

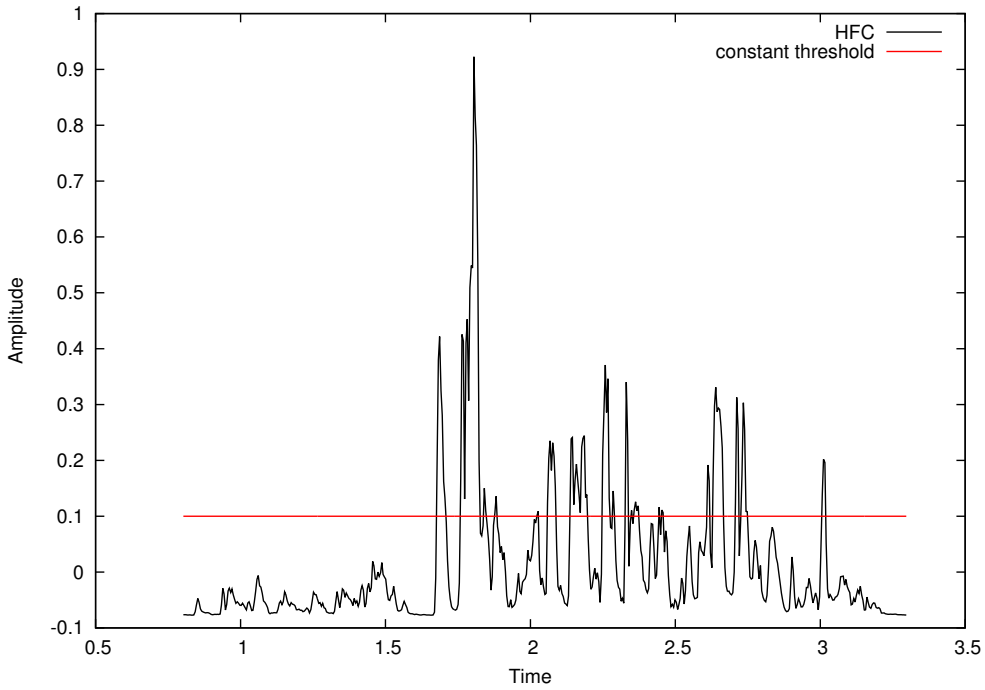

Figure 2.12: Constant threshold.

Since music typically exhibits great dynamics, constant thresholds usually give weak results (Bello et al., 2005), so it is common to use adaptive thresholds (Holzapfel et al., 2010; Dixon, 2006). An adaptive threshold can be constructed in several ways. An option is to make use of a linear LP-FIR filter (Holzapfel et al., 2010; Bello et al., 2005):

$$\tilde{\delta} = \delta + \sum_{i=0}^{M} a_i d(n-i) \tag{2.23}$$

where a_i are the filter coefficients. Other option is to use the square of the detection function (Bello et al., 2005):

$$\tilde{\delta} = \delta + \lambda \sum_{i=-M}^{M} w_i d^2(n-i) \tag{2.24}$$

where λ is a positive constant and w_i is a smooth window of choice.

Although better than the constant threshold, the previous adaptive thresholds still present problems when facing musics with great dynamical change. The best way to overcome this problems is to build a function based on the local mean (see Fig 2.13) or local median of the detection function, d (Kauppinen, 2002):

$$\tilde{\delta}(n) = \delta + \lambda \,\mathrm{mean}(|d(n-M)|, ..., |d(n+M)|) \tag{2.25}$$

or

$$\tilde{\delta}(n) = \delta + \lambda \,\mathrm{median}(|d(n-M)|, ..., |d(n+M)|) \tag{2.26}$$

where λ and δ are positive constants, that can be tweaked, and M is the size of a window around each of the points of the detection function.

These threshold functions based on the mean and on the median are the most robust to signal dynamics (Bello et al., 2005; Dixon, 2006; Kauppinen, 2002).

It is important to notice that the parameters used in thresholding have a large impact on the final results, mainly in the ratio of false positives to false negatives (Dixon, 2006).

2.2.3.3 Peak-picking

After the above procedures, picking the onsets, $o(n)$, is reduced to the identification of local maxima above the defined threshold, which can be summarized as (Eyben et al., 2010):

$$o(n) = \begin{cases} 1 & \text{if } d(n) > \tilde{\delta}(n) \\ & \text{and} \quad d(n-w) \leq d(n) \leq d(n+w) \\ 0 & \text{otherwise} \end{cases} \tag{2.27}$$

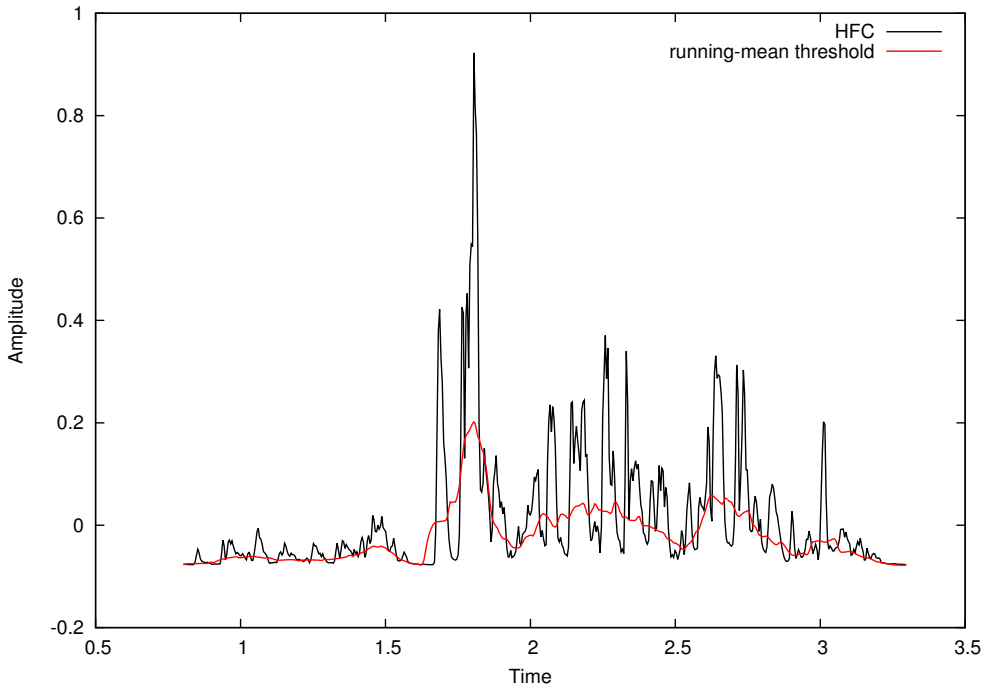

Figure 2.13: Running-mean adaptive threshold.

2.3 Summary

In this chapter, we define the concept of onset, discussing the difficulties and ambiguities in the definition, and the concept of onset class, that is, the classification of onsets.

The complete process that leads from the audio signal to the onsets – pre-processing, reduction and peak selection – is explained and each step focused.

Pre-processing methods, in general, try to emphasize the features relevant to onset detection, in order to help the subsequent stages.

When it comes to the Detection Function step of Onset Detection, it can be of several types. From the perspective of representation of the signal, the Onset Detection functions can be of Time Domain or Spectral Domain type; although, according to other properties, we have also other possibilities for classifying detection functions: probabilistic, pitch-based, and data-driven.

The spectral-domain OSS are the most focused on this chapter, as this is the type of functions created and explored in Chapter 3. The most common spectral-domain OSS are the

High Frequency Function, the Spectral Flux, the Phase Deviation family and the Complex Domain family. The HFC, SF and PD use only information from the magnitude of the spectrum, while the Complex Domain family (CD and RCD functions) and the WPD function use information from both the magnitude and phase of the spectrum.

Next, we describe the Peak Selection step of onset detection, that is, how can one select the peaks of the OSS functions that correspond to onsets. This step usually consists of 3 parts: post-processing, thresholding and peak selection. The post-processing part is responsible for making the OSS more uniform, while the threshold – usually a dynamic and not a constant threshold – separate even-related from non-event-related peaks. Finally, the peak picking is responsible for collecting the position of the peaks that are greater than the threshold.

3 Analysis of Onset Detection Performance

Aiming at providing the most complete possible information about Onset Detection, a selection of six Onset Detection Methods using Spectral Features, described in Chapter 2, were explored in this chapter using two approaches:

- Direct comparison of the performance of each Onset Detection Method;
- Assessment of the influence of the Peak Selection part of Onset Detection process on the results of each OSS.

In order to be able to follow these two approaches, one must define first how the results are evaluated – the evaluation metrics (Section 3.1) – and on which audio signals the methods will be tested – the datasets (Section 3.2).

3.1 Evaluation Metrics

When facing classification problems, the main source for performance measurements is called a Classification Matrix (Olson & Delen, 2008), and it can be defined as in Table 3.1.

	True (Reference)	
Predicted (Inferred)	True Positive (TP)	False Positive (FP)
	False Negative (FP)	True Negative (TN)

Table 3.1: Classification Matrix

The upper-left to lower right diagonal of the matrix represent the correct decisions made while the other diagonal represent the errors made (Olson & Delen, 2008). With the help of this matrix, we can define several useful quantities, that are depicted in Fig. 3.1.

With the help of both the classification matrix (Table 3.1) and Fig. 3.1, it is possible to define Precision, Recall and F-measure, the quantities that will be used to evaluate the results in Section 3.3, Section 3.4 and Section 3.5.

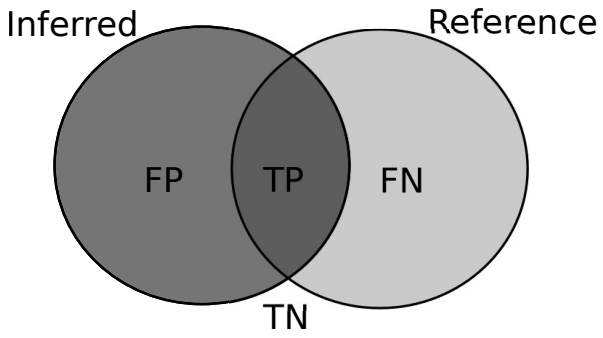

Figure 3.1: Relationship between the quantities defined in the Classification Matrix (Manning et al., 2008)

The Precision (also called overall accuracy (Olson & Delen, 2008)), that is, the fraction of retrieved instances that are relevant, is defined by dividing the total correctly detected positives by the total number of returned results.

$$Precision = \frac{TP}{TP + FP} \tag{3.1}$$

On the other hand, the Recall (also called true positive rate or hit rate), that is, the fraction of relevant instances that are retrieved, is obtained by dividing the correctly detected positives by the number of results that should have been returned.

$$Recal = \frac{TP}{TP + FN} \tag{3.2}$$

In the particular case of onset detection one can interpret the TP as the correctly detected onsets, the FP as falsely detected onsets and the FN as onsets that were not detected.

It is important to note that the Mirex Onset Detection Task specifications (MIREX, 2011), and most of the papers in this area, consider onsets detected as TP if they are in a window of $50ms$ around the annotated onset. On the other hand, if more than one detection fall inside the same tolerance window, only one is counted as TP, the others are considered as FP. When a detection is inside the tolerance window of two onset annotations, one TP and one FN are counted.

We can also define a quantity that joins both precision and recall in a sort of weighted average. It is called F-measure and can be obtained by:

$$\text{F-measure} = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2 \cdot P \cdot R}{P + R} \tag{3.3}$$

There is another form of evaluation, called Correct Detection Rate(CDR) (Liu et al., 2003; Collins, 2005a), defined as:

$$\text{CDR} = \frac{total - FN - FP}{total} \tag{3.4}$$

3.2 Datasets

The main difficulty with evaluating an onset detection algorithm is that of having a significantly large balanced songs for which the onset times are known (Dixon, 2006). This is not easy to achieve, because precise onset times are available only to a small amount of musics, such as produced by computer-monitored pianos (Dixon, 2006), and all the other has to be labelled by hand, which is an error-prone task.

With these difficulties in mind, we chose to datasets: one publicly available, and the other developed for a paper by Bello et al. (2005).

3.2.1 Alicante Dataset

The Alicante Dataset is a publicly available annotated dataset (Alicante, 2012) created by the Pattern Recognition and Artificial Intelligence Group of the University of Alicante (PRAIg-UA). This dataset contains 19 real recordings that cover a relatively wide range of instruments and musical genres (Alicante, 2012) and 2155 onsets. The songs are available in WAV format (sample rate 22.050 kHz, mono, 16 bit) and their onset positions (in seconds) in text format.

We can divide the recordings of the dataset in 3 distinct groups, according to the characteristic of their onsets: Complex Mixture (Mix), Pitched Non-percussive (PNP) and Pitched Percussive (PP), as show in Table 3.2.

The ground truth onset labelling in this dataset was marked and reviewed with the help of the software Speech Filling QSystem (Sciences, 2012).

	No. Songs	No. Onsets
Mix	11	1397
PNP	3	254
PP	5	504
Total	19	2155

Table 3.2: Alicante Dataset Structure

3.2.2 Bello Dataset

The Bello Dataset is a hand labelled and annotated dataset first proposed in Bello et al. (2005) and used in several other works, such as the ones of Dixon (2006) and Holzapfel et al. (2010). It contains commercial and non-commercial recordings, covering a variety of musical styles and instrumentations, in a total of 23 songs (Bello et al., 2005) and 1065 onsets. The songs are available in WAV format (sample rate 22.050 kHz, mono, 16 bit) and their onset positions (in seconds) in text format.

One can group the songs in this dataset just as it was done for the previous dataset, but adding the NPP class that was absent in the Alicante Dataset. This grouping of songs by onset classes is described in Table 3.3.

	No. Songs	No. Onsets
Mix	7	271
PNP	1	93
PP	9	489
NPP	6	212
Total	23	1065

Table 3.3: Bello Dataset Structure

3.3 Comparison of OSS functions Using Spectral Features

In order to compare the performance of the different Onset Detection Methods, the algorithms of 6 Onset Detection Methods using Spectral Features were reproduced: HFC, SF, PD, WPD, CD, and RCD.

Several experiments were created to test and compare the behaviour of these Onset Detection Methods according to the different Onset Classes and to the two datasets described

in the previous section.

3.3.1 Experiments

The following general steps were used to run the experiments:

- Read the data from the WAV file;
- Build the detection functions with the help of a spectral-domain representation of the signal obtained by the calculus of consecutive STFT (with Hamming-window size and hop size as parameters);
- Normalize the detection function so that it will be in the interval [-1, 1] and its average will be zero, that is, for each value one subtracts the mean and divides by the maximum value;
- Create a running-median threshold, as defined in Eq. 2.25, with λ, *delta* and M as parameters;
- Consider as onset every value in the detection function that is bigger than the threshold and is a local maximum (in a window of 3 samples around it).

We ran our experiments in the Alicante and Bello datasets by varying the parameters with the purpose of maximizing the F-measure for each group of onsets (Mix, PP, PNP and NPP) described in the previous subsection. By multiple experiments we found that the maximum values for F-measure appear when we use a window of 1024 samples (that is 46.4ms in this 22.05kHz sampled signals) with hop size of 50% when calculating the STFT and by setting $M = 10$ in Eq. 2.25. Moreover, we found that the best threshold parameters λ and δ varied per group of musics. Hence, the results presented and discussed in the next section were all obtained by varying the parameters λ and δ between 0 and 1, maintaining the other parameters fixed to the values stated.

	F-measure	Precision	Recall
HFC	0.708	0.724	0.692
SF	**0.840**	**0.829**	**0.852**
PD	0.455	0.390	0.557
WPD	0.704	0.693	0.715
CD	0.768	0.781	0.755
RCD	0.771	0.741	0.804

Table 3.4: Overal Results by OSS in the Alicante Dataset

	F-measure	Precision	Recall
HFC	0.781	0.760	0.806
SF	**0.921**	**0.930**	**0.913**
PD	0.589	0.465	0.816
WPD	0.794	0.780	0.813
CD	0.835	0.830	0.843
RCD	0.856	0.856	0.860

Table 3.5: Overal Results by OSS in the Bello Dataset

	PP			PNP			Mix		
	F	P	R	F	P	R	F	P	R
HFC	0.814	0.849	0.801	0.542	0.555	0.529	0.783	0.784	0.782
SF	**0.921**	**0.916**	**0.926**	**0.780**	**0.794**	**0.767**	0.816	0.776	0.861
PD	0.377	0.281	0.668	0.473	0.423	0.534	0.529	0.468	0.610
WPD	0.825	0.822	0.843	0.606	0.612	0.600	0.804	0.791	0.816
CD	0.883	0.923	0.848	0.621	0.651	0.594	**0.824**	**0.827**	**0.821**
RCD	0.893	0.911	0.875	0.694	0.675	0.714	0.797	0.747	0.854

Table 3.6: Results for the Alicante Dataset: Precision(P), Recall(R), and F-measure(F)

	NPP			PP		
	F	P	R	F	P	R
HFC	0.921	0.922	0.920	0.838	0.846	0.830
SF	0.931	0.946	0.926	**0.961**	**0.964**	**0.947**
PD	0.652	0.573	0.819	0.497	0.410	0.734
WPD	0.916	0.959	0.882	0.810	0.796	0.826
CD	**0.947**	**0.978**	**0.923**	0.883	0.892	0.874
RCD	0.933	0.977	0.903	0.882	0.880	0.883

Table 3.7: Results for NPP and PP onset classes in the Bello Dataset: Precision(P), Recall(R), and F-measure(F)

	PNP			Mix		
	F	P	R	F	P	R
HFC	0.553	0.519	0.591	0.812	0.753	0.881
SF	**0.911**	**0.888**	**0.935**	**0.880**	**0.922**	**0.842**
PD	0.615	0.479	0.860	0.540	0.396	0.851
WPD	0.660	0.602	0.731	0.832	0.762	0.811
CD	0.684	0.650	0.720	0.866	0.798	0.854
RCD	0.808	0.745	0.882	0.824	0.823	0.770

Table 3.8: Results for PNP and Mix onset classes in the Bello Dataset: Precision(P), Recall(R), and F-measure(F)

3.3.2 Results & Discussion

The results of our experiments are shown in Table 3.4, to Table 3.8, and in Figures 3.2 and 3.3. These results were obtained by choosing the parameters that maximize the F-measure for each group of onsets.

The results in the two datasets allow us to notice right away that the PNP onset type has the lower performance, that is, soft onsets are harder to detect. This result is in complete agreement with the results in the literature (Bello et al., 2005; Dixon, 2006; Holzapfel et al., 2010).

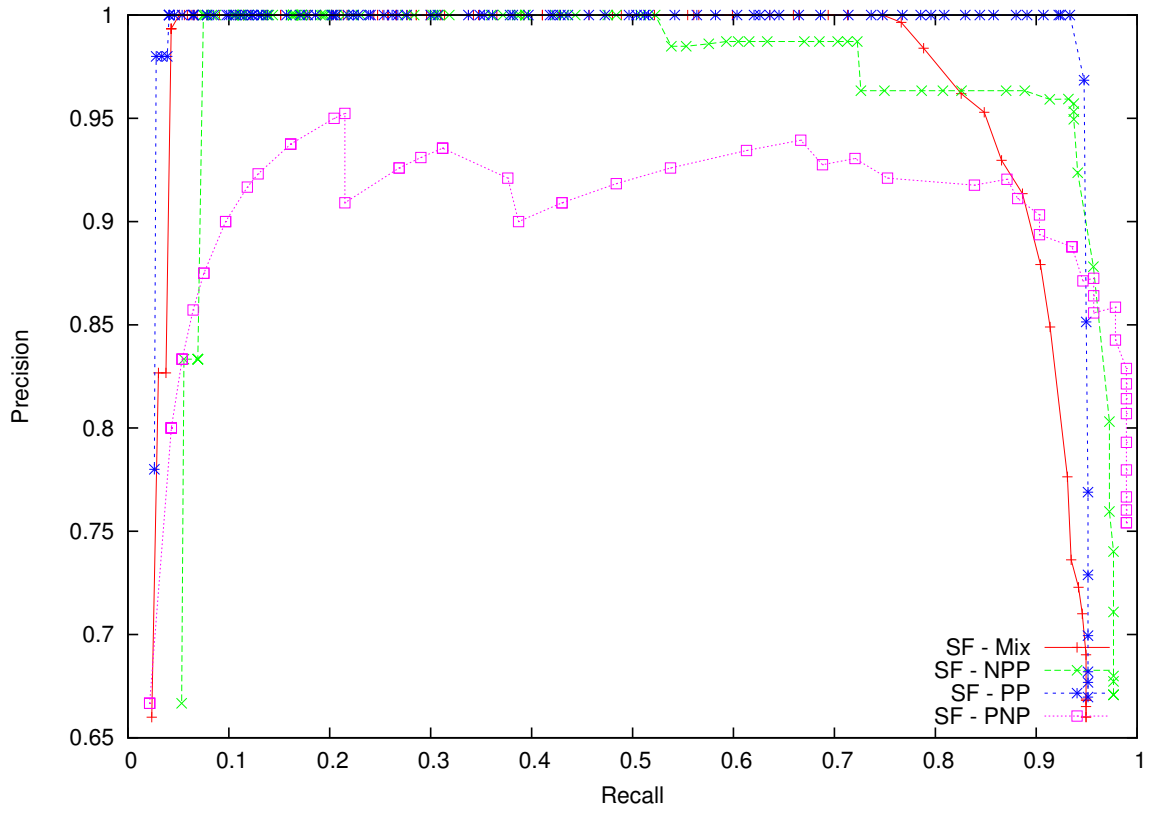

Figure 3.2: Precision vs Recall for the SF OSS in all the onset classes of the Bello Dataset.

In general, the results are better for the NPP class (in the Bello Dataset), followed by the PP, Mix and lastly by the PNP class. Although there are some exceptions: the PD OSS performs best for the Mix onsets on the Alicante dataset and the in the Bello Dataset it performs worse for the PP class.

The organization of the tables allow us to understand easily how each detection function performs on the different onset classes.

Globally, when looking at the results, one can clearly see that the SF was the best overall performer – except for Mix onsets in the Alicante Dataset and NPP onsets in the Bello Dataset – followed closely by the CD, RCD and WPD functions, a result that confirms previously obtained results in the literature (Bello et al., 2005; Dixon, 2006). Also confirmed by previous literature (Dixon, 2006) is the PD as the function with the worst performance.

One can try to explain the overall poor results obtained with the PD detection function with the impact it suffers from the phase distortion of signals where percussive sounds have preponderance. Although this fact counts for explaining well the PD results in the PP and Mix onset types, it does not explain why the PD still performs badly for PNP onsets, where the phase distortion does not have a clear impact. A possible explanation to these poor results by the PD function is some sort of possible filter and/or normalization that the authors that proposed this method used but did not publish in their papers.

Our results show also, in accordance with Dixon (2006), that the WPD function brings great improvement on the results of the PD function. In the dataset used, for PP onset types, this improvement is of more than 30pp[1]. On the contrary, the improvements brought by the RCD function to the CD function are not as relevant and for the Mix onset class in both the datasets the RCD performs 3 to 4pp worse than the CD function.

As expected, the HFC presents very good results – staying very close to the results of the WPD, CD and RCD OSS – for Mix, PP and NPP onsets, that is, those onsets where the percussive components are predominant, lowering its performance for the PNP onsets. All in all, the HFC presents very high results in this two datasets, mainly because they are largely dominated by percussive onsets.

The functions that use information from the magnitude and phase of the spectrum perform relatively well in al onset classes except the PNP, where the results drop most of the times by 20pp. The overall good performance of this functions is in aggreement with the literature (Duxbury, Bello, Davies, & Sandler, 2003; Dixon, 2006; Bello et al., 2005), although the

[1] pp – Percentage Point

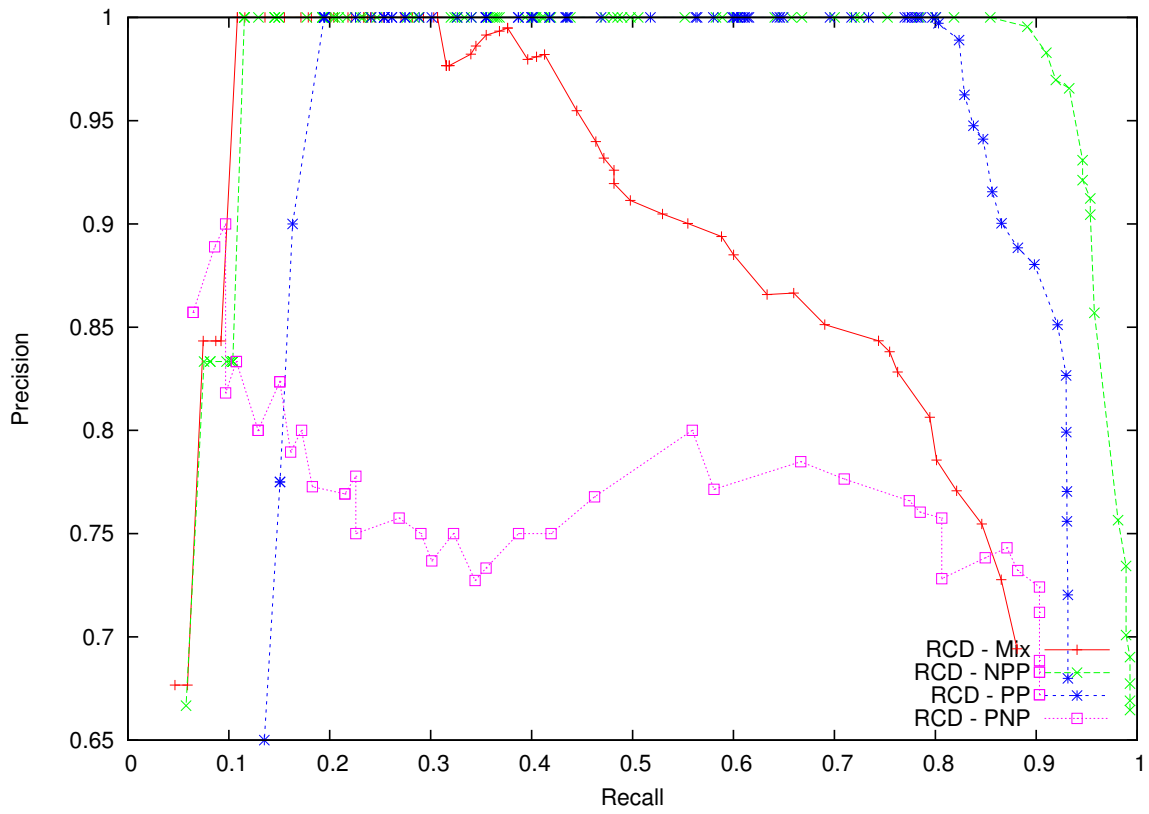

Figure 3.3: Precision vs Recall for the RCD OSS in all the onset classes of the Bello Dataset.

drop of performance for the PNP class was not something to be expected *a priori* from the literature. A possible explanation for this fact is some kind of problem with the implementation done of the algorithms proposed by the authors.

By comparing Table 3.4 with Table 3.5 and Table 3.6 with Tables 3.7 and 3.8, we can see that, in general, the results in the Bello Dataset outperform those in the Alicante Dataset. The only exceptions are the WPD, CD and RCD OSS which perform slightly better in the Alicante Dataset for the PP onset class.

A possible explanation for this result can be the fact that the Bello Dataset has more percussive onsets than the Alicante Dataset, being the percussive onsets – PP and NPP onset classes – easier to detect than the PNP and Mix onset classes.

On a song *per* song basis, the results allow us to understand that even between different songs with the same type of onsets, the results can be quite different. This fact can be used to explain the differences in performance between the two datasets, and means that the dataset

chosen can influence significantly the results.

It is not only the maximum values of F-measure for each detection function that one wants to analyse; the relation of false positives to true positives is also relevant, because it tells us information about the properties of the detection function. One can see the true positives *vs* false positives relationship by drawing a graph TP *vs* FP or by drawing a Precision *vs* Recall curve.

Figures 3.2 and 3.3 depict the precision *vs* recall curve of our results in the case of the SF and RCD detection functions for all onset classes in the Bello Dataset. The figures were obtain by varying δ in small steps.

The ideal case would be a curve touching the top-right corner (Precision = 1, Recall = 1) of a graphic of this type, that is, detecting all onsets without any false positives. That is something very hard to achieve, but one wants a curve as close to the top-right corner as possible.

Since some applications, like *tempo* estimation, require high accuracy on the detected onsets (that is, as few false positives as possible) even if missing a few onsets, and other applications require the maximization of the percentage of detections, even if increasing number of FP, the behaviour of this curve is of great interest when choosing which onset detection function to use. This clearly means that the choice of onset detection function depends on the particular application where one wants to apply the onsets obtained.

3.4 Influence of interpretation style on Onset Detection

An interesting result that was not reported yet in the literature was obtained with the previous experiments.

In Tables 3.9 and 3.10, we show the top 5 results of two violin songs from the PNP class of the Alicante Dataset: the violin song 1 is a simple song played with the most common violin techniques, while the violin song 2 is a piece composed by the Italian composer Niccolò Paganini and it requires several skilful and advanced techniques (like very fast *arpeggios*, and extreme *glissando*) in order to be performed correctly.

	F-measure	Precision	Recall
SF	0.828	0.759	0.911
SF	0.828	0.759	0.911
SF	0.825	0.769	0.888
SF	0.825	0.769	0.888
SF	0.816	0.755	0.888

Table 3.9: Top 5 results for violin song 1 - common interpretation

	F-measure	Precision	Recall
CD	0.810	0.899	0.737
CD	0.806	0.888	0.737
CD	0.804	0.878	0.742
CD	0.803	0.868	0.747
CD	0.803	0.868	0.747

Table 3.10: Top 5 results for violin song 2 - *virtuoso* interpretation

The F-measure results are similar in the two songs, although the OSS function that dominates the top results is not the same in the two cases. For the common interpretation, the SF function dominates, while for the *virtuoso* song, the CD achieves the top results. This means that the advanced violin technique produces softer onsets than the usual technique and that the CD is able to distinguish between closer onsets than the SF in this particular case. These results, however, cannot be generalized, as the dataset only has two songs of this type.

3.5 Influence of Peak Selection Methods on Onset Detection

Several papers in this area (Bello et al., 2005; Dixon, 2006; Holzapfel et al., 2010), mention that the Peak Selection part of the Onset Detection process influences greatly the final results. However, at the present time, to the best of our knowledge, there is no study that deeply explores this influence.

With this in mind, we tested the six Onset Detections developed and compared in Section 3.3 with five different Peak Selection Methods in order to test and compare the influence of four different aspects of Peak Selection:

- Threshold type;

- Normalization type;
- The use of a smoothing filter;
- The use of a local maximum for selecting peaks.

3.5.1 Experiments

In order to assess the influence of peak selection methods on the results of onset detection, different simulations were run each with a particular peak selection method. These methods were selected because they have been used in recent work (Bello et al., 2005; Dixon, 2006; Holzapfel et al., 2010).

We used the following abbreviations to name the used peak selection methods:

norm Normalize the detection function by dividing by the absolute maximum and subtracting the average value, so that the average will be zero.

stdev Normalize the detection function by dividing by the maximum standard deviation and subtracting the average value, so that the average will be zero.

mean Create a running mean threshold (Eq. 2.26).

median Create a running median threshold (Eq. 2.25).

filter Before normalization, smooth the detection function by applying a simple low pass filter (Eq. 2.19).

no-filter Do not apply the low pass filter, that is, do not use smoothing.

local-max Consider as onsets every value in the detection function that is greater than zero, greater than the threshold and is a local maximum in a window of 3 samples around it. I.e., use $w = 3$ in Eq. 2.27.

no-local-max Consider as onset every value bigger than the threshold. In other words, use $w = 0$ in Eq. 2.27.

	A	B	C	D	E
norm	×	×		×	×
stdev			×		
mean		×			
median	×		×	×	×
filter				×	
local-max	×	×	×	×	

Table 3.11: Components of the Peak Selection Methods A, B, C, D and E.

First we run our experiments with the peak selection method median-norm-no-filter-local-max(A), then we replaced the running mean threshold with a running average-threshold with parameter $M = 10$ by running the experiments with the peak selection method mean-norm-no-filter-local-max(B).

After that, in order to assess the influence of the type of normalization, we ran the experiments by replacing the norm type of normalization with the stdev type of normalization, that is, using the peak selection method median-stdev-no-filter-local-max(C).

We also tested the influence of a smoothing step before the peak selection – with the use of a simple lp-filter – by running the experiments with the median-norm-filter-local-max(D) peak selection method.

Finally, to test the final peak picking algorithm's influence, we ran the experiments without the local maximum condition, that is we used the median-norm-no-filter-no-local-max(E) peak selection method.

3.5.2 Results & Discussion

	A			B			C		
	F	P	R	F	P	R	F	P	R
HFC	0.921	0.922	0.920	0.922	0.957	0.901	0.921	0.922	0.920
SF	0.931	0.946	0.926	0.943	0.957	0.937	0.934	0.946	0.932
PD	0.652	0.573	0.819	0.650	0.571	0.819	0.652	0.573	0.819
WPD	0.916	0.959	0.882	0.922	0.933	0.918	0.914	0.945	0.891
CD	**0.947**	**0.978**	**0.923**	**0.946**	**0.987**	**0.913**	**0.943**	**0.970**	**0.923**
RCD	0.933	0.977	0.903	0.933	0.966	0.913	0.936	0.977	0.908

Table 3.12: Results for NPP onsets in the Bello Dataset using the Peak Selection methods A, B, and C: P, Precision, F, F-measure and R, Recall.

Following the results in Section 3.3, while running the experiments we fixed the window

	D			E		
	F	P	R	F	P	R
HFC	0.823	0.913	0.766	0.622	0.525	0.798
SF	**0.939**	**0.953**	**0.933**	**0.782**	**0.709**	**0.903**
PD	0.628	0.586	0.749	0.520	0.417	0.893
WPD	0.828	0.900	0.778	0.603	0.507	0.816
CD	0.872	0.931	0.835	0.583	0.482	0.820
RCD	0.909	0.919	0.904	0.419	0.298	0.824

Table 3.13: Results for NPP onsets in the Bello Dataset using the Peak Selection methods D and E: P, Precision, F, F-measure and R, Recall.

	A			B			C		
	F	P	R	F	P	R	F	P	R
HFC	0.838	0.846	0.830	0.848	0.829	0.867	0.842	0.846	0.839
SF	**0.961**	**0.968**	**0.954**	**0.965**	**0.978**	**0.953**	**0.961**	**0.969**	**0.954**
PD	0.497	0.410	0.734	0.488	0.414	0.740	0.388	0.278	0.823
WPD	0.810	0.796	0.826	0.811	0.793	0.830	0.811	0.793	0.830
CD	0.883	0.892	0.874	0.899	0.876	0.923	0.903	0.883	0.923
RCD	0.882	0.880	0.883	0.891	0.863	0.920	0.881	0.823	0.947

Table 3.14: Results for PP onsets in the Bello Dataset using the Peak Selection methods A, B, and C: P, Precision, F, F-measure and R, Recall.

size of the STFT at 1024 samples (that is 46.4ms in these 22.05kHz sampled signals) with hop size of 50%. The parameters δ and λ are tweaked while running the experiments, in order to obtain the values that maximize the f-measure.

The results obtained by running the experiments in the Bello dataset with all the peak selection methods described in the previous section are shown in Tables 3.12 through 3.19.

In order to do the comparisons, we consider as base the results with the peak selection method A and compare all others with this one.

First we will analyse the influence of the peak selection methods on the results obtained for the different onset classes and next we will analyse the influence of the peak selection methods on each of the OSS, and, finally, we will make a global balance about the significance of the compared results of the different Peak Selection Methods.

3.5.2.1 Onset Classes

The differences between running the experiments by using a running-median threshold – Peak Selection Method A – or a running-mean threshold – Peak Selection Method B – have mixed behaviours according to the onset classes. In the NPP and PP classes, the mean gives slightly

	D			E		
	F	P	R	F	P	R
HFC	0.576	0.607	0.547	0.523	0.437	0.651
SF	**0.876**	**0.878**	**0.874**	**0.893**	**0.868**	**0.921**
PD	0.529	0.323	0.823	0.368	0.256	0.732
WPD	0.470	0.545	0.414	0.666	0.641	0.692
CD	0.441	0.547	0.370	0.543	0.488	0.611
RCD	0.599	0.574	0.625	0.734	0.664	0.820

Table 3.15: Results for PP onsets in the Bello Dataset using the Peak Selection methods D and E: P, Precision, F, F-measure and R, Recall.

	A			B			C		
	F	P	R	F	P	R	F	P	R
HFC	0.553	0.519	0.591	0.552	0.519	0.591	0.553	0.519	0.591
SF	**0.911**	**0.888**	**0.935**	**0.915**	**0.858**	**0.978**	**0.914**	**0.914**	**0.914**
PD	0.615	0.479	0.860	0.615	0.479	0.860	0.615	0.479	0.860
WPD	0.660	0.602	0.731	0.670	0.626	0.720	0.670	0.626	0.720
CD	0.684	0.650	0.720	0.680	0.644	0.720	0.677	0.657	0.699
RCD	0.808	0.745	0.882	0.808	0.745	0.882	0.808	0.745	0.882

Table 3.16: Results for PNP onsets in the Bello Dataset using the Peak Selection methods A,B, and C: P, Precision, F, F-measure and R, Recall.

better results (1pp better) than the median, while it improves for certain OSS it gives worse results for others, but just 1-2pp differences for better or for worse. On the other hand, the running-mean threshold is prone to give worse results by around 2-3pp in the Mix onset class.

To use a normalization based on the maximum standard deviation – Peak Selection Method C – when comparing to a normalization based on the maximum absolute value – Peak Selection Method A – gives mixed behaviours according to the onset classes. In the NPP and PNP onset classes, the results remain almost the same (the changes are less than 1pp) while for the PP the relevant changes are a decrease of around 10pp for the PD function and a performance increase of about 3pp for the HFC and CD functions. When it comes to the Mix onset class, the results for the HFC and PD functions remain just the same, but the other OSS functions have worse f-measure (2-3pp).

When smoothing the Onset Detection Function – Peak Selection Method D – the results become quite different. For the NPP onset class, the SF becomes slightly better (less than 1pp), while for all the other OSS, the results become poorer from 3 to 10pp. In the case of PP onsets, the filter improves about 3pp on the PD function, although it decreases the results significantly (10 to 40pp) for all other OSS. In the PNP onset classes, the behaviour is mixed

	D			E		
	F	P	R	F	P	R
HFC	0.405	0.471	0.355	0.358	0.242	0.688
SF	**0.869**	**0.847**	**0.892**	**0.696**	**0.595**	**0.839**
PD	0.803	0.770	0.839	0.184	0.101	1
WPD	0.465	0.506	0.430	0.463	0.343	0.710
CD	0.388	0.444	0.344	0.409	0.295	0.667
RCD	0.503	0.500	0.505	0.562	0.425	0.828

Table 3.17: Results for PNP onsets in the Bello Dataset using the Peak Selection methods D and E: P, Precision, F, F-measure and R, Recall.

	A			B			C		
	F	P	R	F	P	R	F	P	R
HFC	0.812	0.753	0.881	0.814	0.757	0.881	0.814	0.757	0.881
SF	**0.880**	**0.922**	**0.842**	**0.867**	**0.895**	**0.841**	**0.867**	**0.889**	**0.846**
PD	0.540	0.396	0.851	0.540	0.403	0.818	0.544	0.409	0.808
WPD	0.832	0.762	0.811	0.801	0.797	0.806	0.809	0.791	0.822
CD	0.866	0.798	0.854	0.844	0.807	0.870	0.843	0.792	0.881
RCD	0.824	0.823	0.770	0.795	0.818	0.761	0.814	0.814	0.803

Table 3.18: Results for Mix onsets in the Bello Dataset using the Peak Selection methods A, B, and C: P, Precision, F, F-measure and R, Recall.

	D			E		
	F	P	R	F	P	R
HFC	0.597	0.686	0.528	0.512	0.435	0.626
SF	**0.853**	**0.844**	**0.863**	**0.679**	**0.693**	**0.665**
PD	0.491	0.373	0.718	0.458	0.337	0.713
WPD	0.587	0.630	0.564	0.557	0.625	0.505
CD	0.541	0.586	0.518	0.522	0.545	0.509
RCD	0.715	0.680	0.745	0.650	0.652	0.653

Table 3.19: Results for Mix onsets in the Bello Dataset using the Peak Selection methods D and E: P, Precision, F, F-measure and R, Recall.

according to the onset class. We have a positive boost of around 20pp for the PD OSS while for all the other functions the results get worse from 4pp to 30pp. For the Mix onset class, the results get considerably worse for all the OSS.

Finally, when dropping the local maximum condition in the peak picking algorithm – Peak Selection Method E – the results become quite different, but there is a general trend easy to spot: the results get worse for every OSS without exception. In the NPP the results are 15 to 50pp worse, while for the PP the results are 13 to 25pp worse. For PNP onsets, in general, the results are around 30pp worse while for Mix onsets the results vary from 10pp to 30pp worse.

3.5.2.2 Detection Functions

Moving from running-median threshold to running-mean threshold – Peak Selection Method B – gives, in general, slight improvements for the HFC OSS in all the onset classes, while for the SF OSS the behaviour is mixed. It improves slightly the SF in PP, NPP and PNP onset classes, while decreasing the performance in the Mix class, although these improvements and decreases are very small (1-3pp). We have similar behaviour for the WPD, CD and RCD Onset Detection Functions, with the increases and decreases not going beyond 3pp. In the case of the PD OSS, the results are quite similar for all the onset classes.

By using a normalization based on the maximum standard deviation – Peak Selection Method C – the results are not very different from the results obtained by using a normalization based on the maximum absolute value – Peak Selection Method A. In the case of the HFC, SF, and RCD, we obtain practically the same results (they change by no more than 1pp) for all the onset classes. In the case of the PD OSS, we have losses of about 10pp for the PP onset class but for the other classes the results remain basically the same (they change by less than 1pp). For the WPD and CD functions the behaviour is mixed, that is, for some onset classes the results improve while for others the results get poorer, although the magnitude of the changes in this OSS is less than 2pp, which means that the changes are not very significant. This Peak Selection Method improves the CD in the PP class, but makes its results worse in the PNP and Mix classes. On the other hand, it improves the WPD in the

PNP class, but makes it worse in the Mix class.

The use of a smoothing filter on the Onset Detection Function – Peak Selection Method D – causes the results, in general, to be much different than the results obtained with the Peak Selection Method A. For the HFC OSS, the results decrease from 10 to 25pp and for the SF the tendency is the same, except that for the NPP onset class the results improve slightly (less than 1pp) and the global losses are not so pronounced: they reach at most 9pp. In the case of the PD function we obtain mixed behaviour: for the NPP and Mix onsets the results are 2.5 and 5pp worse respectively while for the PP onsets the results improve by 3pp and for the PNP we have a 20pp improvement. The results get about 2 to 34pp and 7.5 to 44pp worse for the WPD and CD OSS respectively, while for the RCD OSS the results remain similar for NPP class, but get 9 to 30pp worse for the other onset classes. The filter has some kind of "good" effect only on the PD OSS, maybe because this kind of function is the most irregular and the filter brings some positive uniformity, and on the other OSS one obtains an excess of uniformity with the filter, decreasing the precision of the OSS.

Dropping the local maximum condition in the peak picking algorithm – Peak Selection Method E – makes, in general, the results be much worse than the results of the Peak Selection Method A. For the HFC the results are all around 30pp worse while the results can be to 20pp worse for the SF, 40pp worse for the PD and to 34pp worse for the WPD. For the complex domain family, the results can be to 40pp worse for the CD and 50pp worse for the RCD.

3.5.2.3 Balance

Having in mind the discussion of the two previous subsections, we can make a global balance. First of all, in general, the differences between the results obtained by applying a running mean and a running median threshold are not statistically significant ($W = 291,\ p = 0.959$ in the Wilcoxon signed rank sum test with continuity correction) and they are dependent upon the particular onset class and OSS, which implies that for certain applications that need just a certain type of onsets, one specific type of threshold can be chosen in favour of the other.

Concerning the normalization methods, the differences between the results obtained with the two kinds of normalization used are not statistically significant ($W = 290,\ p = 0.975$ in

the Wilcoxon signed rank sum test with continuity correction).

On the other hand, the results obtained by the usage of a smoothing filter get significantly poorer ($W = 427,\ p = 0.004$ in the Wilcoxon signed rank sum test with continuity correction) in most of the cases, except for the single case of the PD OSS. This means that one should not use a smoothing filter at all (except maybe for the single case of the PD function) or try to test a different filter from the one used in this study.

Finally, not using the local maximum condition makes the results get significantly poorer ($W = 500,\ p < 0.001$ in the Wilcoxon signed rank sum test with continuity correction), which means that one should really use the local maximum condition.

3.6 Summary

In this chapter, we introduced the Datasets used in the experiments as well as the Evaluation Metrics used to analyse the results of the experiments.

In order to test the differences between each Onset Detection Method and to see how they behave for different onset classes, a set of experiments was created and run on two distinct datasets.

The experiments' results show that, in general, the SF, WPD, CD and RCD OSS are the ones which perform best, while the PD the worst performer. It was also found that, in general, the NPP and PP are the onset classes on which Onset Detection performs best, while the PNP is the class where the results are worse.

It was also reported the interesting finding about the influence of the performance technique on onset results, that is, the *virtuoso* violin technique produces softer onsets than the common playing technique.

To test the influence of the Peak Selection step of the Onset Detection process, a set of experiments was created and run in the Bello Dataset.

In terms of Onset Classes, with the results of this last set of experiments, it was found that, in general, the differences between using a running-median or a running-mean threshold are not significant, the normalization by using the standard deviation improves the results for some onset classes, but decreases for others. It was also found that the use of a smoothing

low-pass filter and dropping the use of the local maximum condition for peak picking makes the results significantly worse.

The results of these last set of experiments also allowed us to notice the influence of the particular Peak Selection Method on each Onset Detection function. It was found that the differences between using a running mean threshold or a running median threshold are almost absent. The same happened with using a normalization based on the absolute maximum or maximum absolute standard deviation. On the other hand, the use a smoothing filter and the dropping of the local maximum condition for peak picking makes the results significantly worse for all the OSS.

4 Conclusions

To achieve the goal proposed, that is, to explore Onset Detection and improve the information about the complete process that leads from the signal to onsets, we explored two different research directions in this work.

The first approach consisted of a general comparison of the most common Onset Detection methods in order to clarify their differences and similarities. With this comparison, it is possible to assess the influence of the different onset classes on the behaviour of the several onset detection methods methods as well as understand the main differences between OSS; informations that help understand which method suits better each circumstance.

The other approach assessed the influence of the final Peak Selection method on the results of Onset Detection. This approach also allows to know which peak selection method suits best to each OSS and, as in the first approach, is able to supply more information to anyone in need of using onset detection.

The results obtained by comparing six OSS using spectral features underline some important aspects: (i) the distinct OSS perform differently, being the PD the overall worse performer and the SF, followed close by the CD function, the best overall performer; (ii) the results are significantly different between onset classes, being the NPP and PP the easier to detect and PNP the harder; (iii) even between different songs with the same type of onsets, the results can be quite different, which means that the dataset chosen can be of great influence to the results; (iv) it was also noticed that in the case of the violin, the interpretation style influences differently the results of each OSS.

With the analysis of the influence of peak selection methods on Onset Detection, other important aspects were noticed. First, the peak selection step is of major importance for the results obtained, although not all parts of the peak selection process influence equally – the

use of a filter and of a local maximum condition for peak picking have great influence, while the two normalization methods and the two thresholds tested have small influence. Next, some functions are more robust to the change of peak selection methods, that is, their results remain basically at the same level – the SF OSS was found to be the most robust –, while others can have their behaviour completely altered – the PD OSS was found to be the less robust – by the different peak selection methods.

With the information from both the approaches it is possible to choose with more confidence which onset detection method is more appropriate to use for particular applications.

As outcome of this thesis, it was also produced an open-source Audio-Framework [1] built with the Java programming language. The results of the two research directions described above were obtained by using this Audio Framework, so it contains implementations of the six OSS tested – HFC, SF, PD, WPD, CD, and RCD – as well as all of the five Peak Selection Methods tested.

The framework is made freely available to use and improve, in the hope it can help push forward the research in this area.

4.1 Future Work

In this section, we enumerate possible future work. Some of the proposals are driven by limitations of the current work, while others suggest interesting future directions based on the work reported in this thesis.

Comparison of Onset Detection Methods performance When comparing the different Onset Detection Methods, we have only studied methods that make use of Spectral Features of the signal, so this comparison can be extended by adding new methods that do not belong to the Spectral Domain. A possible way to improve, could be by adding methods that use machine learning techniques to the comparison.

Influence of instrument interpretation style on Onset Detection When it comes to the influence of the interpretation style on the results of Onset Detection, we have only explored the violin and for a very short selection of songs, so it will be interesting to increase

[1] With source-code freely available at `https://github.com/Shemahmforash/Audio-Framework`

the number of violin songs to test. Another possibility of improvement could be to compare the interpretation style influence on other instruments of different onset classes.

Influence of the Peak Selection step on Onset Detection Despite having done quite an extensive analysis of the influence of different peak selection methods on onset detection, this analysis did not focus all the existing methods, so one can extend this study by adding some more Peak Selection Methods and/or more Onset Detection methods to the study. It will also be interesting to test other kinds of smoothing filters, because just one type of filter was tested in this study.

Bibliography

Abdallah, S., & Plumbley, M. (2003). Probability as metadata: event detection in music using ICA as a conditional density model. In *Proceedings 4th Int. Symp. Independent Component Analysis and Signal Separation (ICA2003)* (pp. 233–238). Nara, Japan.

Abraham, G. (1979). *The Concise Oxford History of Music.* Oxford University Press.

Alicante, P. R. . A. I. G. U. of. (2012, January). *Onset detection database.* `http://avmediasearch.eu/wiki/index.php/Onset_Detection_Database`.

Barthet, M., Noland, K., Lahne, M., Marsh, W., & Ashworth, R. (2010). Digital signal processing tools to analyse calcium activity in vivo. In *EU Workshop on Zebrafish Neurophysiology and Behavior.* Queen Mary University of London.

Basseville, M., & Nikiforov, I. (1993). *Detection of abrupt changes: theory and application.*

Bello, J., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M. B. (2005). A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 1035–1047.

Bello, J., Duxbury, C., Davies, M., & Sandler, M. (2004, June). On the Use of Phase and Energy for Musical Onset Detection in the Complex Domain. *IEEE Signal Processing Letters*, *11*(6), 553–556.

Bello, J., & Sandler, M. (2003). Phase-based note onset detection for music signals. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. ICASSP'03. Proceedings.* (Vol. 5, p. 4952). Hong Kong, China: IEEE.

Casey, M., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008, April). Content-Based Music Information Retrieval: Current Directions and Future Challenges. *Proceedings of the IEEE*, *96*(4), 668–696.

Collins, N. (2005a). A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions. In *Proceedings of AES118 Convention* (pp. 1–12).

Collins, N. (2005b). Using a pitch detector for onset detection. In *Proceedings of ISMIR2005* (pp. 100–106).

Davies, M. E. P., & Plumbley, M. D. (2007, March). Context-Dependent Beat Tracking of Musical Audio. *IEEE Transactions on Audio, Speech and Language Processing*, *15*(3), 1009–1020.

Ding, J.-j., Tseng, C.-j., Hu, C.-m., & Hsien, T. (2011). Improved onset detection algorithm based on fractional power envelope match filter. In *Proceedings of the 19th European Signal Processing Conference (EUSIPCO 2011)* (pp. 709–713). Barcelona, Spain.

Dixon, S. (2006, September). Onset Detection Revisited. In *Proceedings of the Int. Conf. on Digital Audio Effects (DAFx-06)* (pp. 133–137).

Duxbury, C., Bello, J., Davies, M., & Sandler, M. (2003). A combined phase and amplitude based approach to onset detection for audio segmentation. In *Proceedings 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS-03)* (pp. 275–280). Singapore: World Scientific Publishing Co. Pte. Ltd.

Duxbury, C., Bello, J., Davies, M., Sandler, M., & Others. (2003). Complex domain onset detection for musical signals. In *Proceedings Digital Audio Effects Workshop (DAFx)* (pp. 6–9).

Duxbury, C., Davies, M., & Sandler, M. (2001). Extraction of transient content in musical audio using multiresolution analysis techniques. In *Digital Audio Effects Conf. (DAFX '01)* (pp. 1–4). Limerick, Ireland.

Duxbury, C., Sandler, M., & Davies, M. (2002). A hybrid approach to musical note onset detection. In *Proceedings Digital Audio Effects Conf.(DAFX,'02)* (pp. 33–38).

Ellis, D. (2007, March). Beat Tracking by Dynamic Programming. *Journal of New Music Research*, *36*(1), 51–60.

Eyben, F., Böck, S., Schuller, B., & Graves, A. (2010). Universal Onset Detection with Bidirectional Long Short-Term Memory Neural Networks. In *11th International Society for Music Information Retrieval Conference (ISMIR 2010)* (pp. 589–594).

Foote, J. (2007). Automatic audio segmentation using a measure of audio novelty. *Journal of New Music Research*, *36*(1), 51–60.

Goto, M., & Muraoka, Y. (1994). A beat tracking system for acoustic signals of music. *Proceedings of ACM*(2), 365–372.

Goto, M., & Muraoka, Y. (1996). Beat tracking based on multiple-agent architecture-a real-time beat tracking system for audio signals. In *Proceedings of the Second International Conference on Multiagent Systems* (pp. 103–110).

Hainsworth, S., & Macleod, M. (2003). Onset detection in musical audio signals. In *Proceedings of the International Computer Music Conference (ICMC)* (pp. 163–166).

Holzapfel, A., Stylianou, Y., Gedik, A. C., & Bozkurt, B. (2010, August). Three Dimensions of Pitched Instrument Onset Detection. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(6), 1517–1527.

Kauppinen, I. (2002). Methods for detecting impulsive noise in speech and audio signals. In *14th International Conference on Digital Signal Processing Proceedings. DSP 2002 (Cat. No.02TH8628)* (Vol. 2, pp. 967–970). IEEE.

Klapuri, A. (1999, March). Sound onset detection by applying psychoacoustic knowledge. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999. ICASSP'99. Proceedings.* (Vol. 6, pp. 3089–3092). IEEE.

Klapuri, A., & Davy, M. (Eds.). (2006). *Signal Processing Methods for Music Transcription.* Springer.

Klapuri, A., Eronen, A., & Astola, J. (2006). Analysis of the meter of acoustic musical signals. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(1), 342–355.

Lacoste, A., & Eck, D. (2005). Onset Detection with Artificial Neural Networks for MIREX 2005. In *Extended abstract of the 1 st Annual Music Information Retrieval Evaluation eXchange (MIREX 2005), held in conjunction with ISMIR.*

Liu, R., Griffth, N., Walker, J., & Murphy, P. (2003). Time domain note average energy based music onset detection. In *Proceedings of the Stockholm Music Acoustics Conference* (Vol. 2003, pp. 7–10). Stockholm, Sweden.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *An introduction to information retrieval.* Cambridge University Press.

Masri, P. (1996). *Computer modeling of Sound for Transformation and Synthesis of Musical Signal.* Phd thesis, University of Bristol.

Masri, P., & Bateman, A. (1996). Improved modelling of attack transients in music analysis-resynthesis. In *Proceedings of the International Computer Music Conference* (pp. 100–103). Hong Kong, China.

MIREX. (2005, August). *Mirex 2005: Audio onset detection results.* `http://www.music-ir.org/mirex/wiki/2005:Audio_Onset_Detection_Results`.

MIREX. (2010, August). *Mirex 2010: Audio onset detection results.* `http://nema.lis.illinois.edu/nema_out/mirex2010/results/aod/`.

MIREX. (2011, May). *Mirex 2011: Audio onset detection task.* `http://www.music-ir.org/mirex/wiki/2011:Audio_Onset_Detection`.

Moore, B. C. J., Glasberg, B. R., & Baer, T. (1997). A model for the prediction of thresholds, loudness, and partial loudness. *J. Audio Eng. Soc*, *45*(4), 224–240.

Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques.* Springer.

Oppenheim, A. V., Willsky, A. S., & Hamid, w. S. (1996). *Signals and Systems (2nd Edition).* Prentice Hall.

Orio, N. (2006). *Music Retrieval: A Tutorial and Review.*

Rodet, X., & Jaillet, F. (2001). Detection and modeling of fast attack transients. In *Proceedings of the International Computer Music Conference* (pp. 30–33). Cuba.

Scheirer, E. D. (1998, January). Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America*, *103*(1), 588–601.

Schloss, W. A. (1985). *On the Automatic Transcription of Percussive Music - From Acoustic Signal to High-Level Analysis.* Phd thesis, Stanford University.

Sciences, S. H. . P. (2012, January). *Speech fillin system.* `http://www.phon.ucl.ac.uk/resource/sfs/`.

Stowell, D., & Plumbley, M. (2007). Adaptive whitening for improved real-time audio onset detection. In *Proceedings of the International Computer Music Conference (ICMC'07).*

Wallin, N., Merker, B., & Brown, S. (Eds.). (2001). *The Origins of Music.* A Bradford Book.

Zhou, R., & Reiss, J. (2007). Music onset detection combining energy-based and pitch-based approaches. In *Proceedings MIREX Audio Onset Detection Contest.*

Zölzer, U., Amatriain, X., Arfib, D., Bonada, J., Poli, G. D., Dutilleux, P., et al. (2002). *DAFX:Digital Audio Effects.* Wiley.

A Glossary

This glossary presents the most relevant terminology.

Audio Novelty See OSS.

CD (Complex Domain) One can combine both energy and phase information for the production of a CD function (Duxbury, Bello, Davies, & Sandler, 2003). In this way, amplitude and energy are used together to check for irregularities in the steady state (Dixon, 2006).

CDR (Correct Detection Rate) CDR is an evaluation metric (Liu et al., 2003; Collins, 2005a) parallel to F-measure. It is used in some onset detection studies and can defined as:

$$\text{CDR} = \frac{total - FN - FP}{total} \tag{A.1}$$

Detection Function See OSS.

F-measure F-measure is a quantity that joins both Precision and Recall in a sort of weighted average. It can be obtained by:

$$\text{F-measure} = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2 \cdot P \cdot R}{P + R} \tag{A.2}$$

FN (False Negative) FN correspond to onsets that were not detected.

FP (False Positive) FP can be understood as the falsely detected onsets.

HFC (High-frequency Content) The HFC OSS explores the fact that an energy increase in one or more frequency bands can be a simple indicator of an onset (Dixon, 2006) and that in the high frequencies the interference with other simultaneous components is smaller (Dixon, 2006). It weights each STFT bin by a factor proportional to its frequency.

Low-pass filter See LP-filter.

LP-filter (Low-pass filter) A Low-pass filter, in general, when applied to a function, selects low frequencies up to the cut-off frequency(f_c) and attenuate frequencies higher than f_c (Zölzer et al., 2002) and can be defined as:

$$y_i = \alpha x_i + (1 - \alpha) y_{i-1} \tag{A.3}$$

MIR (Music Information Retrieval) MIR is a recent and emerging research area devoted to fulfil the listeners' music information needs (Orio, 2006). That is, it aims at finding and retrieving relevant information to humans from the musical signal.

Mix (Complex Mixture) One can think of Complex Mixture onset class as containing any polyphonic music where several instruments are played together, something that happens, for instance, in a rock or pop song.

NPP (Non-pitched Percussive) The NPP onset class corresponds to onsets typically produced by percussion instruments such as drums or cymbals.

Onset An onset can be understood as the starting moment of a musical note (Dixon, 2006). For a precise definition see Section 2.1.2.

Onset Detection Function See OSS.

OSS (Onset Strength Signal) An OSS is a function resulting from a process, sometimes called Reduction, that transforms the original signal into a more simplified function which easily expresses the transients (Bello et al., 2005).

PD (Phase Deviation) The PD OSS explores the fact that is unlikely that the frequency components of the new sound are in phase with the previous sound, that is, it uses the fact that irregularities in the phase of several frequency bins can be used to indicate the presence of onsets (Dixon, 2006).

Peak Selection The act of choosing from the OSS the peaks that correspond to onsets. Usually it consists of three steps: Post-processing, Thresholding and Peak-picking.

Pitch Pitch is a perceived sound property closely related to frequency that allows one to organize sounds in a frequency-related scale, i.e., with notions of "higher" and "lower" in the common way associated with melody (Klapuri & Davy, 2006).

PNP (Pitched Non-percussive) The PNP onset class corresponds to onsets that do not have percussive characteristics and have a very well defined pitch; this class contains onsets from instruments such as bowed strings or wind instruments.

Polyphony Polyphony refers to sounds or musical pieces where several acoustic sources are simultaneously present.

PP (Pitched Percussive) The PP onset class corresponds to onsets that have a percussive characteristic but, nonetheless, still maintain a well defined pitch.

pp (Percentage Point) A percentage point is the unit for the arithmetic difference of two percentages.

Precision Precision, also called accuracy, is the fraction of retrieved instances that are relevant (Olson & Delen, 2008). It can be defined by dividing the total correctly detected positives by the total number of returned results.

$$Precision = \frac{TP}{TP + FP} \tag{A.4}$$

RCD (Rectified Complex Domain) Dixon (2006) noted that the CD method does not distinguish well between onsets and offsets, so in 2006 he proposed a RCD function to surpass this problem. This method uses a half-wave rectification (similar to the SF method) in order to preserve only the positive variations of energy in the spectral bins.

Recall Recall, also called true positive rate or hit rate, is the fraction of relevant instances that are retrieved. It is obtained by dividing the correctly detected positives by the number of results that should have been returned (Olson & Delen, 2008).

$$Recal = \frac{TP}{TP + FN} \tag{A.5}$$

Reduction It is a step in the typical Onset Detection methodology that consists in the computation of a detection function, i.e., a function whose peaks should be simultaneous, within a tolerance margin, with onset times (Dixon, 2006).

Sample In Signal Processing, a sample refers to the smallest unit in the discrete signal when coming from the continuous signal. In other words, a sample is the value of the discrete signal at a point in time (Oppenheim et al., 1996).

Sample Rate The sample rate (also called sampling rate or sampling frequency) defines the number of samples per unit of time (usually seconds) taken from a continuous signal to make a discrete signal (Oppenheim et al., 1996).

Sampling In Signal Processing, sampling refers to the process of reduction from a continuous signal to a discrete signal (Oppenheim et al., 1996).

SD Spectral Deviation. See SF.

SF (Spectral Flux) The SF OSS measures the change in magnitude in each frequency bin (Dixon, 2006) and is calculated by computing the difference of two consecutive short-time spectra bin by bin (Eyben et al., 2010)

Spectral-domain Representation of a function with respect to frequency. Usually a magnitude vs frequency representation.

STFT (Short-time Fourier Transform) A STFT is a mathematical function that expresses a function of time in a function of frequency, that is, it transforms from the Time-domain representation to the Spectral-domain representation of a function. For a general signal $x(n)$ and the n^{th} bin of the k^{th} frequency, it can be defined as:

$$X_k(n) = \sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1} [x(nh+m)]^2 w(m) e^{-\frac{-2i\pi mk}{N}} \tag{A.6}$$

where $w(m)$ represents an N-point rectangular window and h the hop-size, i.e., the time-shift between adjacent windows.

Tempo Tempo (Italian word for time) is the speed or pace of a given musical piece.

Threshold A function created from the OSS – typically using the mean or median – that separates the event-related peaks from the non-event-related peaks.

Time-domain Representation of a function with respect to time. Usually a magnitude vs time representation.

TP (True Positive) TP can be interpreted as the correctly detected onsets.

Transient The transient can be understood as a short-time interval in which a significant energy change occurs in the signal (Bello et al., 2005; Klapuri & Davy, 2006).

WPD (Weighted Phase Deviation) The PD OSS considers all frequencies equally, so Dixon (2006) proposed weighting the frequency bins by their magnitude – the WPD – so it can account for "noise introduced by components with no significant energy" (Bello et al., 2005).