Towards Interactive Multimodal Music Transcription

José Javier Valero Mas

Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior

Towards Interactive Multimodal Music Transcription

José Javier Valero Mas

Tesis presentada para aspirar al grado de

DOCTOR POR LA UNIVERSIDAD DE ALICANTE

MENCIÓN DE DOCTOR INTERNACIONAL

DOCTORADO EN INFORMÁTICA

Dirigida por

Dr. José Manuel Iñesta Quereda

Esta tesis ha sido financiada por el Programa Propio del Vicerrectorado de Investigación, Desarrollo e Innovación para el Fomento de la I+D+i en la Universidad de Alicante a través del programa de ayudas para la Formación del Profesorado Universitario (ref. UAFPU2014–5883).

A mis amigos y compañeros, por los momentos vividos
A mi familia, por vuestro apoyo
A María, Ana y Noa, por las risas y la ilusión
A Jose y Simo, por todo

Acknowledgements

Surprisingly, writing this part of the Thesis turns out to be considerably more difficult than what I expected. I guess you constantly imagine yourself producing this part of the manuscript as the pinnacle of the work, but I did not expect these mixed feeling of happiness and nervousness.

In the first place I would like to thank José Manuel Iñesta for all the time spent in me. It has not only been his supervision or guidance in the scientific part but also his efforts towards creating a good environment which made possible my personal and professional development during these years. Many thanks for that, for the good times during our teaching duties, for the endless discussions about music, for giving me the opportunity to develop this work, and for discovering long time ago me this world of technology and music.

When I first started in the Pattern Recognition and Artificial Intelligence Group from the University of Alicante I hardly knew anybody. Nonetheless, now I feel this place as a second home, and this is not beacuse of the many hours spent here but because of the people I met along the way: Antonio Pertusa, Pierre, David Rizo, Carlos Pérez, Luisa Micó, José Oncina, José Bernabeu, Javi Sober, Javi Gallego, and Juan Ramón Rico. I would like to make a separate mention to my office partner Jorge Calvo (and also his wife Bea) for bearing me for so many days, cheering me up in the difficult moments, and spending those many *prog rock* evenings, among much other stuff. Also, I would also like to thank Lea Canales and Daniel Torregrosa, my partners in the PhD program, for all those relief moments and your help with all the logistic stuff. I feel really lucky of having experienced all those moments with all of you.

I am greatly thankful to everybody in the Centre for Digital Music (C4DM) in the Queen Mary University of London for having received me for not only once but twice and having made my stays in London so amusing. Apart from the nice and cool *music atmosphere* that you have created there, the actual important part is how grateful all of you are. A great piece of this work is because of the good times I had the chance to spend there with all of you: Pablo, Elio, Dave(s), Yading, Maria, Julien, Veronica, Siying, Fran, Giulio, Sid, Carlos, Antonella, Katerina, Adrien, Madeleine, Lena, Yvonne, Rodrigo... Besides, I would like to specially thank Emmanouil Benetos for having invited me to be part of that family and for having shared his time and knowledge with me.

Also thanks to Francesc J. Ferri for having invited me to the Universitat de València and to Irene Martín and Max Cobos for having made my days there interesting enough to talk about them.

I would also like to thank all my friends from outside the research and music environment. Lucky me, I would need a lot of space for naming all of you one by one, and I know I would eventually forget somebody. My deepest gratitude to all of you since you are responsible of much of the work here. Maybe not in code lines, written sentences or experiments, but this would have not been possible without all of you.

And last but not least to my family, and specially my parents Jose and Simo, for their endless patience and love. I cannot actually express, not even in my mother tongue, how thankful I am to you for supporting me all these years and all your efforts and sacrifices.

For sure, but not intentionally, I am forgetting many people here. Nevertheless I know you will forgive me since, in the end, this is simply either a sheet of paper or a bunch of bytes altogether.

Moltes gràcies a tots!

José Javier Valero Mas
Elche, a 25 mayo de 2017

The work in this dissertation has been carried out at the Pattern Recognition and Artificial Intelligence Group of the University of Alicante (Spain) between November 2013 and May 2017, with research stays from May 2015 to July 2015 and October 2016 to November 2016 at the Centre for Digital Music of the Queen Mary University of London (UK) and September 2015 as well as July 2016 at the Department of Computer Science of the Universitat de València (Spain). This work has been supported by the Vicerrectorado de Investigación, Desarrollo e Innovación para el Fomento de la I+D+i de la Universidad de Alicante through Formación del Profesorado Universitario (FPU) program (ref. UAFPU2014–5883), the Spanish Ministerio de Economía y Competitividad through Project TIMuL (No. TIN2013-48152-C2-1-R supported by EU FEDER funds), and Consejería de Educación de la Comunitat Valenciana through project PROMETEO/2012/017.

Preface

'[...] who was six years old on 6/6/66 [...]'

With such particular sentence Frank Zappa's guitar solos transcription book titled *The Frank Zappa Guitar Book* began (Zappa & Vai, 1982). Nevertheless, this quotation was not about Frank as it related to the other author of the book, who found really curious this unusual fact historically related to the Devil or the Antichrist. This person was an unknown musician at the time, but he was (and still is) one of most innovative electric guitar players ever. His name was Steven Siro Vai and, despite his concerns about the *evilness* of the birth date, he later claimed they all disappeared once he met Marilyn Manson.

Steve Vai grew up in Long Island (New York) and began playing guitar at the age of 13. His first guitar teacher, also an unknown musician at the time, was only three years older than him but with a great reputation in the area. His name was Joe Satriani and he is nowadays considered one of the main references in electric guitar playing. It is told that, the day of the first class, Steve went to Joe's house (where he would receive the lessons) carrying an acoustic guitar without any strings on it. Nevertheless, as lessons kept on going, Steve became impressed with the possibilities of that instrument. Fallen in love with it, Steve began to practise as much as he could, which eventually led to a great development of both a great technical ability and remarkable musical skills.

As a teenager Steve became literally obsessed with the music of Frank Zappa, up to the point of being totally determined to become a member of his band. At the age of twenty, while attending to the prestigious Berklee School of Music in Boston (Massachussets), Steve sent Frank some material he thought the well-known musician would be interested in. Apart from several tapes showing off his very skilled and mature sense of guitar playing, there was a transcription of one of Zappa's most challenging pieces called *The Black Page*. This piece, composed for the drum kit and originally performed by the renowned drummer Terry Bozzio, receives its name due to the large amount of notes, ornamentations and annotations present in it, which makes the score resemble a black page. Frank, impressed with the talent and abilities of that unknown musician, immediately hired him.

Steve left Berklee and started his career in Frank Zappa's band. Initially, most of his duties consisted in transcribing music, typically guitar solos and drum sections. After some time he became a full-time band member, often playing *impossible* guitar parts credited as *Strat Abuse*. Eventually, some years later he left the band to pursue his own musical career, becoming one of the most acclaimed electric guitar players nowadays.

And, how does this story relate to this work? In general, distinguishing the different instruments, notes, rhythms, chords, tonalities is not a trivial task which certainly requires a great deal of practise. On top of that, people with such skills who also have the necessary expertise to represent this information in an abstract musical format is definitely scarce. Nevertheless, symbolic music representations and codifications of the information present in audio streams are undoubtedly useful for tasks such as preservation, reproducibility, or musicological analysis, among many others.

This dissertation focuses on this issue of retrieving a symbolic high-level representation which abstracts the musical information present in an audio piece using computational approaches. This process is known as Automatic Music Transcription in the Music Information Retrieval community in which it shows large application, not only as a user-end application but also as an intermediate process for other tasks.

Nevertheless, due to its high complexity, this problem is still far from being solved, reason why Frank would still probably require Steve for transcribing some of his improvisations. However, small research contributions as this work should eventually provide more accurate and reliable techniques not only applicable in the research community but also on a daily basis.

Contents

List of Figures

List of Tables

Acronyms

AMT	Automatic Music Transcription
CHC	Cross-generational elitist selection, Heterogeneous recombination and Cataclysmic mutation
CNN	Condensed Nearest Neighbor
DROP3	Decremental Reduction Optimization Procedure 3
ECNN	Edited Condensed Nearest Neighbor
EFCNN	Edited Fast Condensed Nearest Neighbor
ENN	Edited Nearest Neighbor
EM	Expectation-Maximization
F0	Fundamental frequency
FaN	Farthest Neighbor
FCNN	Fast Condensed Nearest Neighbor
FP	False Positive
FN	False Negative
FST	Finite State Transducer
HMM	Hidden Markov Model
ICF	Iterative Case Filtering
IPR	Interactive Pattern Recognition
ISPR	Interactive Sequential Pattern Recognition
kNN	k-Nearest Neighbor
MIDI	Musical Instrument Digital Interface
MIR	Music Information Retrieval

MIReS	Music Information ReSearch
MIREX	Music Information Retrieval Evaluation eXchange
MOP	Multi-objective Optimization Problem
MPE	Multi-pitch Estimation
NCL	Neighborhood Cleaning Rule
NE	Nearest to Enemy
NMF	Non-negative Matrix Factorisation
NT	Note Tracking
ODF	Onset Detection Function
OSF	Onset Selection Function
OR	Overlap Ratio
PG	Prototype Generation
PLCA	Probabilistic Latent Component Analysis
PR	Pattern Recognition
PS	Prototype Selection
RCNN	Repeated Condensed Nearest Neighbor
RENN	Repeated Edited Nearest Neighbor
RNN	Recurrent Neural Network
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
TL	Tomek Links
TP	True Positive

Chapter 1

Introduction

"Music in general is looking for something new overall"

Leslie Edward "Les" Claypool

In general, art is associated with the human desire of communication and expression. Throughout history different artistic disciplines such as painting, writing, or dancing allowed, and still do, the expression of feelings, concerns, or different visions of life. Among them, music constitutes a particular discipline whose main means of communication is sound together with its absence, the silence.

Music has always played a key cultural role across all civilizations and human time periods. Thus, the evolution of our society has always entailed a clear development of this form of artistic expression: from the initial rudimentary expressions of our prehistoric ancestors, more elaborated musical movements such as the complex suites from the Western common practice, the epic rock anthems from the modern era, the eastern folklore pieces or the fusion styles such as bossa nova have addressed the different musical concerns and interests. In this regard, music has always been considered as a fruitful field of study.

Music has been studied through a considerable number of scientific points of view: philosophy (most commonly, through aesthetics), physics, psychology, mathematics and so on. However, during the second half of the past century, with the accessibility and dissemination of computational methods, a new perspective of study emerged: the computational one.

1.1 Music Information Retrieval (MIR)

Since originally coined by Kassler (1966), the term Music Information Retrieval (MIR) and its scope have been thoroughly studied and analyzed by the scientific community. Throughout the years, many authors have devoted a considerable amount of effort in defining it and also clearly describing the aims, boundaries, and implications of this research area.

Among the possible definitions in literature, we may find two representative examples. A rather concise description is proposed by Orio (2006) who defines this field as "*[a] research area devoted to fulfill users' music information need*". More recently, the Music Information ReSearch (MIReS) Consortium describes it in a more technical style as "*a field that covers all the research topics involved in the understanding and modeling of music and that use information processing methodologies*" (Serra et al., 2013).

In addition to these definitions, it is important to consider the *"multicultural, multiexperiential, and multidisciplinary aspects of music"* present in this research field, as denoted by Downie (2003). These elements clearly evince the challenging difficulties when tackling MIR research and proves the non-triviality of achieving new developments in the field.

Given the above, we may define MIR in our own words as the field that aims at extracting and retrieving information from music data considering different scientific perspectives such as engineering, psychology, mathematics, or physics, among others.

In spite of the relatively novelty of this field, a large number of specific research areas has emerged under the scope of MIR. Recently, Schedl, Gómez, and Urbano (2014) grouped the majority of them into a set of four main tasks: *(i) feature extraction*, being music transcription, key estimation or structural analysis some representative subtasks; *(ii) similarity*, in which query by humming/tapping or cover song detection may be located; *(iii) classification*, which deals with issues such as mood recognition, composer identification or audio tagging; and *(iv) applications*, in which tasks such as audio fingerprinting, playlist generation and music recommendation are included.

Considering the previously exposed taxonomy, this dissertation falls within the *feature extraction* group as it focuses on the particular task of Automatic Music Transcription (AMT). The following section shall properly define this research area as well as its underlying problems and difficulties.

1.2 Automatic Music Transcription (AMT)

Among the music literature, the definition of the term *transcription* shows a clear ambiguity (Herbert, 2009). According to Randel (1944), transcription is defined in the same terms as arrangement: *"adaptation of a composition* *for instruments other than those for which it was originally written (thus, in a way, the musical counterpart of a literary translation)"*.

However, the actual definition which suits this work is the one given by Gallagher (2009), who defines transcription as *"a representation of a musical performance in standard music notation or tablature"*. Thus, for the rest of the dissertation, music transcription must be understood as the annotation of the performance of a music piece in some type of symbolic music notation.

This particular duty is unarguably useful for any task related to the musicological analysis of a piece (Nettl, 2015), reason why musicians devote a considerable amount of effort to training and developing such skills. Notwithstanding, it should be noted that this process inherently implies a certain degree of ambiguity: as studied by List (1974), transcriptions on the same performance made by different experts showed some slight differences, although all of them perfectly codified the piece. This uncertainty, although somehow implicit to any field related to music, should be taken into account when tackling transcription and annotation tasks.

In our computational equivalent, Automatic Music Transcription (AMT) may be directly defined as the automated version of the aforementioned task. Nevertheless, given that this process has been largely addressed by the MIR community, we may find a number of definitions for it in the literature. For instance, Bello (2003) describes it as follows:

> *"convert a musical recording or performance into a musical score"*

Klapuri (2004b) defines it in a similar sense but not restricting the outcome to a musical score:

> *"transforming an acoustic signal into a symbolic representation"*

Cemgil (2004) focuses on the need for obtaining a high-level representation capable of being understood by the user:

> *"extraction of a human readable and interpretable description from a recording of a music performance"*

Similarly, Pertusa (2010) also emphasises the need for the abstraction to be understood by the user:

> *"extract a human readable and interpretable representation, like a musical score, from an audio signal"*

As a last example, Benetos (2012) implicitly remarks the need for producing a human-readable encoding as the aim of the task:

> *"process of converting an audio recording into a symbolic representation using some form of musical notation"*

Given these representative definitions in the literature, AMT may be considered, in our own words, as the process of obtaining a high-level abstraction of the music content in an audio piece using information retrieval techniques. To our understanding, this abstraction must be computable in order to allow other MIR fields to take advantage of that information, which may not be necessarily human-readable. Besides, this encoding should also allow the *translation* of that piece of information to any kind music notation. A possible existing format which fulfills the requirements exposed could be the one proposed by the Music Encoding Initiative (MEI).

However, currently the most common representation for the result of practical AMT systems is the piano roll. This representation is basically a two-dimensional graph in which the abscissa axis represents the time evolution of the pieces and the ordinate axis encodes the pitch content of the piece, most commonly as discrete notes events. Thus, each coordinate of the graph shows which note events are active or inactive for each time stamp of the piece. Figure 1.1 shows an example of such representation.

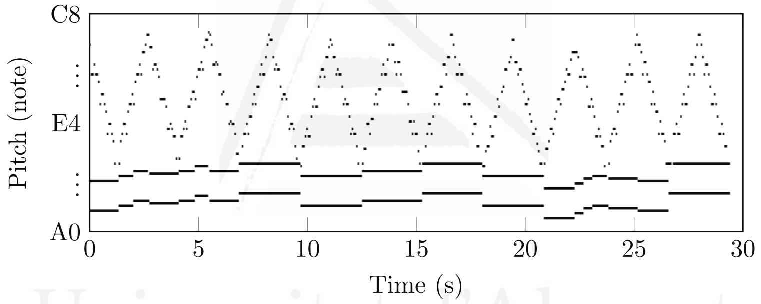


Figure 1.1: Example of a piano-roll representation for Automatic Music Transcription: time and pitch activation are represented by the abscissa and ordinate axes, respectively.

In relation to this representation issue, it is finally important to mention the Musical Instrument Digital Interface (MIDI) standard. In spite of being a communication protocol for the interconnection of *musical* elements as, for example, synthesizers, effects units, or sampling devices, the messages of this protocol can be encapsulated into files that basically encode note events: pairs of messages in which one is used for starting a note at some point and the other one for ending it at some other. In this sense, the music content of a MIDI file is basically a piano roll representation.

The usefulness of AMT is significant in music, up to the point of having been considered *"the Holy Grail in the field of music analysis"* (Benetos, Dixon, Giannoulis, Kirchhoff, & Klapuri, 2012). On the one hand, for tasks such as music preservation through (digital) scores, the abstraction resulting from the AMT process constitutes a goal by itself; on the other hand, for tasks more related to the MIR discipline (e.g., music search, similarity, and retrieval), interactive music systems (e.g., score following) or computational musicological analysis, AMT actually constitutes an intermediate process from which the resulting abstraction is further processed (Klapuri & Davy, 2007). Figure 1.2 graphically summarizes some of these applications.

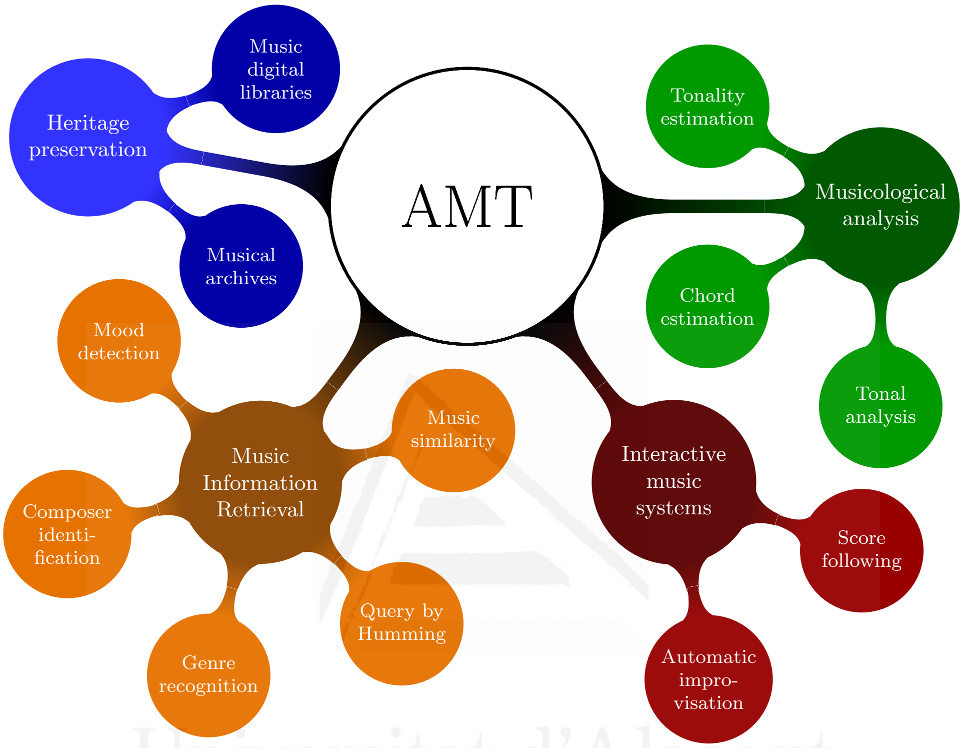


Figure 1.2: Mindmap of representative applications of Automatic Music Transcription.

The majority of AMT systems comprise two stages (Benetos et al., 2012): an initial step called Multi-pitch Estimation (MPE), typically considered the core part of AMT, in which the system estimates the actives pitches present in the signal; and a second stage known as Note Tracking (NT) which processes the result of the MPE in terms of a discrete pitch value, note starting time (onset), and note ending time (offset). Thus, while the former stage aims at retrieving a *raw* pitch description of the signal, the latter acts as both a correction and segmentation stage for obtaining musically-meaningful representations.

The main problematic for MPE methods lies in the polyphony degree of audio piece at issue (Grosche, Schuller, Müller, & Rigoll, 2012). While pitch estimation in monophonic pieces has been largely addressed in the literature, to the point of being considered a solved task by some authors, it still remains an open question for polyphonic audio pieces (Klapuri, 2004b; Argenti, Nesi, & Pantaleo, 2011). In contrast, NT has not received the same degree of attention (Duan & Temperley, 2014), possibly due to its intrinsic dependency with the MPE stage.

However, as reported in the literature, a glass ceiling may have been reached with these classic methodologies (Benetos et al., 2012): apart from the fact that results seem to have stalled with slight improvements, most approaches seem to be quite suited for certain types of data, hence lacking the flexibility expected from such systems, or the fact that very scarce examples of AMT systems succeed in obtaining user-readable scores, which is one of its main purposes. Thus, it seems that a need for a paradigm shift is required (Benetos, Dixon, et al., 2013).

In this regard, some authors have started incorporating additional processes to the aforementioned classic scheme. In most cases, these procedures are directly other MIR tasks which provide additional descriptions of the signal at issue as, for instance, information about harmony, rhythm, sound sources, or instrumentation, among others. These pieces of information impose musical constraints to the MPE and NT stages which narrow the *search space*, imitating the human approach to transcription which significantly relies on prior knowledge and complementary descriptions (multimodal sources of information) of the piece at issue.

Nevertheless, even with the use of additional processes for AMT, one of the main issue lies in the fact that none of those components exhibits an error-free performance, thus being an external user required to manually inspect the errors committed by the system and correct them. Hence, due to the fact that the implication of an external user cannot be neglected for the actual success of the task, some works start to consider users as an active elements within the transcription process. Benetos, Dixon, et al. (2013) summarized these ideas of interaction and multimodality in a conceptual AMT scheme reproduced in Figure 1.3.

Attending to this proposal, the core of the AMT system still lies in the MPE and NT processes. Nonetheless, these process are now complemented with additional descriptions such as information about onset/offset events, rhythm estimation, or harmonic analysis provided to improve the performance of the core of the system. Also, prior musical information such as computational models trained on formal music principles, organology studies or particularities of the genre of the piece at issue is considered for such aim. On top of that, this information may be directly estimated from the signal with computational models but, as aforementioned, approaches considering the user are currently being considered as an alternative to stand-alone MIR methods.

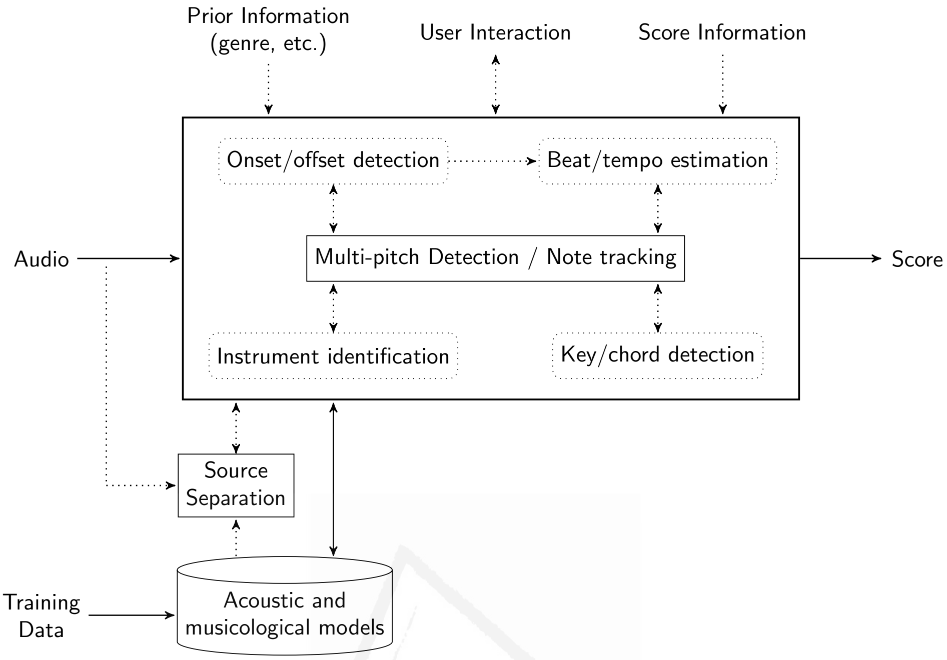


Figure 1.3: General Automatic Music Transcription scheme by Benetos, Dixon, et al. (2013). Dotted lines represent optional subtasks while double arrows point out information fusion or interaction between the subsystems.

1.3 Motivation and aim

The starting point of this dissertation stands in the aforementioned concepts of interactivity and multimodality applied to AMT. The use of different sources of information aims at resembling the manner human experts act towards this transcription issue: MIR processes such as chord extraction, tempo estimation, onset/offset detection or instrument separation may be useful for the success of the task (Benetos, Dixon, et al., 2013).

However, no existing MIR can be considered as error-free, reason why an external agent is required for correcting the information estimated by these additional tasks so that it can be reliably used by the AMT engine. Avoiding the discussing of whether these system shall ever be totally autonomous, nowadays there is a need for a user to be part of the system.

Taken for granted this need for a human agent to be part of the system, it seems interesting to explore interactive modalities that allow to include the user as an active part of the system rather than as an external correction agent (Toselli, Vidal, & Casacuberta, 2011). Within these interactive methodologies, the success of the task can be guaranteed, at least up to the expertise of the user, whilst the main and challenge lies in efficiently exploiting and reducing the user effort towards the completion of the task.

In our case we focus on the idea of interaction with onset events of the audio signal and its application, as a particular multimodal source of information, to the NT stage of the AMT process. Onset events constitute an important source of information for the temporal segmentation and the rhythmic description of the signal, and have proved to be particularly helpful in the commented NT stage (Grosche et al., 2012). However, as onset detection systems still do not exhibit flawless performance, an efficient practice to improve this performance gap may be tackling it with human intervention.

This human intervention task may be approached from a number of different points of view, being a considerable amount of the work in this dissertation developed from a Pattern Recognition (PR) perspective. Hence, some of the studies are tackled from a general viewpoint whose contributions and conclusions are applicable to both MIR and PR fields.

1.4 Thesis structure

The rest of the this manuscript is structured as follows:

Chapter 2: Background in Music Information Retrieval. Provides the Music Information Retrieval concepts necessary for the rest of the dissertation: an introduction and revision to existing Automatic Music Transcription approaches and onset detection methods, as well as their evaluation methodologies.

Chapter 3: Pattern Recognition in Music Information Retrieval. Introduces the fundamentals of Pattern Recognition and classification required for the understanding of this work as well as its application in the field of Music Information Retrieval. Special emphasis is done on the k-Nearest Neighbor classifier due to its large application along the dissertation.

Chapter 4: Studies on Prototype Selection for the k-Nearest Neighbor classifier. Presents the studies carried out for the k-Nearest Neighbor classifier in terms of Prototype Selection and imbalanced data.

Chapter 5: Approaches for Interactive Onset Detection and Correction. Exposes the works related to the interactive methodologies developed in this dissertation for the particular task of onset detection. This includes the definition of a novel set of metrics for the quantitative evaluation of these tasks as well as a set of particular techniques proposed for the interactive detection/correction paradigm.

Chapter 6: On the use of Onset Information for Note Tracking. Presents the research work focusing on the use of multimodal information for improving transcription. More precisely, this chapter presents two studies focusing on the use of onset information for correcting the transcription obtained by Multi-pitch Estimation systems as a post-processing stage.

Chapter 7: Conclusions and future perspectives. Summarizes the contributions of this dissertation and studies future perspectives that could be addressed from this work.

CHAPTER 2

Background in Music Information Retrieval

"A lot of music is mathematics.
It's balance"

MELVIN KAMINSKY "MEL BROOKS"

This chapter provides the necessary MIR basis on which the rest of the work is grounded: first, an in-depth discussion of AMT systems including interactive and multimodal strategies is included; then, the task of Onset Detection is introduced and reviewed; after that, a section is devoted to the presentation of the evaluation methodologies of the aforementioned tasks; finally, the last section presents a general discussion of the topics discussed in the chapter.

2.1 Review on AMT systems

As commented in Chapter 1, the core of practical AMT systems comprises two phases, the Multi-pitch Estimation (MPE) stage and the Note Tracking (NT) step. Although the aim of each stage is different, both compute and produce an abstraction of the input signal: MPE obtains a mid-level abstraction know as *frame-level* transcription[1] that indicates the active pitches present at each analysis frame of the input signal; NT obtains a high-level abstraction known as *note-level* transcription that describes the events in the signal in

[1] It might be arguable to consider such frame-by-frame analyis as transcription rather than detection as it does not retrieve any high-level symbolic equivalent of the audio content. Nevertheless, we shall restrict to this definition as it is commonly considered in the AMT field.

terms of a discrete pitch value, onset, and offset (Cheng, Dixon, & Mauch, 2015). Figure 2.1 shows a graphical example of such scheme.

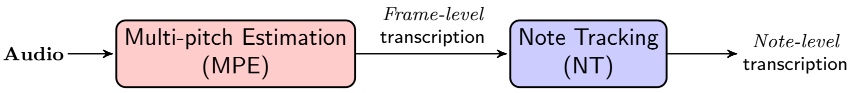


Figure 2.1: Diagram the core tasks in an Automatic Music Transcription system.

Given the relevance of these stages in AMT systems, the following sections are devoted to its explanation and literature revision. Besides, two additional sections are dedicated to the issues of interactivity and multimodality applied to AMT.

2.1.1 Multi-pitch Estimation (MPE)

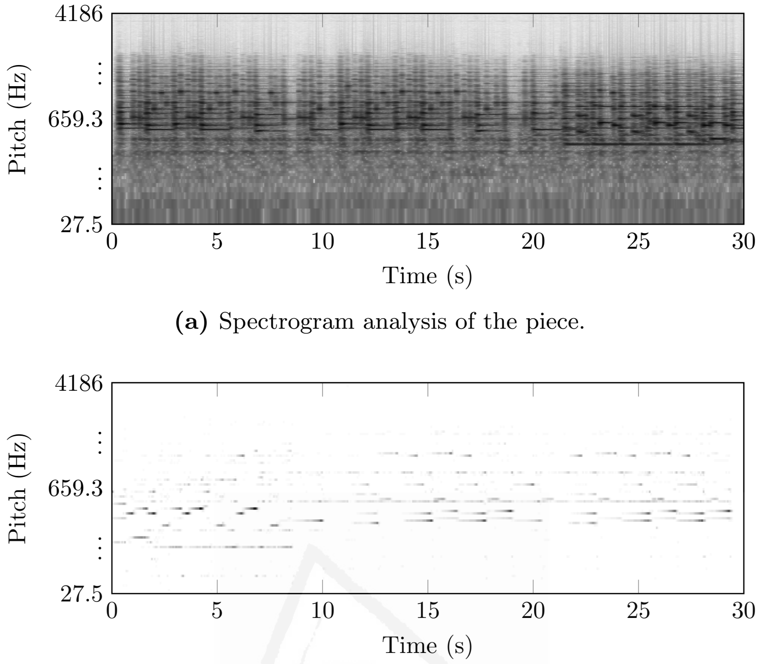
The aim of the MPE process is describing the input signal in terms of the fundamental frequencies (F0s) or pitches[2] present at each analysis frame considered. The output of such systems is commonly referred to as *pitch activation matrix* or, in cases involving probabilistic frameworks, *posteriogram*. Figure 2.2 shows an example of such analysis applied to a piece of piano music.

While this task is considered to be practically solved for cases in which monophonic music is considered, for the case of polyphonic data it is still far from being solved since the number of simultaneous pitches is, a priori, unknown. Furthermore, this task still gets more complicated when tackling polytimbral pieces since the harmonic and spectral structures of the simultaneous pitch values from different instruments may not be the same, thus making it difficult to make any assumption about the possible spectral structure of the mixture.

Given the above, MPE systems are usually complex methods that combine several processing principles. Thus, as reported by Klapuri (2004a), it is not trivial to propose a single taxonomy which properly classifies the different existing methods into isolated categories.

A classical differentiation found in works such as Brossier (2006) or Yeh (2008) classifies MPE methods depending on whether the signal is processed in the time domain or in an alternative one (typically frequency domain, but also other transformed domains such as the Wavelets, Mel filter-banks or the Constant Q transform may be considered). A third category based on combinations of both principles is also considered in some cases (de Cheveigné,

[2] Although *fundamental frequency* and *pitch* refer to the same physical concept, the latter term implies some psychoacoustic connotations of human perception. Nevertheless, as in other similar AMT works, we are obviating these perceptual nuances to use both terms indifferently.



(a) Spectrogram analysis of the piece.

(b) Pitch activation matrix from an multi-pitch analysis of the piece.

Figure 2.2: Example of multi-pitch analysis applied to a piece of piano music.

2006). Nevertheless, the main issue with this taxonomy is that, given the aforementioned complexity of MPE, methods rarely rely on time-domain representation.

Yeh (2008) classifies MPE systems as either an *iterative* or a *joint* pitch estimation: iterative strategies estimate a single pitch value and then cancel its residual harmonic structure before starting the process again, until a converge criterion is reached; joint estimation methods check different pitch combinations and hypotheses, without performing any cancellation, until a solution is achieved. Iterative methods usually provide computationally efficient solutions, but joint strategies often yield more accurate results. Nowadays most MPE techniques fall on the joint category due to the improvements in terms of computation.

Finally, other authors such as Klapuri (2004a) or Pertusa (2010) classify these systems in terms of their core processing technique. In this work we consider the taxonomy proposed by Benetos, Dixon, et al. (2013) that classifies MPE techniques into three categories. Note that, even with such categorisation, some methods may not exclusively belong to one of the families due to the aforementioned issue. We shall now introduce these three categories and provide brief review of the most relevant methods for each of them:

Feature-based methods

This first category comprises the methods which perform the MPE process relying on a set of features typically obtained using signal processing techniques without the further consideration or use of any specific model. Early examples of AMT systems belong to this category as, for instance, the work by Moorer (1977) in which comb filters were used to perform pitch tracking in the frequency domain, but limited to musical duets of monophonic instruments.

A major limitation in MPE systems is the assumption that pitch components are totally harmonic while in practise this may not be always true. In this regard, Klapuri (2003) proposed a system which does not assume ideal harmonicity: the spectrum is divided in 18 single bands and pitch is tracked individually in each band; then, a weighting function combines the decisions of the different bands and a general pitch value is obtained; this process is applied iteratively until a convergence criterion is satisfied.

Pertusa and Iñesta (2008) proposed an iterative scheme for retrieving the pitch salience function by tracking the single pitch values as the ones whose spectral structure (harmonic spectrum) smoothly evolves over time. Yeh (2008) stated the need for considering the noise components of the spectrum for a proper pitch tracking. In this sense, noise amplitude is modelled using a Rayleigh distribution while the source is modelled as quasi harmonic; iteratively, the spectral peaks are classified as either source or noise; eventually, pitch values are estimated using a joint estimation scheme which minimises inharmonicity and maximises spectral smoothness. Recent approaches such as Kraft and Zölzer (2015) still consider peak selection and pitch salience to obtain the prominent pitch values in the audio source.

Some authors have also considered MPE methodologies that do not exclusively rely on frequency information but also consider temporal cues for the tracking process. For instance, Emiya, David, and Badeau (2007) consider a temporal analysis based on the autocovariance function and a harmonic spectral analysis, both of them biased towards piano tones by considering the inharmonicities present in the resonances of that instrument. Another example is the one by Su and Yang (2015) which also considers harmonic spectral search while the periodicity analysis is performed in the *quefrency* domain.

Other techniques that may also be categorised within this ground are the ones that, instead of manually processing the initial features, use some type of classification scheme. One of the earliest examples of such systems is the one proposed by Marolt (2004) in which a set of networks comprising adaptive oscillators were used for MPE and another group of neural networks was considered for pitch tracking.

Poliner and Ellis (2007) used a scheme based on Support Vector Machines (SVMs) to perform the pitch tracking: 87 one-versus-all SVM classifiers were trained using a set of coefficients directly derived from the spectrum of the signal at issue. Additionally, a Hidden Markov Model (HMM) was considered for smoothing the output of the classifier. Similarly, Nam, Ngiam, Lee, and Slaney (2011) proposed a system based on SVM classification and HMM postprocessing but considering a set of features learned from a Deep Belief Network (DBN) applied to the spectrogram of the signal at issue. More recently, Kelz et al. (2016) studied the influence of the input representation for MPE when considering deep neural networks and convolutional architectures.

Böck and Schedl (2012) consider the use of Recurrent Neural Network (RNN) for the estimation process: two spectrogram analyses of the same input signal (only differing in the analysis parameters) are obtained and processed through a set of semitone band-pass filters; the resulting coefficients are the features for the network, which consists of a set of Bidirectional Long Short-Term Memory (BLSTM) neural networks that models the temporal dependencies in the pitch trajectories.

Finally, a remarkable example of unsupervised learning applied to AMT may be found in Berg-Kirkpatrick, Andreas, and Klein (2014) in which spectral profiles and temporal envelopes are jointly learned through the use of random variables for modeling temporal envelopes, spectral structure, duration, velocity, and activation state.

Statistical model-based methods

This second family of techniques models the MPE problem within a statistical framework. The basic idea is to, given a set of possible pitch activations, finding the subset that maximises a particular statistical criterion in each analysis frame.

Formally, being $\mathbf{x}$ a frame of the signal at issue and $\mathbf{C}$ the set containing all possible pitch combinations, the probability of a subset $C \in \mathbf{C}$ of properly describing $\mathbf{x}$ can be described in terms of Bayesian statistics as:

$$P(C|\mathbf{x}) = \frac{P(\mathbf{x}|C) \cdot P(C)}{P(\mathbf{x})} \tag{2.1}$$

where $P(\mathbf{x}|C)$ stands for the likelihood of frame $\mathbf{x}$ given subset C, $P(C)$ for the prior probability of having subset C, and $P(\mathbf{x})$ for the marginal probability of frame $\mathbf{x}$.

Having modeled the MPE as a statistical problem, the estimation is now ideally reduced to a maximum-a-posteriori (MAP) estimation (Emiya, Badeau, & David, 2010; Benetos, Dixon, et al., 2013):

$$C_{\mathrm{MAP}} = \arg\max_{C \in \mathbf{C}} \frac{P(\mathbf{x}|C) \cdot P(C)}{P(\mathbf{x})} = \arg\max_{C \in \mathbf{C}} P(\mathbf{x}|C) \cdot P(C) \tag{2.2}$$

Furthermore, in case no prior probability is known, this problem is typically reduced to a maximum likelihood (ML) estimation :

$$C_{\mathrm{ML}} = \underset{C \in \mathbf{C}}{\arg\max}\, P(\mathbf{x}|C) \tag{2.3}$$

Some particular examples following Bayesian principles for MPE may be found in works such as Cemgil (2004) or Yoshii and Goto (2012). Alvarado and Stowell (2016) constitutes a particular approach due to its consideration of Gaussian processes for modelling the statistical distributions.

Other examples relying on a statistical framework may be found in the literature: Duan, Pardo, and Zhang (2010) proposed an iterative scheme in which a general likelihood function is derived as likelihood of peak and non-peak regions in the spectrum and a maximum likelihood criterion is applied; Peeling and Godsill (2011) proposed a set of generative models for solving the task, being one of them devoted to modeling the spectral shape of sinusoids and noise floor in the spectrum and another one to evaluate the pitch candidates using a likelihood function; or the system by Koretz and Tabrikian (2011) who proposed an iterative scheme in which, for each iteration of the estimation process, a general criterion based on the combination of maximum likelihood and maximum a posteriori tracks one single pitch while considering the rest as interferences.

Spectrogram factorisation-based methods

This last family of methods is based on the idea that an initial matrix can be (approximately) decomposed into the product of two simpler matrices. Due to its large application, such techniques have been largely studied in many disciplines such as algebra or signal processing.

Matrix factorisation was first introduced to MPE by Smaragdis and Brown (2003) under a Non-negative Matrix Factorisation (NMF) framework. Conceptually, in the context of AMT, one of the matrices models the set of pitch activations (the actual MPE) and the other one approximates the spectral bases of the instruments in the signal. Formally, the non-negative spectrogram $\mathbf{S} \in \mathbb{R}_{+}^{K \times N}$ of K frequency bins and N analysis frames of signal $\mathbf{s}$ can be approximated as:

$$\mathbf{S} \approx \mathbf{W}\mathbf{H} \tag{2.4}$$

where matrix $\mathbf{W} \in \mathbb{R}_{+}^{K \times R}$ models the aforementioned spectral bases, matrix $\mathbf{H} \in \mathbb{R}_{+}^{R \times N}$ the pitch activations and the dimensions satisfy $R << K, N$. Figure 2.3 exemplifies this decomposition process.

This particular approach proved to be quite effective for MPE, thus many research works in the literature have considered similar approaches with particular adaptations. For instance, Vincent, Bertin, and Badeau (2010) propose an NMF-based decomposition that includes additional constraints about the level of harmonicity and spectral smoothness in the resulting matrices. Similar to this, Weninger, Kirst, Schuller, and Bungartz (2013)

$$\begin{pmatrix} \mathrm{S}_{11} & \mathrm{S}_{12} & \dots & \mathrm{S}_{1\mathrm{N}} \\ \mathrm{S}_{21} & \mathrm{S}_{22} & \dots & \mathrm{S}_{2\mathrm{N}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{S}_{\mathrm{K}1} & \mathrm{S}_{\mathrm{K}2} & \dots & \mathrm{S}_{\mathrm{KN}} \end{pmatrix} \approx \begin{pmatrix} \mathrm{W}_{11} & \dots & \mathrm{W}_{1\mathrm{R}} \\ \mathrm{W}_{21} & \dots & \mathrm{W}_{2\mathrm{R}} \\ \vdots & \ddots & \vdots \\ \mathrm{W}_{\mathrm{K}1} & \dots & \mathrm{W}_{\mathrm{KR}} \end{pmatrix} \cdot \begin{pmatrix} \mathrm{H}_{11} & \mathrm{H}_{12} & \dots & \mathrm{W}_{1\mathrm{N}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{H}_{\mathrm{R}1} & \mathrm{H}_{\mathrm{R}2} & \dots & \mathrm{H}_{\mathrm{RN}} \end{pmatrix}$$

Figure 2.3: Example of a matrix factorisation process.

also proposed a supervised NMF with spectral templates for pairs of pitch and instrument, together with constraints about harmonicity, sparseness, and temporal continuity; additionally, the results of the NMF analysis are then used as features for an SVM classifier that performs the pitch-tracking task due to the reported robustness of such schemes. Dessein, Cont, and Lemaitre (2010) explored NMF in real-time transcription applications, leading to accuracy results comparable to offline NMF systems. The work by Arı, Şimşekli, Cemgil, and Akarun (2012) proposes a method for efficiently training an NMF-based transcription model in the context of large scales music collections. As a last example, Cheng, Mauch, Benetos, and Dixon (2016) proposed a system that, in addition to the NMF decomposition, provides an note attack/decay model biased towards piano sounds to improve the obtained results.

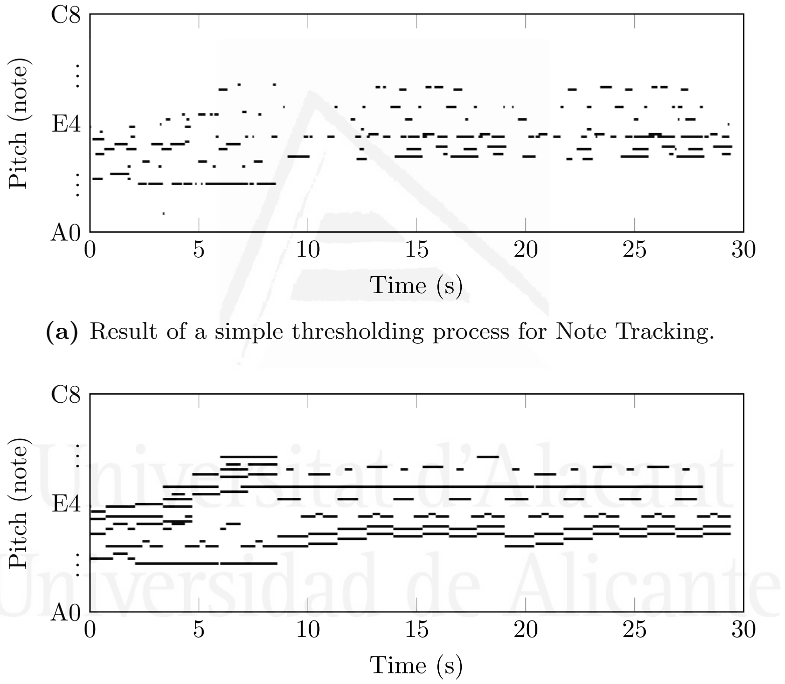
The Probabilistic Latent Component Analysis (PLCA) model, which is the probabilistic extension of NMF, has also attracted the attention of a number of authors for years. The basic idea is that the input time-frequency representation is considered a bivariate probability distribution (time and frequency dimensions) to be able to use statistical techniques for its analysis (Smaragdis, Raj, & Shashanka, 2006). Some examples of remarkable works considering this framework are the ones that follow: Fuentes, Badeau, and Richard (2011) proposed a system that considers the temporal variation of both the spectral envelope and pitch of the harmonic events; Grindlay and Ellis (2011) addressed the issue of multi-instrument polyphonic transcription and considered the spectral structure of different instruments during the training stage; Benetos and Dixon (2012) introduced a similar multi-instrument transcription approach but introducing multiple spectral templates per pitch and instrument; this last approach has subsequently been improved introducing note models with the use of HMMs (Benetos & Dixon, 2013); finally, due to the computational cost this technique implies, some works such as Benetos and Weyde (2015) have addressed this issue and proposed efficient versions of PLCA schemes.

Another commonly consider variant to the NMF methodology is the so-called *sparse coding*. This framework basically performs very restrictive constraints about the sparseness in the resulting matrices obtained. Some works addressing the MPE problem from this perspective are the ones by Lee, Yang, and Chen (2012), O'Hanlon, Nagano, and Plumbley (2012) and

Cogliati, Duan, and Wohlberg (2015), among others.

2.1.2 Note Tracking (NT)

As previously commented, the aim of the NT stage is to process the results obtained in the MPE one to provide a note-level description of the signal in terms of discrete pitch values, onset points, and offset events such as the piano-roll description shown in Fig. 2.4. However, although this stage is the one providing musically-meaningful representations, it has not been as thoroughly studied as the MPE one. Due to the relevance of this process in the dissertation, we now provide a revision of the different techniques considered for this process.



(a) Result of a simple thresholding process for Note Tracking.

(b) Ground truth data of the piece.

Figure 2.4: Example of Note Tracking analysis applied to the piece in Fig. 2.2.

A very simple but yet commonly considered method consists of binarising the MPE result by directly applying a global threshold to the pitch activations: values over the threshold are considered active pitch elements while the ones below it are considered silence. Some works considering this approach are the ones by Vincent et al. (2010) and Grindlay and Ellis (2011). However, due to its simplicity, this type of approaches are not robust enough against errors that might occur in the MPE stage as, for instance, false voice alarms or over-segmentation of long activations.

In this regard, alternative techniques which postprocess the initial binarisation are also considered to palliate those types of errors. Most commonly, these techniques are based on combinations of *minimum-length pruning* processes for eliminating spurious detections and, occasionally, *gap-filling stages* for removing small gaps between between consecutive note events. Quite often, these techniques are implemented as rule-based systems. For example, works by Dessein et al. (2010) and Benetos and Weyde (2015) considered simple pruning stages for removing false detections, while the system in Bello, Daudet, and Sandler (2006) studied a more sophisticated set of rules comprising both pruning and gap-filling stages.

Probabilistic models have also been considered for this NT process. In this regard, HMMs have reported remarkably good results in the literature: the work by Ryynänen and Klapuri (2005) considered HMMs to model note events in terms of their attack, sustain, and noise phases; Cheng et al. (2015) also proposed an HMM with four stages to model the phases of a musical note; finally, other works such as Poliner and Ellis (2007); Benetos and Dixon (2013); Cañadas-Quesada, Ruiz-Reyes, Vera-Candeas, Carabias-Orti, and Maldonado (2010) proposed systems in which binary HMM models used for modeling events as either active or inactive.

Alternative methodologies to the commented ones may also be found in the literature. For instance, Raczyński, Ono, and Sagayama (2009) proposed a probabilistic model based on dynamic Bayesian networks that takes as input the result of an NMF analysis. Other examples are the proposal by Duan and Temperley (2014) that models the NT issue as a maximum likelihood problem, the one by Pertusa and Iñesta (2012) that addresses this task by favoring smooth transitions among partials, or the work by Weninger et al. (2013) that proposed a classification-based approach for the NT stage based on SVMs taking as features the results from an NMF analysis.

It must be noted that, in general, MPE systems are rather imprecise in terms of timing. Examples of typical issues are their tendency to miss note starts, mainly due to the irregularity of the signal during the attack stage, the over-segmentation of long notes or the merge of repeated notes (e.g., tremolo passages) into single events. Hence, the use of timing information in this context is clearly necessary and useful.

Under this premise some works have considered the use of onset information to palliate such issues. Examples of works taking advantage of onset information may be found in Marolt and Divjak (2002), which considers onset information for tackling the problem of tracking repeated notes, the work by Emiya, Badeau, and David (2008), in which onset information is used for segmenting the signal before the pitch estimation phase, the proposal by Iñesta and Pérez-Sancho (2013), which postprocesses the result of the MPE stage with the aim of correcting timing issues with onset information, or the system by Grosche et al. (2012), which also considers onset information under an HMM framework. In addition, some authors such as Benetos and Dixon (2011) have considered offset information additional to the onset one to still improve the obtained results.

2.1.3 User interaction

As introduced, user interaction has been recently considered as an alternative framework to tackle some issues found in classical autonomous systems. However, this change of paradigm has some implications (Kirchhoff, 2013): firstly, one of the challenges in interactive systems is to find areas in which the user input can be beneficial; secondly, this paradigm is not applicable to the analysis of large databases given that these systems are not completely autonomous; lastly, as the success of the task can be guaranteed at the expense of user effort, one of the challenges remains in developing approaches that optimize the user effort invested (Iñesta & Pérez-Sancho, 2013).

Among the MIR literature, a large number of authors have explored the use of interactivity in the particular field of source separation. For example, Smaragdis and Mysore (2009) proposed a system in which the main melody is separated aided by its hummed version provided by an external user. With the same aim of separating the main melody of the signal, Fuentes, Badeau, and Richard (2012) considered the use of a mid-level representation of the signal in which the user notes such melody. Another example may be found in Ozerov, Vincent, and Bimbot (2012) in which a framework for incorporating prior information about the number and types of sources to a source separation scheme is implemented.

In contrast to the source separation problem, few authors have considered the use of interactivity applied to AMT. One of the first examples is the one in Dittmar and Abeßer (2008) who proposed an interactive system for the transcription of melody, bass, chord, and percussion that allows to adjust the tracked notes to an estimated beat grid and to the diatonic scale derived from the tonality (particularly, the key of the piece) specified by the user. Nevertheless, no formal evaluation was proposed in this work.

Kirchhoff, Dixon, and Klapuri (2012) proposed an study comparing two types of user input for an NMF transcription system: in the first variant the user specifies the instruments present in the piece, thus a set of learned instrument spectra are used for the decomposition; in the second variant the user is allowed to label notes in the piece as belonging to different instruments, and thus the instrument spectra is directly estimated from those annotated notes. As a conclusion, the latter variant provided better results than the former one. In a further development, Kirchhoff, Dixon, and Klapuri (2013) expanded the previous one to minimise the user intervention by reducing the amount of data required to be labelled. More precisely, this approach reduced the need for labeling examples at each pitch value and instrument with methods such as replicating spectra from other labelled examples, interpolating partial amplitudes from adjacent notes or adapting pre-learned spectra, among others.

Iñesta and Pérez-Sancho (2013) proposed an AMT system for monotimbral polyphonic music transcription in which the user is allowed to interact with note onsets: assuming that the user corrects the estimation in a left-to-right fashion, an interaction at a certain point implicitly validates all estimations before that mark and the results after that point are recomputed somehow taking into consideration the implicitly validated information.

A recent example is the one by de Andrade Scatolini, Richard, and Fuentes (2015) in which a PLCA system is proposed in which the user interaction consists in providing a transcription of an excerpt of the signal for training the model, being the remaining part transcribed using this model.

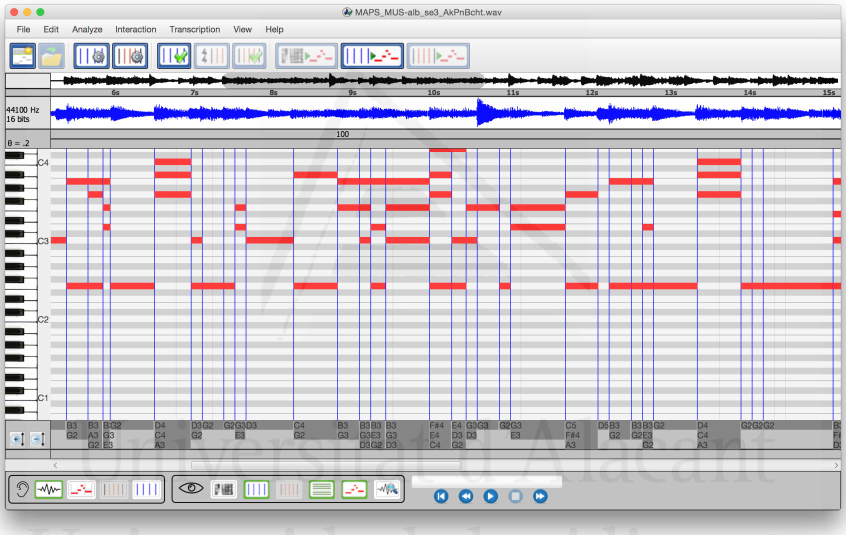


Figure 2.5: Screenshot of the computer-assisted transcription tool by Pérez-García et al. (2011) for monotimbral polyphonic music.

Finally, it must be pointed out that, besides research examples, some authors have also developed tools for performing interactive transcription of music pieces. For instance, Pérez-García et al. (2011) implements the monotimbral polyphonic transcription system[3] described in the paper by Iñesta and Pérez-Sancho (2013) and shown in Fig. 2.5. Also, the work by Mauch et al. (2015) presents the so-called Tony[4] tool (Fig. 2.6 shows a screenshot of the system) specifically designed for melody transcription. Finally, the work by Dixon (2001) presents a visualisation tool for interactively correcting beat

[3] `http://miprcv.prhlt.upv.es/index.php?option=com_content&task=view&id=234&Itemid=205`

[4] `https://code.soundsoftware.ac.uk/projects/tony`

information obtained from a beat tracking system.

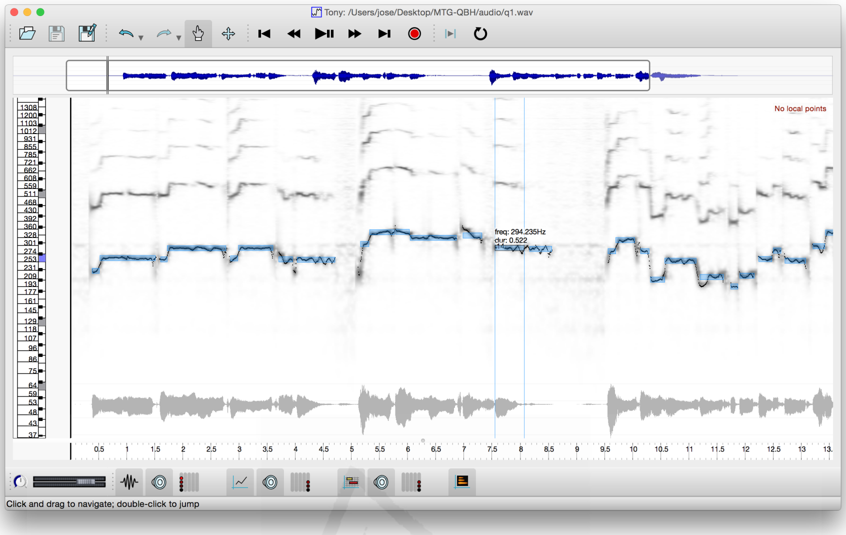


Figure 2.6: Screenshot of the Tony software (Mauch et al., 2015) for computer-assisted melody transcription.

2.1.4 Multimodality

In practice, human transcribers do not only rely on a single *description* of the piece to transcribe. Intuitively, any complementary information that narrows the uncertainty of performing a transcription from scratch clearly simplifies this task: for instance, knowing the tonality and key of a certain piece makes particular notes more likely to appear than others (Krumhansl, 2001). Similarly, other descriptions such as chords, instrumentation or beat information may also provide certain constraints in the search space of the AMT system (Cambouropoulos, 2010). Actually, in practice most AMT systems make *silent* assumptions about those parameters to make the transcription problem tractable (Kirchhoff, 2013). Thus, the use of MIR processes to provide the required additional descriptions or even considering interactive methodologies in which the user provides these high-level pieces of musical information stands as a clear need to develop complete AMT systems.

So far, onset information has been one the most commonly considered complementary descriptions (Benetos, Dixon, et al., 2013). As aforementioned, onset information is essential for the correct temporal description of the signal, and it is typically considered in the NT stage of AMT systems. Offset information has been also considered in some works, but this particular task has, by far, received less attention than the former one.

Timbre information has been also been studied as a way of improving transcription, especially in factorisation-based MPE approaches: although those approaches are commonly trained following an unsupervised fashion, some authors include timbre templates in the process. An example of this methodology may be found in the system by Cazau, Revillon, Krywyk, and Adam (2015) in which the authors propose a PLCA-based system that incorporates information about the three types of instruments considered: piano, guitar, and zither. Also, Cazau, Wang, Chemillier, and Adam (2016) explored the use of timbral information as a prior for an MPE system based on PLCA for the particular case of the *marovany* zither. This last work also studied the use of additional constraints for the AMT system by incorporation both music language models, which shall be later introduced, and modelling the style of the player.

Beat information is another source of information whose utility in AMT has been considered by researchers. For instance, Raphael (2005) considered a system for singing voice transcription through a graphical model that jointly estimates pitch, rhythm, temporal segmentation, and tempo information. Other examples of works considering beat information but focusing on polyphonic music is the one by Kameoka, Ochiai, Nakano, Tsuchiya, and Sagayama (2012), which combines tempo and onset information with the MPE considering a Bayesian framework, and the work by Kameoka, Nakano, et al. (2012), which incorporates musically-meaningful constraints based on note onsets, beat locations, and tempo to the activation matrix of an NMF process, among others.

As commented previously, tonality and key implicitly provide a certain probability distribution of a particular pitch value to appear in the piece. In this regard, some researchers have considered the use of that principle to further incorporate restrictions and knowledge to AMT schemes. Ryynänen and Klapuri (2005) proposed a system in which key information is used for estimating possible note transitions in the transcription of melody and bass lines. More recently, Benetos, Jansson, and Weyde (2014) considered a scheme in which key information is considered within a PLCA-based acoustic model rather than in a postprocessing fashion.

Chord information has also been studied as it also provides a description of the piece in terms of harmony. Laaksonen (2014) proposed a system for melody transcription that considers chord information for segmenting the audio signal into single units and then applying a rule-based approach for estimating the notes of the melody in each segment. Raczyński, Vincent, Bimbot, and Sagayama (2010) proposed a probabilistic framework that jointly models the temporal dependencies between the notes and the underlying chords based on musicological models in a Bayesian framework. More recently, the same authors proposed an improvement of this work for solving a series of limitations and approximations in the previous work (Raczyński, Vincent, & Sagayama, 2013).

Finally, in a similar way to language models for speech recognition, music language models have been considered as a possible tool for improving AMT results by modeling music dependencies in a symbolic domain (Cemgil, 2004). However, the long-term dependencies found in music as well as the complexity of modeling concurrent pitches in polyphonic music due to the combinatorial issue it supposes have limited the use of models typically considered in speech recognition such as n-grams or HMMs. Boulanger-Lewandowski, Bengio, and Vincent (2012) proved that an appropriate scheme for tackling them is considering RNNs for modeling time dependencies and energy-based methods as Restricted Boltzmann Machines (RBM) for the polyphony issue. Based on that, Sigtia et al. (2014) proposed a music language model in which prior information given by the RNN-based symbolic model is incorporated to a PLCA-based MPE scheme that improves transcription results when compared to exclusively relying on the acoustic model. More recently, the same authors proposed a graphical probabilistic model that gathers the acoustic model (frame-level classifier) with the symbolic one within a hybrid architecture so that training can be jointly done (Sigtia et al., 2015). Finally, a last example can be found in the work by Ojima, Nakamura, Itoyama, and Yoshii (2016) in which a hierarchical Bayesian model that fuses an NMF acoustic model with an HMM symbolic one that relates pitch and chord information is studied.

2.2 Onset Detection

Onset detection stands for the automatic estimation of the starting points of note events in music audio signals (Bello et al., 2005). Despite its conceptual simplicity, onset information has proved to be undoubtedly useful for a wide range of MIR tasks as, for example, *beat detection* (Ellis, 2007), *tempo and meter estimation* (Alonso, Richard, & David, 2007) or *audio transformations* (Dorran & Lawlor, 2004), among others. In this regard, the particular interest of this information in this work is its utility in the field of AMT (Benetos & Dixon, 2011). Figure 2.7 shows an example of audio signal together with its onset events.

Onset estimation approaches may be categorised in two families depending on the principle in which the process is based (Chuan & Chew, 2008; Schlüter & Böck, 2014): methods considering a signal processing approach or schemes based on machine learning techniques.

While most research efforts have been typically devoted to the signal processing paradigm, some examples may be found in the literature for the machine learning case. In this context, one of the earliest examples of this paradigm is the work by Marolt, Kavcic, and Privosnik (2002) in which a set of neural networks were considered for the task taking as input the bands of the signal obtained from a bank of auditory filters. Lacoste and

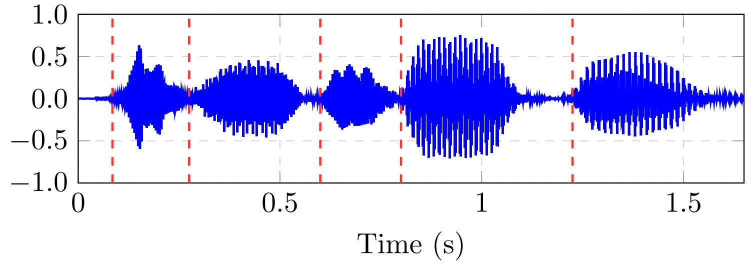


Figure 2.7: Example of onset detection process. Dashed vertical lines represent the onset events in the signal.

Eck (2007) also exploited the use of neural networks but meant for obtaining an enhanced spectrogram representation for the onset detection process. Kapanci and Pfeffer (2006) proposed a system for tracking soft onsets based on a hierarchical architecture of SVM classifiers. Giraldo, Ramírez, and Rollin (2016) proposed an ensemble method evaluating different classifiers for this particular task, being an SVM-based ensemble the one achieving the best results. Finally, more recent approaches based on Deep Learning have also been considered as in the works by Schlüter and Böck (2013, 2014) in which Convolutional Neural Networks are applied to the spectrogram of the signal at issue by considering it an image, or the works by Eyben, Böck, Schuller, and Graves (2010); Marchi et al. (2014) which exploit the use of RNNs, more precisely Long Short-Term Memory neural networks, for performing a time-aware classification scheme to track onset events.

On the other hand, signal processing schemes generally base its performance on a two-stage approach (Glover, Lazzarini, & Timoney, 2011; Zhou, Mattavelli, & Zoia, 2008): a first stage known as Onset Detection Function (ODF) that processes the target signal computing a time series $O(t)$ whose peaks represent the positions of the estimated onsets by measuring the change in one or more audio features; and a second stage called Onset Selection Function (OSF) that evaluates $O(t)$, selects the most promising peaks as onsets and retrieves them as a list of L time stamps, $(o_i)_{i=1}^{L}$. Figure 2.8 graphically shows this process.

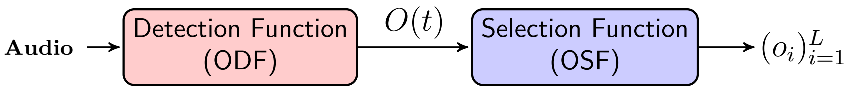


Figure 2.8: Diagram a two-stage onset detection system. The Detection stage processes the initial audio piece, thus retrieving time series $O(t)$; this function is then evaluated by the Selection stage that eventually retrieves the list of estimated onset events $(o_i)_{i=1}^{L}$.

Due to its relevance in this work, we shall thoroughly describe this process in the following sections.

2.2.1 Onset Detection Function (ODF)

As aforementioned, the ODF stage computes time series $O(t)$ by measuring changes in one or more audio features that depict the presence of onsets in the signal. This $O(t)$ function is typically meant to depict likely onset positions as local maxima in the function, thus a spiky shape is generally expected for this time series. Figure 2.9 shows an example of this $O(t)$ function.

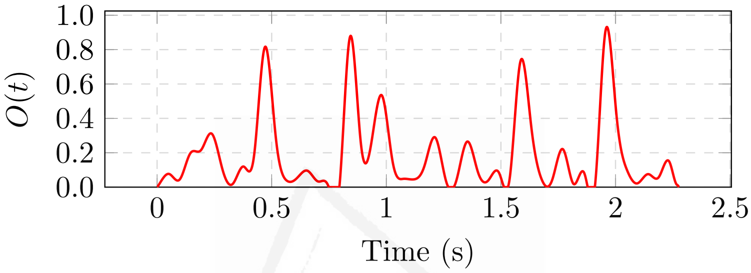


Figure 2.9: Example of time series resulting from an Onset Detection Function analysis.

Among the different possible features, one of the most commonly considered ones is signal energy. Under this principle the idea is tracking energy rises in the signal that might depict a note onset event. Some works considering such representation are the ones by Klapuri (1999); Goto (2001); Duxbury, Sandler, and Davis (2002); Pertusa, Klapuri, and Iñesta (2005).

The use of signal energy for onset detection has been reported to achieve remarkably good results in the case of sounds with sharp attack phases, such as plucked instruments (e.g., guitar, harpsichord or balalaika) or struck instruments (e.g., piano, marimba or clavichord). Nevertheless, for instruments depicting soft attack phases (e.g., non-staccato violin or hurdy-gurdy), this approach does not generally perform that well. In such situations, onset events are alternatively tracked considering changes in the phase information of the signal. Examples of works considering such principle are the one by Bello and Sandler (2003) or the work by Holzapfel, Stylianou, Gedik, and Bozkurt (2010).

Alternatively, pitch information has been also considered for this task. In this case, the idea is to obtain a contour of the pitch information of the signal to then track changes in that contour. An example of this principle is found in the work by Collins (2005) in which an onset detector and segmentation scheme was proposed for monophonic data.

Finally, it must be mentioned that some methods consider the use of combinations of the different approaches previously described. Some examples of works following this approach are the ones by Bello, Duxbury, Davies, and Sandler (2004); Zhou and Reiss (2007); Benetos and Stylianou (2010).

2.2.2 Onset Selection Function (OSF)

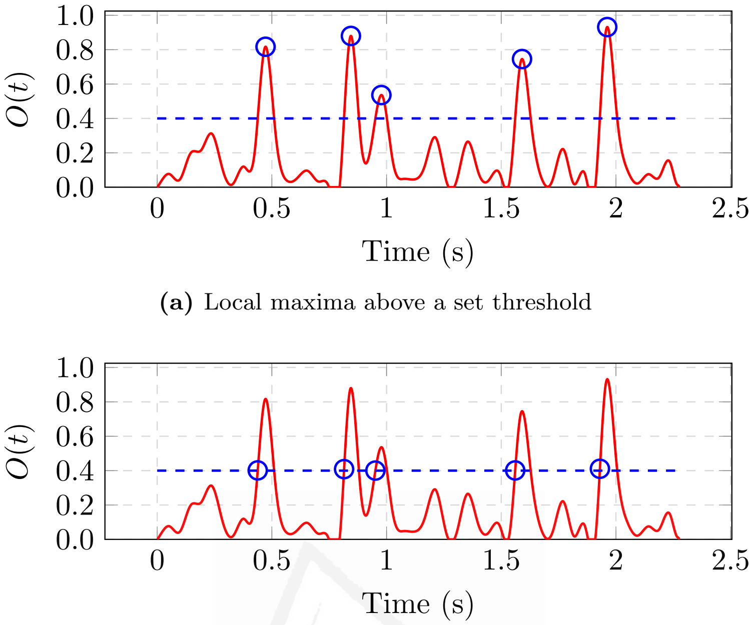
The OSF stage evaluates the $O(t)$ time series resulting from the ODF process and selects the most promising points as onsets. Ideally, if the ODF method would perfectly track the onset events present in the signal, this stage would not be required. However, given that no ODF is capable of doing so, the OSF is required in order to discriminate between actual onsets and artifacts, thus constituting a key point in the overall performance of the onset tracking system (Rosão, Ribeiro, & Martins de Matos, 2012; Dixon, 2006).

In general, OSF methods evaluate the $O(t)$ function by seeking for peaks above a certain threshold, which is generally considered to eliminate noisy elements and artifacts that may have appeared during the ODF stage. Peaks are generally tracked by seeking for local maxima in the $O(t)$ function, but as this process is not causal, some authors considered threshold-triggering processes so that they can be used in real-time processing systems (Stowell & Plumbey, 2007). Figure 2.10 shows the difference between these two OSF methods applied to an $O(t)$ function when considering a basic thresholding process.

Regardless of the peak picking methodology considered, OSF techniques generally differ in the way this threshold is obtained. The most basic approach consists in manually establishing a static threshold value (Klapuri, 1999), as the one in Fig. 2.10. This technique does not consider any particularities of the signal in the sense that no prior knowledge is taken into account but the value is simply heuristically set.

In order to consider the particularities of the signal, other authors consider the possibility of computing some statistical descriptor on the $O(t)$ function and set it as the static threshold value (Böck et al., 2012). Typical descriptors considered are the mean or the median of the function, being the latter a commonly addressed one as it has been thoroughly studied for noise deletion (Kauppinen, 2002).

The issue with the aforementioned strategies is that they do not consider the temporal evolution of $O(t)$: once the threshold value is set, this value does not change throughout the function, which may be inappropriate. Thus, adaptive techniques that consider the temporal evolution of the $O(t)$ are also used. Instead of obtaining a global static threshold value, these methods consider a sliding window, whose size is generally set in an empirical way, for performing the analysis at each point of the $O(t)$ function (Duxbury, Bello, Davies, & Sandler, 2003). As before, a statistical descriptor is considered for obtaining the threshold value for each window, being the median value a typically considered one. Figure 2.11 shows an example of both the sliding window analysis and the resulting $O(t)$ function.



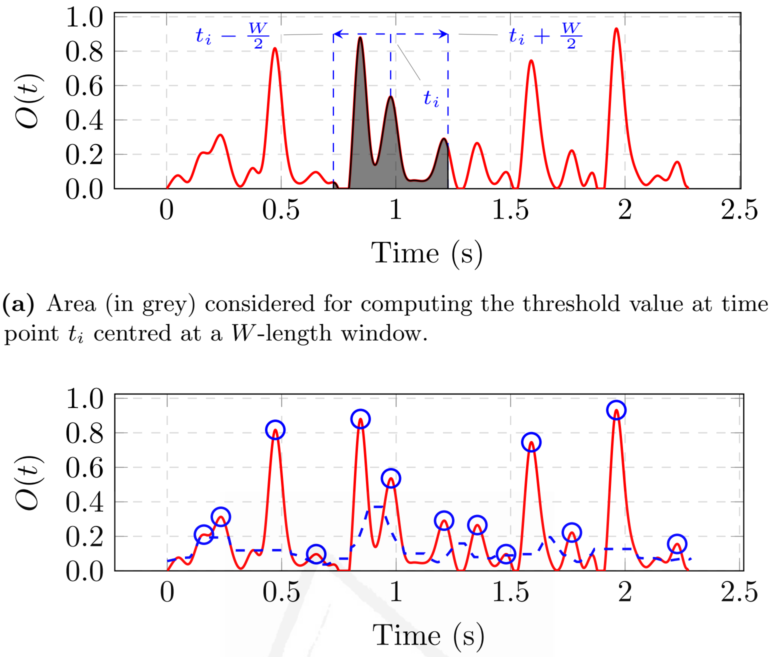
(a) Local maxima above a set threshold

(b) Threshold surpassing

Figure 2.10: Example of static threshold Onset Selection approach. Figures compare the obtained onsets (○) when the Selection Function estimates onsets as local maxima above a certain threshold **(a)** or at the point in which a threshold is surpassed **(b)**.

Extending these ideas, some authors have considered more sophisticated ideas to improve the performance of the systems. Examples of such work comprise the one by Bello et al. (2006), who consider an OSF strategy that includes low-pass filtering for $O(t)$ to additionally remove noisy peaks in the function, or the work by Dixon (2006), who considers an additional set of thresholds for further discarding noisy elements.

Finally, it must be mentioned that machine learning techniques have been also considered in opposition to the commented *hand-crafted* methodologies. Some examples are the work by Abdallah and Plumbley (2003), which considers an HMM-based clustering approach to evaluate the result of an ODF process, or the work by Böck, Schlüter, and Widmer (2013), in which an RNN (unidirectional and bidirectional for non-real time and causal processing, respectively) is trained for evaluating the result of a state-of-the-art ODF method.



(a) Area (in grey) considered for computing the threshold value at time point t_i centred at a W-length window.

(b) Resulting threshold (dashed line) when considering the median statistical descriptor and a window of $W = 0.5\ s$.

Figure 2.11: Example of adaptive threshold Onset Selection based on a sliding window.

2.3 Evaluation methodologies

A large number of the tasks addressed by the MIR community imply a high degree of subjectivity: assessment for tasks such as *music similarity*, *genre classification* or *mood detection* is not easily addressable since even for human beings there is no a clear consensus about what makes two pieces similar or where the actual frontier of a music genre is.

Ideally, AMT should not exhibit such degree of subjectivity: one would expect to obtain an exact copy of the original score of a piece after applying an AMT to its audio version. However, as proved by List (1974), even expert human transcribers differ usually when annotating music pieces, especially in relation to rhythmic aspects of music. This fact points out that different annotations may be perfectly valid despite showing variations among them.

Nevertheless, the MIR community has been largely studying the development of measures to objectively assess and compare different MIR systems, being the annual Music Information Retrieval Evaluation eXchange (MIREX) contest[5] the most representative example of such efforts.

The following sections introduce the MIREX measures related onset

[5] http://www.music-ir.org/mirex/wiki/MIREX_HOME

detection and to AMT schemes as they constitute the ones of interest for the rest of the work.

2.3.1 Onset detection

As frequently reported in the literature, the start of a musical event is not a specific point in time but rather a time lapse known as *rise* or *transient* time (Lerch & Klich, 2005; Bello et al., 2005). This is graphically shown in Fig. 2.12: a sound with a typical Attack-Sustain-Release envelope is shown; the onset position o_i is marked as being in the middle of the transient, but any of the points fulfilling $o_i \pm W_o$ is equally valid as they belong to the transient time as well.

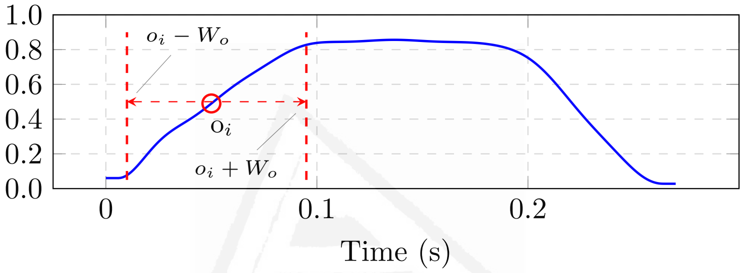


Figure 2.12: Example of onset event in a sound showing a typical Attack-Sustain-Release envelope. Position o_i reveals the point selected as onset while range $[o_i - W_o, o_i + W_o]$ contains the rest of the possible positions for the onset event.

Owing to this *loose* definition, onset detection algorithms are given a certain time lapse in which the detection is considered to be correct. Most commonly, this acceptance window has been set to $W_o = 50\ ms$ following the criterion in the MIREX contest. More recent works have considered more restrictive tolerance windows as, for instance, the one by Böck et al. (2012) in which this value is lowered to $W_o = 30\ ms$ as the authors point out it represents a proper time lapse for human beings to be able to detect onsets. Nevertheless, except where noted, this Thesis work considers the $W_o = 50\ ms$ tolerance window considered by the MIREX contest.

Let us now define True Positives (TPs) as the detected onsets that match a reference annotation within the W_o tolerance window, False Positives (FPs) as the detected onsets not matching any reference annotation, and False Negatives (FNs) as the reference annotations not matching any detected element. With such concepts, onset detection systems are typically assessed in terms of their Precision and Recall using Eqs. 2.5 and 2.6, respectively.

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2.5}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.6}$$

Given that these two measures are generally opposed, the F-measure is usually considered for obtaining a global measure that properly summarizes the performance of system. This metric is obtained as the harmonic mean of both Precision and Recall measures:

$$\mathrm{F}_{\beta} = (1 + \beta^2) \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\beta^2 \cdot \mathrm{P} + \mathrm{R}} \tag{2.7}$$

being β a parameter for giving more relevance to one of the previous measures. Most often, a weight of $\beta = 1$ (both measures are equally weighted) is typically considered, giving the following expression:

$$\mathrm{F}_1 = \frac{2 \cdot \mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}} = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{2.8}$$

Finally, Fig. 2.13 shows the best and worst results obtained for the three measures considered for the onset detection task in each of the editions of the MIREX contest. In general these figures suggest that some systems show a remarkably accurate performance, especially in the latest editions of the contest with results of $\mathrm{F}_1 \approx 0.9$, although there is still room for improvement. Nevertheless, it must be noted that the MIREX dataset remains exactly the same since its inception, thus not being realistic results in the sense of representing real-life applications.

2.3.2 Automatic Music Transcription (AMT)

Evaluation of AMT systems is typically performed considering a piano-roll representation, such as the one in Fig. 1.1: for each analysis frame, the system outputs the set of both active and inactive pitch elements, which are then compared to the set of annotations.

With such representation two different types of evaluation may be performed (Bay, Ehmann, & Downie, 2009): a *frame-based* assessment that evaluates the correctness of the estimation in a frame-by-frame basis, or a *note-based* evaluation that characterizes the events in the piano roll as notes defined by an onset, an offset, and a pitch value and compares them to the events in the annotations.

These two assessment methodologies, which also constitute the ones considered in the MIREX contest, are explained in the following sections.

Frame-based

A first figure of merit defined for the frame-based analysis is the so-called *Accuracy*. As in the onset detection case, this figure is defined in terms of

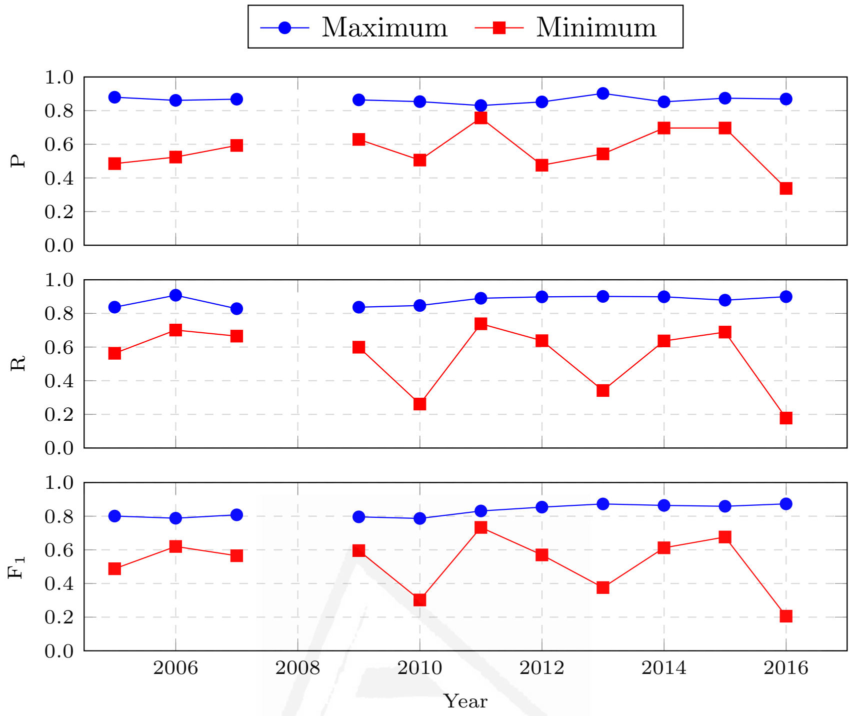


Figure 2.13: Results of the onset detection task for the different MIREX editions. Only best and worst results per year are shown. No information about the 2008 edition is included as the task was not considered for the contest.

TPs, FPs, and FNs as follows:

$$Accuracy = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{2.9}$$

being now TP a pitch element detected as active that matches an active pitch annotation within less than half semitone (±3 % in terms of pitch value).

However, as reported by Bay et al. (2009), this *Accuracy* measure is not descriptive in terms of error analysis. In that sense, other figures of merit have been devised to perform such type of analysis and thus complement the aforementioned metric.

The first of them is the *frame-level transcription error* and was initially used by Poliner and Ellis (2007) for AMT. This metric summarizes the total number of errors in the estimation as a single value and is defined as:

$$E_{tot} = \frac{\sum_{t=1}^{T} \max(\mathrm{N}_{\mathrm{GT}}(t), \mathrm{N}_{\mathrm{EST}}(t)) - \mathrm{N}_{\mathrm{COR}}(t)}{\sum_{t=1}^{T} \mathrm{N}_{\mathrm{GT}}(t)} \tag{2.10}$$

where $\mathrm{N}_{\mathrm{GT}}(t)$ stands for the total number of active pitches at frame t in the reference annotations, $\mathrm{N}_{\mathrm{EST}}(t)$ is the total number of estimated active pitches at frame t, and $\mathrm{N}_{\mathrm{COR}}(t)$ is the number of correct TPs for that frame. E_{tot} is not bounded as errors in the estimation may suppose this metric to span over the unit. However, the value of 0 occurs when a perfect transcription is reported and a value of 1 when no single pitch value is correctly tracked.

To further analyze the errors of the system, E_t may be decomposed into three types of different errors: substitution errors, missed elements errors, and false alarm errors. We shall now introduce these types of errors.

The substitution error rate accounts for the number of annotated pitch elements for which an incorrect pitch value was returned instead. This rate is obtained as:

$$E_{sub} = \frac{\sum_{t=1}^{T} \min(\mathrm{N}_{\mathrm{GT}}(t), \mathrm{N}_{\mathrm{EST}}(t)) - \mathrm{N}_{\mathrm{COR}}(t)}{\sum_{t=1}^{T} \mathrm{N}_{\mathrm{GT}}(t)} \tag{2.11}$$

The missed errors figure of merit describes the amount of annotated pitch elements that are missed by the AMT system but for which no substitution value is given. This metric is related to the Recall one in Eq. 2.6 and is obtained as:

$$E_{miss} = \frac{\sum_{t=1}^{T} \max(0, \mathrm{N}_{\mathrm{GT}}(t) - \mathrm{N}_{\mathrm{EST}}(t))}{\sum_{t=1}^{T} \mathrm{N}_{\mathrm{GT}}(t)} \tag{2.12}$$

The last component of these errors is the false alarm error rate. This metric accounts for the extra pitch elements detected by the AMT system that are not part of the substitution elements set and is related to the Precision one in Eq. 2.5. This measure may be obtained as:

$$E_{fa} = \frac{\sum_{t=1}^{T} \max(0, \mathrm{N}_{\mathrm{EST}}(t) - \mathrm{N}_{\mathrm{GT}}(t))}{\sum_{t=1}^{T} \mathrm{N}_{\mathrm{GT}}(t)} \tag{2.13}$$

Finally, Fig. 2.14 shows the results obtained in the different MIREX editions in terms of these frame-based evaluation measures introduced. Due to the large number of results, only the best and worst results obtained for each metric and edition of the context are included. As a very broad and qualitative analysis, it can be seen that the accuracy of the systems has not significantly improved for a number of years, as if a glass ceiling had been reached as suggested by Benetos et al. (2012). Nevertheless, it can be checked that the total number of errors committed by the systems (E_{tot}) has remarkably decreased, effect that seems to be mostly due to the reduction in both the number of false alarm errors (E_{fa}) and the number of missed events (E_{miss}).

Note-based

As aforementioned, note-based metrics assess AMT systems by considering note events described by an onset time, an offset time, and a pitch value

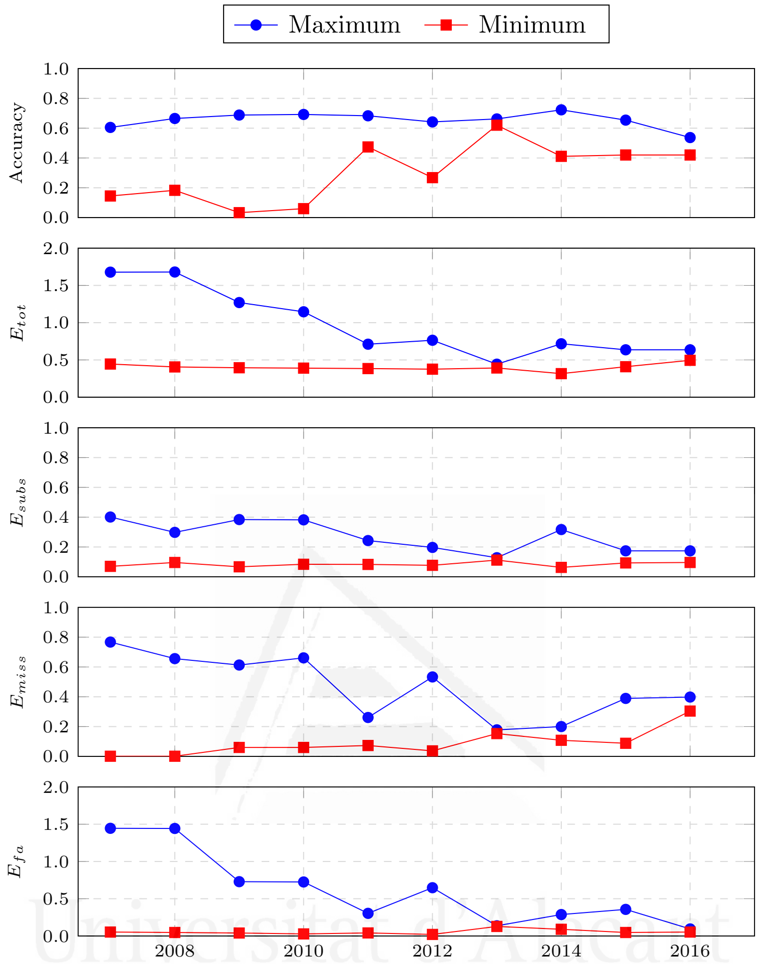


Figure 2.14: Results of the frame-based metrics for the multi-pitch detection task for the different MIREX editions for general music. Only best and worst results per year are shown.

rather than single frames in the piano roll.

In such terms, an annotated note event is considered to be correctly tracked (a TP event) if there exists an estimated note event that accomplishes three conditions: (i) their pitch values differ in less than a quarter of tone (±3 % in terms of frequency values); (ii) the onset time of the estimated event is within a 100ms range of the onset of the annotated event; and (iii) the offset value of the estimated event is within a 20 % range of the offset of the annotated event. In a similar sense to onset detection a to the frame-base AMT measures, FPs occur when estimated events do not have a corresponding note among the annotated corpus and FNs when no annotated event has a corresponding detected note. With these definitions, the figures of Precision, Recall, and F-measure are derived using Eqs. 2.5, 2.6, and 2.8, respectively.

Alternatively, an evaluation methodology that avoids the offset criterion is considered. This is referred to as a onset-based evaluation and, although this relaxation in the assessment clearly implies an improvement in the figures, this is typically done as onset events are considered to be more musically relevant than offset information.

In both cases, a last figure of merit is considered to further analyze the results. This is known as Overlap Ratio (OR) and basically compares how well estimated and annotated note events overlap. This measure is only computed for the TP events as this overlapping assessment requires of a correspondence between estimated and annotated events. For each i note event of the set, this measure is obtained as:

$$\mathrm{OR}_i = \frac{\min(t_{i,\mathrm{off}}^{\mathrm{GT}}, t_{i,\mathrm{off}}^{\mathrm{EST}}) - \max(t_{i,\mathrm{on}}^{\mathrm{GT}}, t_{i,\mathrm{on}}^{\mathrm{EST}})}{\max(t_{i,\mathrm{off}}^{\mathrm{GT}}, t_{i,\mathrm{off}}^{\mathrm{EST}}) - \min(t_{i,\mathrm{on}}^{\mathrm{GT}}, t_{i,\mathrm{on}}^{\mathrm{EST}})} \tag{2.14}$$

where $t_{i,\mathrm{on}}$ and $t_{i,\mathrm{off}}$ refer to onset and offset times of the i-th event, respectively, and super indexes EST and GT indicates whether the note event comes from either the estimated or the annotated set. A global OR is eventually obtained as the average of all individual scores.

Finally, the results obtained for the note-based assessment in the MIREX contest are now included. The evaluation in MIREX is done using two different datasets: a first one considering a general set of instruments (bassoon, clarinet, flute, horn, oboe, violin, cello, guitar, saxophone, and electric bass guitar) and a second one comprising a set of piano recordings. The best and worst results obtained for each of the two datasets in the different editions of the context are shown in Figs. 2.15 and 2.16 for the general and piano sets, respectively.

In a qualitative analysis of such figures, it can be checked that Note Tracking (NT) still shows a remarkable room for improvement: although in the last editions (2014, 2015, and 2016) there is a remarkable improvement with respect to the previous editions, especially in terms of the F_1 measure, results still are far from being perfect. This effect is even more accused when considering the onset-offset evaluation criterion as results in terms for the F_1 rarely go above the value of 0.5, especially for the case of piano music. Thus, given the observed results, NT constitutes a topic to further explore and improve.

2.4 General discussion

AMT is considered to be one of the most powerful, yet challenging, tasks in the field of MIR. The possibility of obtaining symbolic versions from audio

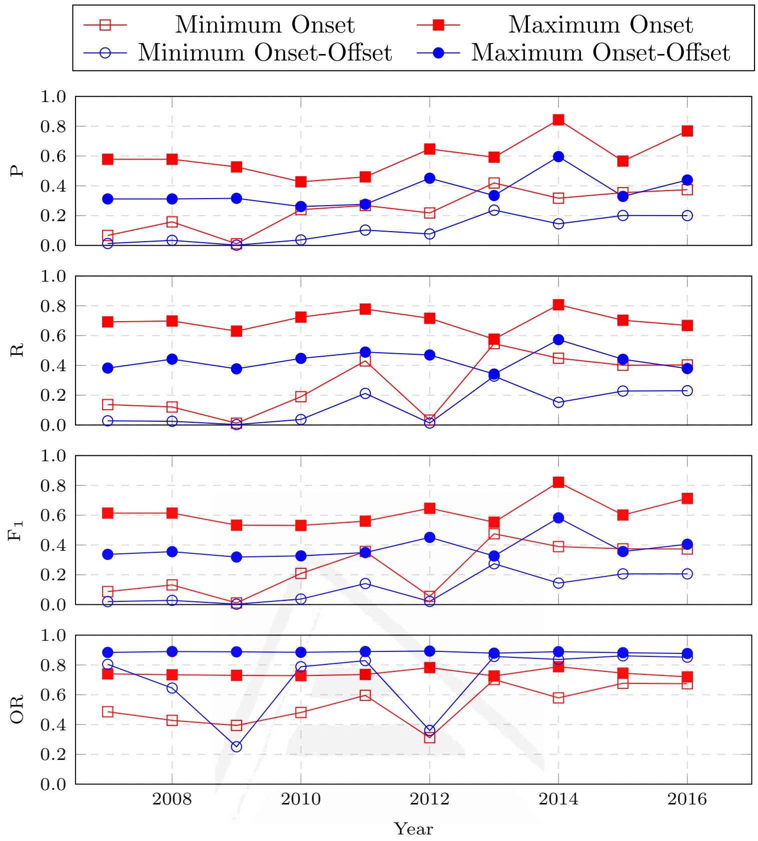


Figure 2.15: Results of the Note Tracking detection task for the different MIREX editions for general music. Only best and worst results per year are shown.

pieces is unarguably appealing for music-related tasks such as musicological analysis, music preservation, similarity-based retrieval and so on.

As commented, AMT systems generally follow a sequential two-stage basis for carrying out the task: an initial MPE stage devoted to retrieving the pitch values of the signal in a frame-based analysis; and second NT phase that retrieves a note-level description of the signal.

From the literature review, it is clear that MPE approaches have been throughly studied in contrast to NT methods. While a wide range of studies and methods may be found for MPE schemes, the set of existing NT approaches is significantly smaller. This is a remarkable fact given that both stages play, in principle, a key role in the overall success of the transcription task. In this sense, it seems interesting to further the influence of NT in transcriptions duties and also contribute with new approaches for it.

Current results, as seen in the MIREX scores and also pointed out by some authors like Benetos et al. (2012), seem to be stagnant for a number

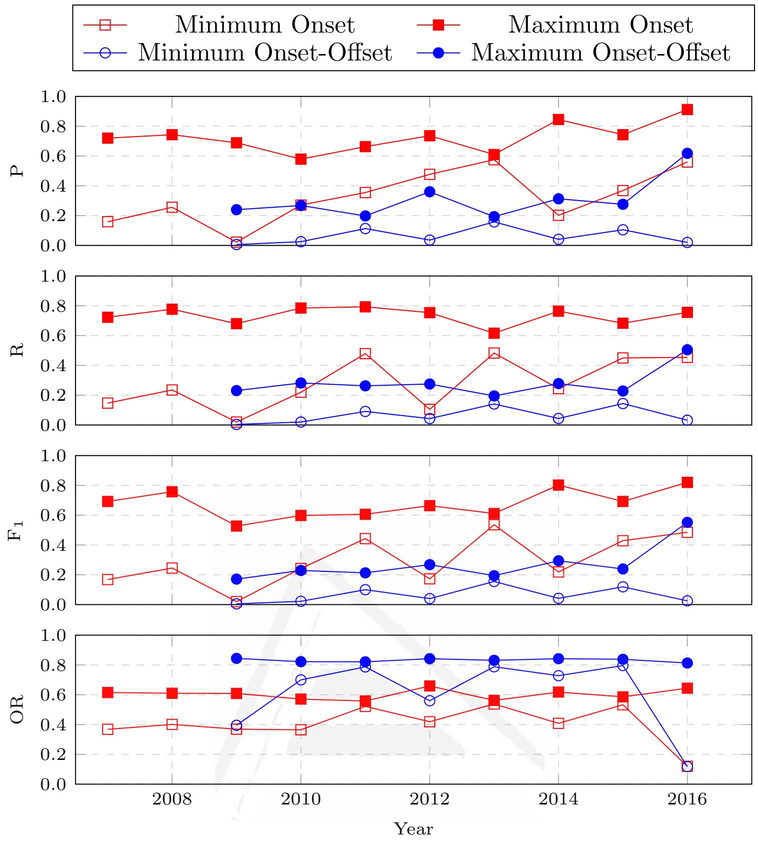


Figure 2.16: Results of the Note Tracking detection task for the different MIREX editions for piano music. Only best and worst results per year are shown. Onset-offset based evaluation was first introduced in 2009.

of years, with slight and marginal improvements being achieved, at least in the typical benchmark datasets. This lack of improvement suggests that the classically considered paradigm may have reached a limit in performance, a *glass ceiling*, and alternative paradigms should be considered to further develop AMT systems.

Given such limitations, in practical scenarios, AMT users need to manually post process the output of such systems to retrieve an accurate enough score-like representation. In that sense, assuming this need for a human presence in the process, the idea of tackling AMT from a *human-computer interaction* perspective (or paradigm) is totally justified and hence addressed in this work.

Multimodality, the use of different sources of information for solving a particular task, is also starting to be relevant among AMT systems, as stated in the literature review. This fact makes perfect sense given that this premise somehow mimics the human methodology: not only the use of harmony or rhythmic information may improve the result of the transcription, but also information about the composer (e.g., preference for scales, modes, or influences), the genre of the piece or the instrumentation basically narrow the amount of possibilities, that is alternative transcriptions. From the existing possibilities we restrict ourselves to onset information as they both constitute a remarkably valuable source of information for fixing timing issues in the detection and allow to work on an interactive basis, as it shall be explained in this Thesis.

Chapter 3

Pattern Recognition in Music Information Retrieval

"Oh, people can come up with statistics to prove anything. Forty percent of all people know that"

Homer J. Simpson

This chapter introduces the concepts of Pattern Recognition relevant for the rest of the work: first, Pattern Recognition is defined and contextualized, being also introduced their interactive variants of particular relevance for the current work; then, a brief introduction of MIR works addressed using Pattern Recognition techniques is introduced; thirdly, the k-Nearest Neighbor is thoroughly explained given its relevance in the present work; finally, a general discussion about the topic in this chapter is presented.

3.1 Pattern Recognition and Machine Learning

Pattern Recognition (PR) is defined as the process of discovering regularities and patterns in sets of data in an automatic fashion, typically with the use of computer algorithms (Bishop, 2006).

A possible strategy for finding such regularities in data is the use of heuristics and handcrafted approaches. These strategies are designed considering the particularities of the problem at issue and lead to *ad-hoc* solutions for it. Although effective, these approaches generally imply very complicated solutions to the task that, in most cases, are not general enough but rather overfitted to the data at issue.

In opposition to these approaches, PR has often taken advantage of Machine Learning techniques in order to automatically infer the aforementioned relations. Such automatic inference does not only solve the previous drawbacks but also opens the possibility to extend those processes to incommensurate amounts of data not manageable in with such manual approaches.

Formally, let $\mathcal{X}$ and $\mathcal{Y}$ represent two data distributions (origin and target) related by function $f : \mathcal{X} \rightarrow \mathcal{Y}$. In general, these data distributions are unknown and simply a set of examples or instances $\mathcal{T} = \{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathcal{X},\ y_i \in \mathcal{Y}\}_{i=1}^{|\mathcal{T}|}$ known as *training set* is accessible. The aim of supervised PR is to obtain a function $\hat{f}$ that approximates as much as possible to f using the set of examples $\mathcal{T}$.

Under this generic premise, different types of PR schemes may be considered depending on the representation of the input elements $\mathbf{x} \in \mathcal{X}$, the model used for estimating the approximated function $\hat{f}$, and the representation of the output elements $y \in \mathcal{Y}$. These concepts are now discussed.

Depending on the **representation for the input data** $\mathbf{x} \in \mathcal{X}$, PR schemes are typically divided into two categories (Duda, Hart, & Stork, 2001): a first category known as *syntactical* or *structural* schemes, in which data is represented using flexible high-level data structures (e.g., strings, trees or graphs) that are only tractable by certain PR techniques; and a second one known as *statistical* or *feature-based* representations, in which data is encoded with numerical feature vectors with limited representation flexibility but addressable by any PR technique (Bunke & Riesen, 2012).

The **nature of the output elements** $y \in \mathcal{Y}$ defines two types of PR tasks: when the target distribution comprises a set of discrete categories or labels ($\mathcal{Y} = \{C_1, C_2, ..., C_{|\mathcal{Y}|}\}$), the problem is known as *classification*; alternatively, when the target distribution is takes continues values ($\mathcal{Y} \in \mathbb{R}$), this task is known as *regression* (Murphy, 2012)

In terms of the **strategies for estimating function** $\hat{f}$, methods may be distinguished as being *parametric* or *non-parametric*: in the former case, some type of assumption over the underlying function f (e.g., the number of free parameters) is assumed, whereas in the latter no assumption is considered (Russell & Norvig, 2010).

Additionally, these methods may also be considered as being *eager* or *lazy* learners depending on when the generalization over training data $\mathcal{T}$, that is the estimation of function $\hat{f}$, is performed: the former methods estimate function $\hat{f}$ before any query is made to the system whereas the latter ones infer function $\hat{f}$ each time a query is made. While it is clear that *eager* strategies show a superior time efficiency as function $\hat{f}$ is only derived once, *lazy* learning allows to derive a local approximations of $\hat{f}$ for each query (Mitchell, 1997), which may lead to better results. A particular case of the *lazy* paradigm is the so-called *instance-based* learning in which no explicit generalization process is performed but new instances are instead directly compared to all the examples in the training set $\mathcal{T}$.

Finally, PR may also be pursued in scenarios in which $\mathcal{T}$ only has examples of the origin distribution $\mathcal{X}$, that is $\mathcal{T} = \{(\mathbf{x}_i) : \mathbf{x}_i \in \mathcal{X}\}_{i=1}^{|\mathcal{T}|}$. In such cases the process is known as *unsupervised learning* or *clustering* and its aim is to automatically group the input elements into *clusters* (groups) of elements that maximize both the *similarity* among them and the *dissimilarity* to elements in other *clusters* (Duda et al., 2001).

3.1.1 Interactive Pattern Recognition (IPR)

In general, PR systems are still far from been perfect and error-free. While this imprecision is tolerable for certain applications, in the cases in which accuracy is a must, human agents are required to verify and correct the results obtained by the PR system to retrieve a completely accurate result.

Assuming this need for a human agent to be part of the system, an alternative to PR known as Interactive Pattern Recognition (IPR) has recently emerged (Toselli et al., 2011). Instead of considering the human agent as simply a verification and correction element in the system, IPR aims at studying ways of actively exploiting the correction feedback by the agent and iteratively improve the core PR model. Figure 3.1 shows a graphical description of this idea: an initial set of data is given to a PR model, which proposes a hypothesis about it; this hypothesis is presented to the user, who assesses it and returns that information to the PR model, which modifies its performance according to this feedback provided.

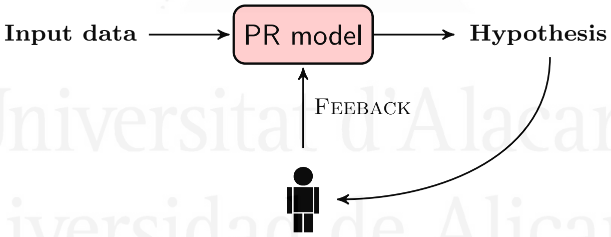


Figure 3.1: Diagram of an IPR system. The hypotheses proposed by the model are validated by the user as in a classic PR task, but in IPR the model is given a certain feedback so that its performance is modified accordingly.

As commented, IPR approaches are able to obtain fairly accurate models at the expense of human intervention. In that sense, performance evaluation must be assessed in terms of the user effort invested in the corrections instead of accuracy since the latter one is guaranteed by the expertise of the user in the field (Vidal, Rodríguez, Casacuberta, & García-Varea, 2008). Thus, for the same input data, the number of interactions required for correcting the result of an IPR system is expected to be lower than the amount required to manually correct the result of a classic PR one as the model in the former case somehow *learn* from the mistakes committed.

A remarkable drawback in IPR is that, each time the user performs a correction, the model is expected to incorporate this new piece of information by training the system. This need for constantly training supposes a clear limitation, especially when considering that within an interactive system the user should not perceive any kind of delay between the correction and the new hypothesis. In this sense, instance-based algorithms are clearly advantageous as incorporating new information to the model is achieved by just adding new instances to the training set.

Interactive Sequential Pattern Recognition (ISPR), a particular case of IPR in which the system deals with sequential data (Calvo-Zaragoza & Oncina, 2017), is of large interest in this Thesis as music information has this nature. In ISPR it is assumed that the user verifies the result following the order of the sequence, most typically a left-to-right fashion. Under this premise, a user correction at a certain point implicitly validates all data between this interaction and the previous interaction performed. Thus, a single interaction provides a larger amount of data to update the model compared to the general IPR case, which is expected to benefit the overall performance of the system.

Finally, given that a large amount of real-world data follows some type of structure, ISPR may be found in a large number of contexts. Some examples of such disparate tasks are human karyotyping (Oncina & Vidal, 2011), computer-assisted translation (Barrachina et al., 2009) or image-to-text transcription (Toselli, Romero, Pastor, & Vidal, 2010), among others.

3.2 Applications to Music Information Retrieval

Given the usefulness of PR for automatically discovering patterns in data, MIR has considerably taken advantage of the different methods and techniques available and adapting them to the particular needs of each task.

A straightforward application of PR to MIR is the task of music genre classification. Remarkable examples of such tasks are the works by Tzanetakis and Cook (2002) in which audio data was considered and a model based on Gaussian Mixture Models (GMMs) was used, the work by Conklin (2013) in which ensemble-based methods were considered for the classification of a set of folk tune songs in terms of both their genre and their geographic localization, or the work by Lidy, Rauber, Pertusa, and Iñesta (2007) in which, considering audio data, a set of descriptors derived from both the audio representation and its symbolic equivalent obtained using an AMT system is considered for training a genre classifier.

Emotion identification and recognition in music constitutes another example of MIR tasks typically addressed with PR techniques. Yang and Chen (2012) performs a review in this topic, highlighting the use of PR as a successful approach.

Music similarity has been also studied from a PR point of view. In this context, some remarkable works are the ones by Rizo (2010) in which symbolic music melodies are encoded as tree data structures and similarity is measured as distances among trees, the work by Bernabeu, Calera-Rubio, Iñesta, and Rizo (2011) in which the previous work is expanded to consider probabilistic tree automata, or the work in Bellet, Bernabeu, Habrard, and Sebban (2016) that also expands the previous ideas to consider tree edit distances directly learned from data.

Some other particular examples of PR in MIR are, for instance, the use of stochastic language models, and particularly, n-grams, for genre, style, and composer identification (Pérez-Sancho, 2009), melody identification in symbolic MIDI files modeling the problem as a classification task (Ponce de León, 2011) or the use of PR techniques for Optical Music Recognition (OMR) as an alternative to heuristic-based analysis methods (Calvo-Zaragoza, 2016). In addition, the work by Poliner (2008) is of particular interest to this work as it considers PR methods for AMT, more precisely using classification-based methods for both the MPE an NT stages.

While the commented works consider the use of supervised learning, unsupervised PR has been also considered for MIR. As a particular example it seems interesting to highlight the use of Self-Organizing Maps (SOM). These methods are a special type of neural network for unsupervised learning that map data in a high-dimensional space to a lower-order one (most usually, a two-dimensional space) while preserving the topological relations of the initial domain as faithfully as possible. In this sense, SOM has been considered for both the unsupervised organization of music genre (Frühwirth & Rauber, 2002) and style (Ponce de León & Iñesta, 2002). Figure 3.2 shows a conceptual example of an SOM process.

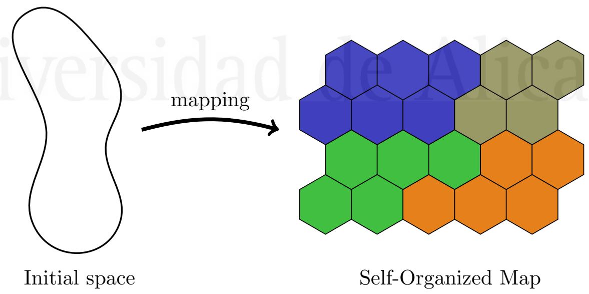


Figure 3.2: Example of an Self-Organizing Map process. The initial high-dimensional data distribution, which is conceptually represented by an amoeba, is mapped into a two-dimensional space in which classes are separated as clusters.

Finally, it is important to highlight the presence of some of the aforementioned tasks as part of the annual MIREX contest, mostly for audio data. Examples of such considered tasks are genre classification (for family genres of US, latin, and K-POP music), mood detection (for both general and K-POP music), classical composer identification, tag classification, and music/speech classification.

3.3 k-Nearest Neighbor (kNN)

Since initially proposed by Fix and Hodges (1951), the k-Nearest Neighbor (kNN) constitutes one of the most well-known instance-based algorithms in PR for supervised non-parametric classification (Duda et al., 2001). Most popularity for kNN in classification tasks is due to its conceptual simplicity and straightforward implementation, as it basically relies on distance comparisons between instances: given a query element, the kNN rule assigns it the most frequent label among the k-nearest prototypes of the training set, where k is a parameter to be set. In addition, the probability of error of this classifier is bounded by twice the Bayes error rate for the case of binary classification (Cover & Hart, 1967). Figure 3.3 shows an example of this classification rule.

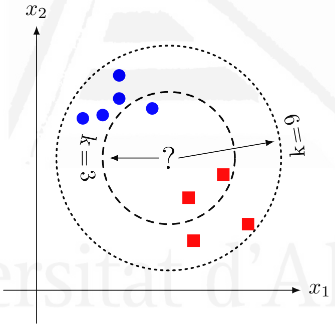


Figure 3.3: Example of the kNN classification rule. As shown, when setting $k = 3$ or $k = 9$, query point is assigned the classes circle or square, respectively.

Again, let $\mathcal{T} = \{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathcal{X},\ y_i \in \mathcal{Y}\}_{i=1}^{|\mathcal{T}|}$ define our training set where $\mathcal{Y}$ is a set of discrete labels or classes. Also let $\zeta(\mathbf{x})$ be a function that retrieves the corresponding label $C_\mathbf{x}$ of instance $\mathbf{x}$ from training set $\mathcal{T}$ and $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+ \cup \{0\}$ a dissimilarity measure. Given a query instance $\mathbf{a}$, class $C_\mathbf{a}$ is estimated using the 1NN rule[1] as:

$$C_\mathbf{a} = \zeta\left(\arg\min_{\mathbf{x}\in\mathcal{T}} d(\mathbf{x}, \mathbf{a})\right). \tag{3.1}$$

This can be generalized to the kNN rule by finding k neighbors instead and assigning to $C_\mathbf{a}$ the most frequent label among them.

[1] Particular case of kNN when considering $k = 1$.

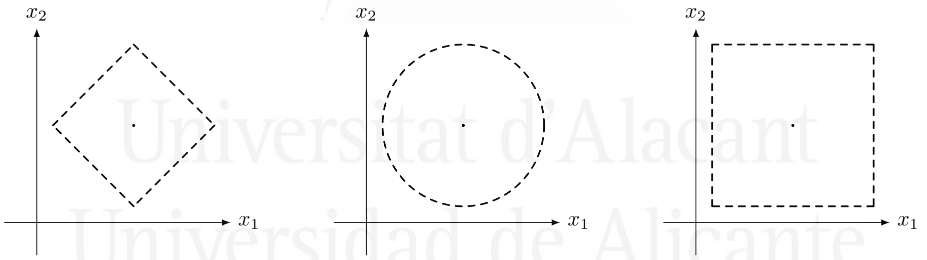
As in any PR scheme, the kNN rule contains a set of parameters to be tuned for the different classification problems, which in this case is two: the number of neighbors k and the dissimilarity measure d. We shall now discuss their implications on the context of this classifier.

On the one hand, the number of neighbors k generally depends on the noise present in the training set T: for hardly noisy sets, the value of k is usually kept low but, as noise figures increase, k is typically raised so as to cope with outliers in the data.

On the other hand, dissimilarity measures are totally dependent on the input representation considered. In this sense, while structural representations show a scarce collection of measures to be considered (e.g., the Edit Distance for strings (Wagner & Fischer, 1974)), in statistical representations we may consider the Minkowski distance:

$$d(\mathbf{x}, \mathbf{a}) = \left(\sum_{i=1}^{|\mathbf{x}|} |x_i - a_i|^p \right)^{1/p} \tag{3.2}$$

being the particular cases of $p = 1$ and $p = 2$ the well-known Manhattan and Euclidean distances, respectively. Figure 3.4 shows some examples of Minkowski dissimilarity measures.



(a) Manhattan distance ($p = 1$).

(b) Euclidean distance ($p = 2$).

(c) Chebyshev distance ($p = \infty$).

Figure 3.4: Examples of Minkowski measures on a two-dimensional space for a fixed distance value (dashed line). Values in parenthesis show the corresponding p exponent of the Minkowski distance.

3.3.1 Limitations

Despite its aforementioned popularity, kNN suffers from several drawbacks that limit its application (García, Derrac, Cano, & Herrera, 2012): (i) as an instance-based classifier, storage memory requirements tend to be high for keeping all training data; (ii) the method shows low computational efficiency as, for each new query, many distance computations are repeated due to the lack of a model; (iii) it may sometimes be sensitive to noisy instances, especially for low k values.

These shortcomings have been widely analyzed in the literature and several strategies have been proposed to tackle them. In general, they can be divided into three categories:

a) Fast Similarity Search: these methods aim at improving the speed issues in kNN with the creation of search indexes for fast consulting in the training set. Some examples of such techniques are the k-dimensional tree (Friedman, Bentley, & Finkel, 1977) or the *Approximating and Eliminating Search Algorithm* (AESA) (Vidal, 1986).

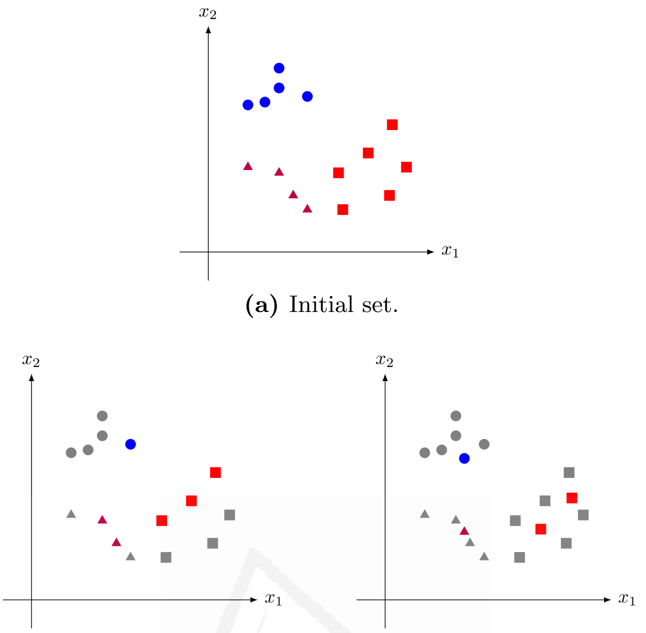
b) Approximated Similarity Search: the main issue with the former strategies is that they do not scale well to high-dimensional spaces (Liu, Moore, Yang, & Gray, 2004); to avoid such issue, approximated similarity search methods work on the premise of searching sufficiently similar prototypes to a given query in the training set instead of retrieving the exact nearest instance. Among the existing proposals, a very successful technique is the Local Sensitive Hashing (Gionis, Indyk, & Motwani, 1999).

c) Data Reduction: opposite to the previous strategies, this set of techniques are commonly considered for kNN to reduce the size of the training set while maintaining, if not improving, the classification accuracy as with the original data (García, Luengo, & Herrera, 2015).

While the two first approaches focus on improving time efficiency, they do not have any consideration towards the reduction of memory consumption or the noise removal, thus limiting their application especially for real-world scenarios. In this sense, the Data Reduction framework rises as a suitable option to consider as it is conceptually capable of tackling all the drawbacks previously introduced by removing both redundant and noisy instances from the initial training set. In fact, the resulting set from such reduction process should, in principle, require a lower k value than the unprocessed initial training set due to the noise removal capabilities of the process (Pekalska, Duin, & Paclík, 2006).

Data Reduction for instance-based classification can be divided into two basic approaches (Nanni & Lumini, 2011): Prototype Generation (PG) and Prototype Selection (PS). The main difference is that the former approach creates new artificial data to replace the initial training set while the latter one simply selects certain elements from that set[2]. Figure 3.5 shows a graphical example of these two families of processes.

[2] In some sense, PS can be seen as a particular case of PG in which the process is constrained to selecting instances rather than creating new ones.



(a) Initial set.

(b) Reduced set after PS.

(c) Reduced set after PG .

Figure 3.5: Comparison between the reduction methodologies of PS and PG. Shaded elements depict instances discarded for the reduced set while the colored ones represent the result of the reduction process.

Finally, as reported in the literature, PG generally leads to better results than PS but such methods also show considerable constraints, especially in terms of the representation used for encoding the data (Calvo-Zaragoza, Valero-Mas, & Rico-Juan, 2016a). Due to this constraint, in this Thesis we focus on the use of PS techniques. These techniques are now thoroughly contextualized in the following section, being a particular methodology for their assessment also introduced.

3.3.2 Prototype Selection (PS)

Given the relevance of PS for kNN, it is possible to find a large number of strategies devoted to perform the aforementioned size reduction and noise removal tasks in the sets. Given that, different authors have proposed taxonomies to group such strategies under different criteria. In this work we rely on the one proposed by García et al. (2012) that divides PS strategies into three groups: Condensing, Editing and Hybrid techniques. These families are now introduced.

The **Condensing** family of strategies is based on the idea of keeping only the most representative prototypes of each class and reducing the size of the set as much as possible. While accuracy on training set is usually maintained, generalization tends to be decreased, especially in noisy scenarios. Some successful examples of such techniques are the Condensed Nearest Neighbor (CNN) by Hart (1968) or its fast and deterministic version, the Fast Condensed Nearest Neighbor (FCNN) by Angiulli (2007).

Methods based on the **Editing** approach focus on eliminating instances that produce some class overlapping, typical situation of elements located close to the decision boundaries or noisy data. For these algorithms, set size reduction is usually lower than the one achieved with Condensing-based strategies but generalization accuracy tends to be higher. The Edited Nearest Neighbor (ENN) by Wilson (1972) constitutes the most representative example of this family, being also the Repeated Edited Nearest Neighbor (RENN) algorithm a commonly-considered variant based on repeating ENN until a certain convergence criterion is achieved.

Hybrid approaches seek for a compromise between Condensing and Editing strategies, that is obtaining the smallest set size while improving, or at least maintaining, the generalization accuracy of the former set. A straight-forward implementation of this idea is found in the Repeated Condensed Nearest Neighbor (RCNN) by Dasarathy, Sánchez, and Townsend (2000), which basically performs an RENN followed by a CNN stage. Also in this category, a very successful approach has been the use of genetic algorithms for accomplishing such objectives, as in the work by Cano, Herrera, and Lozano (2006) with the Cross-generational elitist selection, Heterogeneous recombination and Cataclysmic mutation (CHC) genetic algorithm by Eshelman (1990).

Additionally, a new family of approaches known as **rank methods** has been recently proposed additionally to the commented ones. For such cases, instances of the training set are ordered in terms of their relevance with respect to classification accuracy, which is a score obtained following a particular heuristic. Eventually, prototypes are selected starting from the highest score until a certain point when a certain criterion is accomplished. Examples of such techniques are the NE and the FaN algorithms by Rico-Juan and Iñesta (2012)

Finally, it is important to highlight that, unfortunately, PS methods commonly carry an accuracy loss with respect to directly using the original training set. In this sense, PS methods have been occasionally hybridized with other paradigms to somehow solve those issues. Examples of hybrid schemes may be found in the work by García-Pedrajas and de Haro-García (2014) in which PS was combined with ensemble methods, or the works by Derrac, Cornelis, García, and Herrera (2012); Tsai, Eberle, and Chu (2013) in which feature selection processes were considered.

Multi-objective Optimization Problem (MOP) for Prototype Selection (PS) evaluation

In general, evaluation of PS algorithms is not a trivial issue[3]. Most often, minimization of the set size and maximization of the classification accuracy are opposing goals as improving one of them generally implies a deterioration of the other one.

From this point of view, PS-based classification can be seen as a Multi-objective Optimization Problem (MOP) in which two functions are meant to be optimized at the same time: minimization of prototypes in the training set and maximization of the classification success rate (Calvo-Zaragoza, Valero-Mas, & Rico-Juan, 2015a). The usual way of evaluating this kind of problems is by means of the *non-dominance* concept. One solution is said to dominate another if, and only if, it is better or equal in each goal function and, at least, strictly better in one of them. The set of non-dominated elements, which is known as Pareto frontier, represents the different optimal solutions to the MOP. Each of them is referred to as Pareto-optimal solution, being all of them considered the best without any particular order.

In formal terms, let $\mathcal{U} = \{\mathbf{u}_1, ..., \mathbf{u}_{|\mathcal{U}|}\}$ be the set of solutions of a given MOP. Each solution $\mathbf{u}_i$ of the set is a vector with the results for the M evaluation criteria considered, $\mathbf{u}_i = \left(u_i^{(1)}, ..., u_i^{(M)}\right)$. Out of it, the Pareto-optimal set of solutions $\mathcal{U}^* \subseteq \mathcal{U}$ is obtained as:

$$\mathcal{U}^* = \left\{ \mathbf{u}_i \in \mathcal{U} : u_i^{(m)} \leq u_j^{(m)} \land \exists \, m_o : u_i^{(m_o)} < u_j^{(m_o)} \right\} \tag{3.3}$$

being $i, j \in [1, ..., |\mathcal{U}|]$, $i \neq j$, and $m, m_o \in [1, ..., M]$.

Note that Eq. 3.3 considers the optimization of the functions by seeking for minima. Nevertheless, MOP is totally equivalent when optimization is achieved by seeking for maxima, or even for combinations of both maxima and minima, in the evaluation functions.

As previously described, for the case of PS evaluation, there are $M = 2$ functions to be optimized: set size and classification accuracy, which are optimized by minimizing and maximizing processes, respectively.

Finally, Fig. 3.6 shows a graphical representation of a two-dimensional MOP highlighting the Pareto-optimal set of solutions.

3.4 General discussion

The field of PR has shown large application for generic data analysis and, in particular interest of this Thesis, in MIR. In general, finding patterns in sets of data to initially categorize and then further processing them significantly

[3] This problem also takes place when assessing PG techniques; nevertheless, in this dissertation we focus on the exclusive use of PS methods.

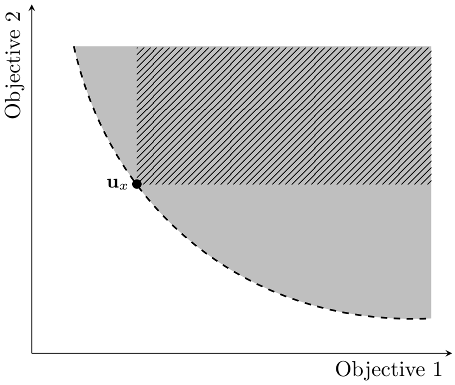


Figure 3.6: Example of a two-dimensional Pareto graph. The global optimal solution is found at the origin of coordinates. Dashed line shows the Pareto frontier with the non-dominated solutions ($\mathcal{U}^*$) whereas the gray area covers the entire set of posible of the problem ($\mathcal{U}$). Additional, the stripped regions covers the area dominated by Pareto-optimal point $\mathbf{u}_x$.

reduces the complexity of the data analysis task. In the case when PR is combined with Machine Learning, this discovery process is remarkably improved as it allows its application to large amounts of data not addressable on a manual basis.

In general, PR systems are not capable of retrieving perfect results, being human post processing hence required to manually correct the errors committed. In this context IPR stands as an appealing alternative to efficiently exploit this human interaction and eventually reduced the workload of a potential user of such systems. Moreover, of particular interest to this work is the case of ISPR as it particularizes the general IPR idea to the case of data following a sequential structure, for which music information constitutes a particular example.

The major limitation in interactive systems is that the user must not perceive any delay in the response of the system. Thus, when a user performs a correction in an IPR system, the core PR model has to be updated as fast as possible, which may not be always possible for very complex models. In this context, instance-based algorithms stand as a remarkably interesting alternative as the model update is done by simply including new instances in the training set. Among them, the one of particular interest for this work in the kNN, mainly because of its fairly conceptual simplicity and the good results typically obtained with it.

CHAPTER 4

Studies on Prototype Selection for the *k*-Nearest Neighbor classifier

"Do not fear mistakes – there are none"

MILES DAVIS

A considerable deal of the work in this dissertation entails the use of supervised classification for time-series analysis. As shall be presented in Chapter 5, we address the estimation of onset events in audio signals as a classification task: each analysis frame of the piece, which typically spans for tens of milliseconds, may be labelled as either containing an onset or its absence. Given that onset events are generally scarce, the onset detection task eventually results in a problem of imbalanced classification in which the non-onset class is remarkably more frequent than the onset one. Such imbalance situations tend to harm the classifier as they bias the performance of the method towards the class representing the majority of examples.

As previously introduced, the use of the *k*-Nearest Neighbor (*k*NN) classifier is of particular interest for this Thesis as it is naturally suited to interactive environments due to being an instance-based method: for adapting its behaviour this classifier simply needs to modify its training set without the further need for a training stage. Nevertheless, *k*NN exhibits low efficiency figures as the number of instances in the training set grows due to the commented lack of generalization model, which is a frequent situation in the aforementioned onset estimation problem.

This chapter presents two studies on Prototype Selection (PS) for tackling the commented issues of low efficiency and imbalanced classification in the context of the *k*NN classifier. The first piece of research focuses on analyzing the rather novelty of the *rank methods* introduced in Chapter 3 and comparing them to the performance of conventional PS strategies. Then, the second part studies the consequences of considering PS in large-scale class-imbalance scenarios, that is, cases in which the number of instances for each class in the dataset is not balanced but there are enough prototypes to require a reduction in the amount of data for improving the efficiency of the kNN rule. Finally, a last section is included to discuss the main ideas gathered from the studies.

For a compact explanation of the methods discussed in this chapter, we shall now introduce some notation. Let $\mathcal{T}$ represent a training set of data and $\zeta(p)$ a function that retrieves the class corresponding to an instance $p \in \mathcal{T}$. We define *friends of a prototype* f_p as the set of instances in $\mathcal{T}$ that share the same class as p, that is $f_p = \{p' \in \mathcal{T} \setminus \{p\} : \zeta(p') = \zeta(p)\}$, being the rest of the set the *enemies of prototype* p, $e_p = \{\mathcal{T} - f_p\} = \{p' \in \mathcal{T} \setminus \{p\} : \zeta(p') \neq \zeta(p)\}$. Consider $d(\cdot, \cdot)$ a dissimilarity function between two data prototypes and $k\mathrm{NN}(p, \mathcal{X}, k)$ the method that retrieves the k closest instances to prototype p in space $\mathcal{X}$. Lastly, we define the nearest enemy to p as its closest prototype with a different class, that is $e_p|_{min} = \arg\min_{p' \in e_p} \; d(p', p)$.

4.1 Experimental study of rank methods for Prototype Selection

This first work performs an experimental study on Prototype Selection (PS) techniques for tackling the low-efficiency issues of the kNN classifier when tackling large-scale data collections such as the onset estimation case introduced. The precise idea of the study is to assess the performance of conventional PS methodologies for large collections of data with a particular emphasis on the so-called *rank methods* introduced in Chapter 3. Given the relative novelty of the latter family of PS algorithms, it is unclear its competitiveness against conventional methodologies. We shall therefore perform a comprehensive experimental study on the performance of rank methods for PS compared to a representative series of conventional PS algorithms selected from the literature in a number of scenarios differing in their size and amount of mislabelled samples (to simulate noisy conditions).

For this study we shall initially introduce the gist of rank methods; then the different conventional PS methods against which rank methods are compared to shall be introduced; after that we shall explain the experimentation scheme proposed, introducing the datasets considered, the algorithms to be compared to, and the evaluation measures; then the results are introduced and commented; finally, a discussion section closes the study.

4.1.1 Introduction to rank methods

The main idea behind rank methods is that instances in the training set are not selected but simply ordered (Valero-Mas, Calvo-Zaragoza, Rico-Juan, & Iñesta, 2016). Following a certain type of heuristic, instances are given a score that indicates its relevance with respect to classification accuracy and ranks them according to this relevance score. Eventually, a selection process is performed by keeping elements starting from the top of the rank (instances with higher relevance) until a certain criterion is accomplished.

A particular approach for rank methods is to follow a voting heuristic in which instances in the training set vote for the rest of the prototypes that help them to be correctly classified. After all instances in the training set have performed this voting process, the score is normalized to produce a relevance rate so that the sum over these rates for all the prototypes of a given class equals the unit. Then, the training set is sorted according to those values and the best candidates are selected until their accumulated score exceeds an external manual parameter $\alpha \in (0, 1]$ that allows the performance of the rank method to be tuned. Low values of this parameter shall lead to a higher reduction of the size of the training set, while high values shall remove just the most irrelevant prototypes. While an external tuning parameter may, in principle, be considered an inconvenient in a data preprocessing framework, it must be pointed out that this characteristic allows the user to enhance a particular objective (either reduction or accuracy) depending on the requirements of the system.

The experimental study presented here focuses on the voting heuristics proposed by Rico-Juan and Iñesta (2012): Farthest Neighbor (FaN) and Nearest to Enemy (NE). Both strategies are based on the aforementioned idea of each instance in the training set voting to the prototype that fulfills a certain criterion, simply differing in the policy for performing such search. These techniques are now introduced.

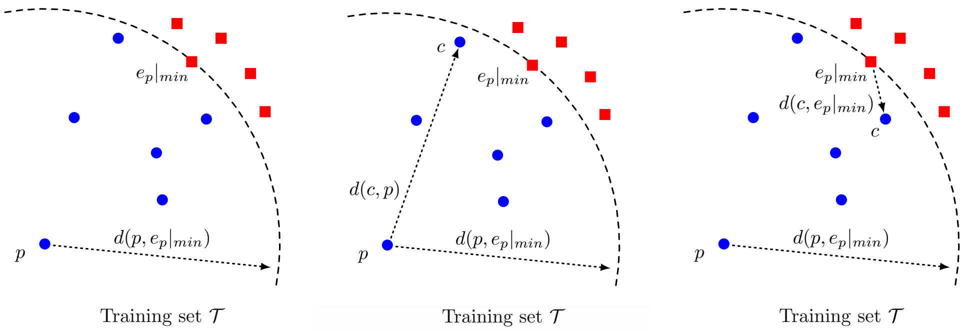
The FaN policy searches for an instance $c \in f_p$ that is the farthest friend of prototype p but closer than nearest enemy $e_p|_{min}$. That is, instance p emits a vote for prototype c that fulfills:

$$c = \arg\max_{p' \in f_p} d(p', p) : d(p', p) < d(p', e_p|_{min}) \ . \tag{4.1}$$

On the other hand, in the NE strategy prototype p votes for an instance $c \in f_p$ that is the closest element to nearest enemy $e_p|_{min}$ with the same class as p. Additionally, this friend must also be within the area centered at p and radius $d(e_p|_{min}, p)$. Formally, p votes to the prototype c that fulfills:

$$c = \arg\min_{p' \in f_p} d(p', e_p|_{min}) : d(p', p) < d(e_p|_{min}, p) \ . \tag{4.2}$$

It is also important to remark that these methods may be extended by letting $e_p|_{min}$ be the n-nearest enemy instead of the first one to reduce the influence of possible outliers in the set. These strategies shall be denoted by n-FaN and n-NE. Figure 4.1 shows a graphical example of the 1-NE and 1-FaN algorithms for PS.



(a) Initial situation.

(b) Farthest Neighbor (FaN).

(c) Nearest to Enemy (NE).

Figure 4.1: Examples of the Farthest Neighbor (FaN) and Nearest to Enemy (NE) schemes (Rico-Juan & Iñesta, 2012) for Prototype Selection on a two-dimensional case: Figure 4.1a shows the nearest enemy $e_p|_{min}$ of prototype p; Figure 4.1b highlights the selected prototype c using FaN; Figure 4.1c depicts the selected prototype c using NE. These examples consider Euclidean distance as dissimilarity measure.

Once the voting stage is finished, a normalization process is applied so that the sum of all votes accumulated by all the prototypes of a particular class sums up to the unit. Finally, the most relevant prototypes are selected using the external α parameter, referred to in these methods as *probability mass*: instances from each class are selected from the top to the bottom of the rank until their accummulated probability exceeds the α parameter. In terms of notation, these parameter is usually shown as a subscript to the actual rank-based PS strategy, that is n-NE$_\alpha$ and n-FaN$_\alpha$.

4.1.2 Conventional Prototype Selection schemes

In terms of conventional PS schemes, we shall consider a set of methodologies that sufficiently cover the different families introduced in Chapter 3.

As of *condensing* schemes, we consider the Condensed Nearest Neighbor (CNN) proposed by Hart (1968). This method is based on the following principle: *(i)* it creates an empty set $\mathcal{T}_{\text{CNN}}$; *(ii)* a prototype p is extracted from $\mathcal{T}$; *(iii)* prototype p is classified with the kNN rule but using set $\mathcal{T}_{\text{CNN}}$; *(iv)* if the estimated class for p mismatches the actual class, p is added to $\mathcal{T}_{\text{CNN}}$; *(v)* the process is repeated from *(ii)* until no more prototypes are left.

The counterpart with CNN is that the reduction in size is not guaranteed as this is highly dependent on the order in which $\mathcal{T}$ is queried. For that, we also include the Fast Condensed Nearest Neighbor (FCNN) by Angiulli (2007) that solves such inconveniences as well as improves its time performance. For this explanation, let $Centroids(\mathcal{T})$ be the set containing the centroids for each class in $\mathcal{T}$. Also consider the Voronoi set as the set of elements of $\mathcal{T}$ that are closer to p than to any other element $p' \in \mathcal{T}_{\text{FCNN}}$, that is $Vor(p, \mathcal{T}_{\text{FCNN}}, \mathcal{T}) = \{p' \in \mathcal{T} : \forall q \in \mathcal{T}_{\text{FCNN}},\ d(p, p') \leq d(q, p')\}$. Additionally, let the so-called *Voronoi enemies* set be $Voren(p, \mathcal{T}_{\text{FCNN}}, \mathcal{T}) = \{p' \in Vor(p, \mathcal{T}_{\text{FCNN}}, \mathcal{T}) : \zeta(p') \neq \zeta(p)\}$. Taking those concepts into consideration, FCNN is described in Algorithm 4.1.

Algorithm 4.1: Description of the Fast Condensed Nearest Neighbor (FCNN).

```
Data: Training set $\mathcal{T}$
Result: Reduced set $\mathcal{T}_{\text{FCNN}}$
$\mathcal{T}_{\text{FCNN}} \leftarrow \emptyset$; $\Delta\mathcal{T} \leftarrow Centroids(\mathcal{T})$ ;
while $\Delta\mathcal{T} \neq \emptyset$ do
    $\mathcal{T}_{\text{FCNN}} \leftarrow \mathcal{T}_{\text{FCNN}} \cup \Delta\mathcal{T}$ ;
    $\Delta\mathcal{T} \leftarrow \emptyset$ ;
    foreach $p \in \mathcal{T}_{FCNN}$ do
        $\Delta\mathcal{T} \leftarrow \Delta\mathcal{T} \cup \{k\text{NN}(p, Voren(p, \mathcal{T}_{\text{FCNN}}, \mathcal{T}), k = 1)\}$ ;
    end
end
```

Regarding *editing* methodologies, we have considered the Edited Nearest Neighbor (ENN) introduced by Wilson (1972). This algorithm is based on the following procedure: *(i)* creates a set $\mathcal{T}_{\text{ENN}}$ equal to the initial set $\mathcal{T}$; *(ii)* a prototype p is extracted from $\mathcal{T}$; *(iii)* prototype p is classified using the kNN rule on set $\mathcal{T}$; *(iv)* if the estimated class for p mismatches the actual class, prototype p is removed from set $\mathcal{T}_{\text{ENN}}$; *(v)* the process is repeated from *(ii)* until all prototypes in $\mathcal{T}$ have been queried.

We have also studied a collection of techniques belonging to the *Hybrid* approaches. The most straight-forward methods are based on combinations of algorithms from the previous families: initially, ENN is applied to set $\mathcal{T}$ and the result is processed using CNN or FCNN, known as Edited Condensed Nearest Neighbor (ECNN) and Edited Fast Condensed Nearest Neighbor (EFCNN) respectively.

In addition to the conventional combinations previously described, we have also considered more sophisticated methodologies. One of them is the Decremental Reduction Optimization Procedure 3 (DROP3) hybrid approach by Wilson and Martinez (2000), which also implements an initial ENN for then performing a condesing-based reduction methodology. This approach is described in Algorithm 4.2.

Another algorithm considered for the comparative study is the Iterative

Algorithm 4.2: Description of the Decremental Reduction Optimization Procedure 3 (DROP3).

```
Data: Training set $\mathcal{T}$
Result: Reduced set $\mathcal{T}_{\text{DROP3}}$
$\mathcal{T}_{\text{DROP3}} \leftarrow$ ENN$(\mathcal{T}, k)$ ;
foreach $p \in \mathcal{T}$ do
    $a(p) = \{p' \in \mathcal{T}_{\text{DROP3}} \setminus p : p \in k\text{NN}(p', \mathcal{T}_{\text{DROP3}}, k)\}$ ;
    $Wi = \#$ of $a_p$ elements classified correctly considering $\mathcal{T}_{\text{DROP3}}$ ;
    $Wo = \#$ of $a_p$ elements classified correctly considering $\mathcal{T}_{\text{DROP3}} \setminus p$ ;
    if $Without > With$ then $\mathcal{T}_{\text{DROP3}} \leftarrow \mathcal{T}_{\text{DROP3}} \setminus p$ ;
end
```

Case Filtering (ICF) proposed by Brighton and Mellish (2002). As in DROP3, this algorithm performs an initial ENN stage and then implements a strategy for data reduction to obtain a more compact set. The explanation of this method is described in Algorithm 4.3. For understanding it, consider $L(p) = \{p' \in T : d(p', p) < e_p|_{min}\}$ the set that contains all instances inside the largest hypersphere around prototype p containing only instances with the same class as p.

Algorithm 4.3: Description of the Iterative Case Filtering (ICF).

```
Data: Training set $\mathcal{T}$
Result: Reduced set $\mathcal{T}_{\text{ICF}}$
$\mathcal{T}_{\text{ICF}} \leftarrow$ ENN$(\mathcal{T}, k)$ ;
do
    foreach $p \in \mathcal{T}_{ICF}$ do
        $C(p) = \{p' \in \mathcal{T}_{\text{ICF}} : p \in L(p')\}$ ;
        $R(p) = \{p' \in \mathcal{T}_{\text{ICF}} : p' \in L(p)\}$ ;
    end
    $progress$ = false ;
    foreach $p \in \mathcal{T}_{ICF}$ do
        if $|R(p)| > |C(p)|$ then
            $\mathcal{T}_{\text{ICF}} \leftarrow \mathcal{T}_{\text{ICF}} \setminus p$ ;
            $progress$ = true ;
        end
    end
while $progress$;
```

Finally, due to its considerable good results reported in the literature, we have also considered the use of the Cross-generational elitist selection, Heterogeneous recombination and Cataclysmic mutation (CHC) genetic algorithm applied to PS as in the work by Cano et al. (2006). This algorithm obtains a reduced set $\mathcal{T}_{\text{CHC}} \subseteq \mathcal{T}$ by following the following steps, which are iteratively repeated for a series of generations fixed by the user: *(i)* an initial population of N instances is selected; *(ii)* then, the N individuals are randomly paired and used to generate N potential offspring; *(iii)* finally, a survival stage is held and the best N chromosomes from the parent and offspring populations are selected for the next generation. For our experiments, we have considered the same parameters as in the work by Cano et al. (2006), which consists of $10,000$ generations with populations of $N = 50$ elements.

4.1.3 Experimentation

Regarding the experimental set-up, five multiclass corpora were considered for the experiments: the *National Institute of Standards and Technology Special Database 3* (NIST3) of handwritten characters (Wilkinson et al., 1992), from which a subset of the upper case characters was randomly selected; the *United States Postal Office* (USPS) handwritten digits dataset (Hull, 1994); the *Handwritten Online Musical Symbols* (HOMUS) dataset (Calvo-Zaragoza & Oncina, 2014) with images of handwritten isolated music figures; and two additional corpora of the UCI collection (Lichman, 2013), the *Penbased* compilation of handwritten isolated digits and the *Letter* collection of handwritten capital characters from the English language. A 4-fold cross validation scheme over each dataset was performed.

For the NIST3 and USPS cases, contour descriptions with Freeman Chain Codes (FCC) (Freeman, 1961) were extracted and the Edit Distance (ED) (Wagner & Fischer, 1974) was used as dissimilarity measure. In the case of the HOMUS set, Dynamic Time Warping (DTW) (Sakoe & Chiba, 1990) was used due to its good results in the baseline experimentation. Since datasets from the UCI may contain missing values in the samples, the Heterogeneous Value Difference Metric (HVDM) (Wilson & Martinez, 1997) was used for the two datasets considered from this family. Table 4.1 shows a summary of the main features of these collections.

Table 4.1: Description of the datasets used the experimentation for the assessment of ranking-based methods for Prototype Selection.

Name	Instances	Classes	Dissimilarity
USPS	9,298	10	ED
NIST3	6,500	26	ED
HOMUS	15,200	32	DTW
Penbased	10,992	10	HVDM
Letter	20,000	26	HVDM

Additionally, in order to test the robustness of PS methods, synthetic noise was artificially induced in the datasets by swapping labels between pairs of instances randomly chosen from the training set. The noise rates (percentage of prototypes that change their label) considered were 0 %, 20 %, and 40 % as they constitute typical values in this kind of experimentation (Natarajan, Dhillon, Ravikumar, & Tewari, 2013).

For the comparative we have the conventional PS strategies explained is Section 4.1.2 and we also included the ALL case in which no PS process is performed and the initial training set is entirely kept.

The configurations tested for the rank-based strategies have been 1-FaN, 2-FaN, 1-NE, and 2-NE. For each of them values of α within the range $(0, 1)$ with a granularity of 0.1 were considered, being extreme values discarded since $\alpha = 0$ would mean an empty set and $\alpha = 1$ is equivalent to not performing any selection process (equivalent to the ALL case).

Finally, the evaluation of the results is performed by assessing the classification accuracy achieved after the PS process as well as the size of the resulting set. Additionally, the Multi-objective Optimization Problem criterion introduced in Chapter 3 is considered to assess both accuracy and reduction figures under a single optimization premise.

4.1.4 Results

Table 4.2 shows the results obtained for the experimentation. Each figure represents the arithmetic mean of the accuracy and set size values obtained for the considered datasets. Bold values represent the non-dominated solutions, which can be graphically seen in Figs. 4.2 and 4.3 for the different induced noise cases considered and shall be later discussed in the Multi-objective Optimization Problem assessment part.

Non-induce noise scenario

Let us pay attention first to the case when no induced noise is considered. It can be observed that, when no information was discarded (ALL scheme), conventional kNN achieved some of the highest accuracy values for all k configurations. Note that increasing this k parameter did not have any noticeable effect. Given that the datasets considered are hardly noisy, the ENN algorithm did not significantly reduce the size of the set (a reduction rate around 10 %), maintaining similar accuracies to those achieved by the conventional kNN strategies.

On the other side, the condensing family of algorithms (CNN and its extensions) showed some remarkable results: all of them achieved great reduction rates, especially ECNN and EFCNN, which simply required around a 10 % of the set size, and performed well in terms of accuracy (only around 3 % lower than the ALL configurations).

DROP3 also achieved high reduction rates (around 9 % of the maximum

Table 4.2: Average figures of the results obtained for the assessment of rank-based methods for Prototype Selection in terms of their achieved classification accuracy (Acc) and resulting set size (Size), in percentage. Bold values represent the non-dominated elements defining the Pareto frontier.

PS strategy	Noise 0 %		Noise 20 %		Noise 40 %		PS strategy	Noise 0 %		Noise 20 %		Noise 40 %	
	Acc	Size	Acc	Size	Acc	Size		Acc	Size	Acc	Size	Acc	Size
ALL ($k = 1$)	93.4	100	76.3	100	63.3	100	ALL ($k = 3$)	93.5	100	86.3	100	75.7	100
ALL ($k = 5$)	93.4	100	90.9	100	86.1	100	**ALL (k = 7)**	93.1	100	**91.5**	**100**	**89.0**	**100**
CNN	90.3	18.0	67.8	57.3	55.7	72.6	**ENN**	92.3	93.3	**91.0**	**67.1**	**88.4**	**48.7**
FCNN	90.4	17.7	67.5	55.1	55.5	71.2	**ECNN**	**90.0**	**10.4**	87.6	9.0	84.4	8.5
EFCNN	90.1	10.5	**88.0**	**8.9**	**84.3**	**8.1**	DROP3	84.6	9.5	74.4	9.9	63.5	10.7
ICF	77.3	15.3	68.2	17.1	59.0	18.4	**CHC**	**84.4**	**3.1**	**71.5**	**2.6**	**60.2**	**2.3**
1-FaN$_{0.10}$	80.8	3.6	83.1	4.2	**83.5**	**4.9**	**1-NE$_{0.10}$**	**71.7**	**1.3**	**81.6**	**3.3**	**83.4**	**4.4**
1-FaN$_{0.20}$	**86.2**	**8.3**	87.1	10.0	85.7	11.5	**1-NE$_{0.20}$**	79.9	3.3	**86.6**	**8.1**	86.0	10.7
1-FaN$_{0.30}$	88.5	14.2	88.3	16.8	81.7	19.3	**1-NE$_{0.30}$**	**85.3**	**6.4**	**88.6**	**14.4**	82.3	18.4
1-FaN$_{0.40}$	90.1	20.3	86.0	24.9	73.4	29.3	1-NE$_{0.40}$	89.1	10.7	86.9	22.4	75.0	28.4
1-FaN$_{0.50}$	91.3	28.3	80.4	34.9	68.1	39.3	**1-NE$_{0.50}$**	**91.3**	**17.3**	80.8	32.3	68.8	38.4
1-FaN$_{0.60}$	92.0	38.3	76.9	44.9	64.1	49.2	**1-NE$_{0.60}$**	**92.2**	**27.8**	76.7	42.2	64.5	48.3
1-FaN$_{0.70}$	92.6	48.4	75.2	54.9	62.9	59.2	1-NE$_{0.70}$	92.8	41.8	74.5	52.2	62.8	58.3
1-FaN$_{0.80}$	93.0	60.6	73.3	64.9	60.0	69.2	**1-NE$_{0.80}$**	**93.2**	**60.2**	72.7	62.6	60.0	68.4
1-FaN$_{0.90}$	93.4	80.1	74.0	80.1	59.8	80.5	**1-NE$_{0.90}$**	**93.5**	**80.1**	74.3	80.1	60.0	80.3
2-FaN$_{0.10}$	80.4	3.6	**82.9**	**3.9**	83.4	4.3	**2-NE$_{0.10}$**	71.2	1.3	**80.5**	**2.7**	**82.9**	**3.6**
2-FaN$_{0.20}$	85.7	8.2	86.8	9.1	85.3	10.3	**2-NE$_{0.20}$**	79.6	3.3	**86.0**	**6.8**	**86.1**	**9.0**
2-FaN$_{0.30}$	88.2	14.0	87.8	15.7	84.6	17.3	2-NE$_{0.30}$	84.7	6.2	88.0	12.3	85.3	15.8
2-FaN$_{0.40}$	89.8	19.8	86.8	22.8	76.8	26.6	2-NE$_{0.40}$	88.7	10.4	88.3	19.3	77.7	25.0
2-FaN$_{0.50}$	90.9	27.5	80.3	32.7	68.1	36.7	**2-NE$_{0.50}$**	**91.0**	**16.4**	81.4	28.9	68.8	35.0
2-FaN$_{0.60}$	91.7	37.5	75.9	42.7	63.6	46.6	**2-NE$_{0.60}$**	**92.1**	**26.9**	75.8	38.8	63.8	44.9
2-FaN$_{0.70}$	92.4	47.9	73.3	52.7	61.2	56.6	**2-NE$_{0.70}$**	**92.7**	**41.2**	72.0	48.8	60.6	54.9
2-FaN$_{0.80}$	93.0	60.4	71.2	62.8	58.1	66.6	2-NE$_{0.80}$	93.2	60.1	70.9	60.7	57.5	64.9
2-FaN$_{0.90}$	93.4	80.1	73.5	80.1	59.0	80.1	2-NE$_{0.90}$	93.5	80.1	74.0	80.1	59.3	80.1

size), but with a significant drop in accuracy when compared to the conventional kNN algorithm (decreased around 10 % with respect to the scores in the ALL cases). ICF, however, achieved neither a high reduction nor a remarkable accuracy. The CHC evolutionary algorithm obtained one of the highest reduction rates, as it only required around a 3 % of the total amount of prototypes. The accuracy achieved, although lower than in most of the previous cases, was close to an 84 %, which is a good result given the high data reduction performed.

The NE and FaN rank methods showed a very interesting behavior. When considering their probability mass parameter $\alpha \leq 0.5$, the reduction figures obtained covered a similar range to the reductions obtained with the other strategies: for instance, 1-NE$_{0.20}$ achieved a similar reduction to CHC (around 3 % of the initial set size) or 1-FaN$_{0.40}$ is comparable to FCNN (approximately, 20 % of the total amount of prototypes). As it can be seen, these configurations can produce an *aggressive* reduction in the set size, which is often paired with a substantial accuracy loss (e.g., 2-NE$_{0.10}$ which reduces the set to approximately 1 % of its size achieving an accuracy figure around 70 %). However, more *conservative* configurations such as when considering $\alpha = 0.5$ achieved results quite close to the ALL case, with around a third or a fourth of the total number of prototypes.

When considering $\alpha > 0.5$, these methods progressively tend to the ALL case as they also include prototypes located at the lowest positions of the rank (*i.e.*, the ones with the least number of votes). This increase in the reduced set size (up to an 80 % of the complete set size when $\alpha = 0.9$) did not carry a remarkable accuracy improvement (less than a 3 % of improvement with respect to the $\alpha = 0.5$ cases). Nevertheless, it should be noted that the 1-NE$_{0.90}$ improved the accuracy of the ALL case with 80 % of the initial set size, possibly because the method discarded noisy instances in the datasets.

In summary, rank methods proved their capability of producing a good trade-off between reduction and classification accuracy in terms of their reduction parameter α. This way, the user is able to tune the reduction degree prioritizing either accuracy or reduction depending on the particular requirements of the application.

Induced noise scenario

The following lines present the analysis of the performance when noise is induced in the set. As results show qualitatively similar trends, remarks will not focus on a particular noise configuration but on the general behavior.

The mislabelling noise in the samples dramatically changed the previous situation. Accuracy results for conventional kNN suffered an important drop as noise figures raised. Nevertheless, the use of different k values palliated this effect and improved the accuracy rates. Especially remarkable is the $k = 7$ case in which kNN scored the maximum classification rate compared to the other schemes in both noisy configurations considered.

ENN algorithms proved their robustness in these noisy environments, as their classification rates were always among the best results obtained. Moreover, the reduction rates achieved were higher than in the noiseless scenario, since the prototypes these approaches remove are the ones actually producing class overlapping.

Results with CNN and FCNN schemes depicted their sensitiveness to noise as they obtained some of the worst accuracies in these experiments. Due to the impossibility of discarding noisy elements, the reduction is not properly performed, leading to a situation in which there is neither an important size reduction nor a remarkable performance. Furthermore, the use of different k values did not upturn the accuracy results.

EFCNN and ECNN, on the contrary, were less affected than CNN and FCNN due to the introduction of the editing phase in the process. This improvement is quite noticeable as, while the latter approaches obtained accuracy rates of around 50 % and 60 % with a reduction rate between 50 % and 70 %, the former algorithms achieved precision rates over 80 % with roughly 10 % of the prototypes.

Hybrid algorithms DROP3 and ICF, just like the CNN and FCNN approaches, were not capable of coping with noisy situations either. Accuracy rates obtained were quite poor as, for instance, the case of the ICF method with a 40 % of synthetic noise was not able to reach a 60 % of accuracy. However, it must be pointed out that, despite achieving similar accuracy rates, hybrid algorithms still showed better reduction figures than the CNN and FCNN strategies. For example, for an induced noise rate of 40 %, CNN obtained an accuracy of 55.7 % with 72.6 % of reduction while DROP3 achieved 63.5 % with only a 10.7 % of prototypes.

Results obtained with the CHC evolutionary scheme showed its relative sensitivity to noise. In these noisy scenarios, although it still depicted one of the highest reduction figures amongst the compared methods with rates around 2 %, its classification performance was significantly affected as no result was hardly higher than 70 %.

The NE and FaN rank-based methods demonstrated to be interesting algorithms in the noiseless scenario: for low α values, the reduction rates achieved, together with the high accuracy scores obtained, are very competitive against other methods; at the same time, high α values achieved accuracy figures comparable to, or even higher than, the ALL case with just 20 % to 40 % of the initial amount of prototypes. Results in the proposed noisy situations reinforce these remarks for the former case: on average, none of these algorithms showed accuracy rates lower than 80 % while, at the same time, the number of distances computed never exceeded the 20 % of the maximum. It is also important to point out that, while ECNN and EFCNN schemes also showed a remarkable reduction rate with good accuracy figures, these approaches internally incorporate an editing process for tackling the noise in the data, whereas the rank methods depicted a clear robustness to these situations by themselves, as long as α remains low. Nevertheless, if α is increased, the accuracy of these methods noticeably lowers since the algorithm is forced to include all prototypes in the computed rank, which progressively leads to the ALL case. In such situation, the 1NN search is not able to cope with the noise, resulting in the low accuracy figures obtained.

Multi-objective Optimization Problem assessment

In addition to the commented results, we now tackle the PS-based classification from the point of view of a Multi-objective Optimization Problem problem. Considering the case with no induced noise (Fig. 4.2), the solution portraying the maximum accuracy result with the least number of prototypes is the 1-$\text{NE}_{0.90}$, defining the right-hand end of the Pareto frontier[1]. The ALL and ENN configurations do not belong to this frontier as, although they achieved roughly the same accuracy as the previous method, they required a larger amount of prototypes. The rest of the solutions, in spite of exhibiting lower accuracy results, in some cases the loss was not so accused. Examples of this behavior can be checked in the non-dominated algorithm EFCNN, which achieved accuracy results around 3 % lower than the maximum, computing roughly a fifth of the maximum number of distances, respectively. Regarding the proposed rank methods (in red), it can be observed that in the non-dominated frontier, in the region of up to 20 % of the total of distances (the one in which most of the PS algorithms studied lie) there is a clear balance between them and the rest of the strategies. This proves the competitiveness of these methods with respect to other classic strategies. Additionally, rank methods also cover the region above the 20 % of distances since the probability mass α allows the selection of the amount of prototypes to maintain.

With respect to the datasets with induced noise (see Fig. 4.3), the first difference is that the ALL case (with $k = 7$) belongs to the Pareto frontier for both noise figures considered. However, other schemes were equally capable of achieving the same accuracy with a lower computational cost. For instance, when considering the case of inducing 40 % of noise in the datasets, both the 7NN and ENN configurations achieved very similar accuracies but, while the former method requires the computation of all the distances, the latter requires less than a half of them.

Rank methods depicted remarkable compromises between accuracy and number of prototypes when considering low α values. An important number of configurations proved to be capable of dealing with these noise figures since they constituted part of the non-dominance frontier. For instance, in

[1] As a reminder, the Pareto frontier represents the set of optimal solutions to the Multi-objective Optimization Problem. Reader is referred to Chapter 3 for a formal description of this concept.

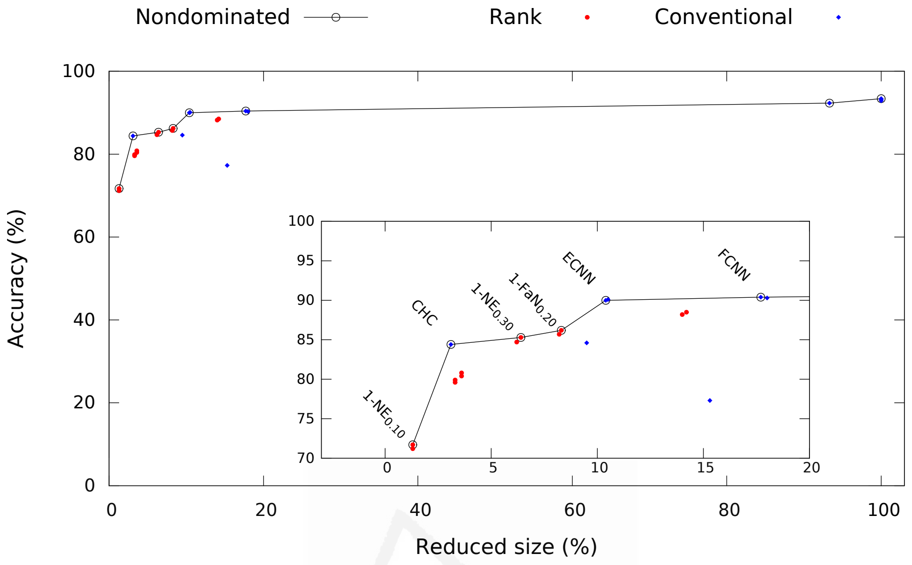


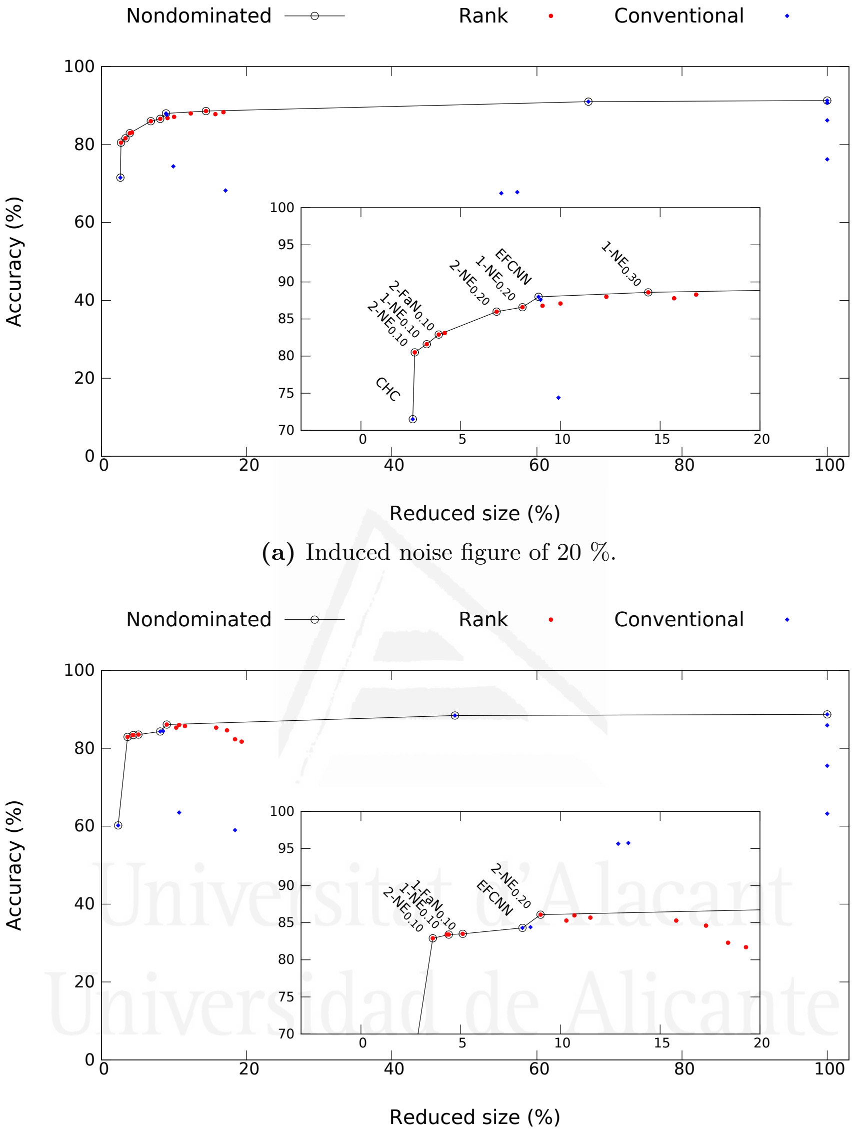
Figure 4.2: Graphical representation of the results obtained for the experimental comparative of rank methods against conventional Prototype Selection techniques. No noise is considered. Circled symbols remark the non-dominated elements defining the Pareto frontier. The area with the largest number of points has been enlarged for its better comprehension.

the 20 % of noise situation, the 1-NE$_{0.30}$ configuration only differed in a 3 % of accuracy with respect the maximum (given by 7NN) but computes roughly a 15 % of the total amount of distances. However, when setting α to a high value, accuracy was noticeably affected since the algorithms were forced to include noisy prototypes with fewer votes located at the lower parts of the rank. In this case, points moved away from the Pareto frontier, proving not to be interesting configurations for such amount of noise.

4.1.5 Discussion

This comparative study of rank-based and conventional PS methods points out several interesting insights to remark. A first one is that, in scenarios without any artificial noise, rank methods are able to achieve considerably small set sizes comparable to the ones obtained by more sophisticated techniques, as for instance the CHC, algorithm without much accuracy loss.

When noise is induced in the sets, these rank methods seem to properly manage the confusion introduced by the mislabelling of the samples, as opposed to other techniques such as CHC or DROP3. More precisely, when rank methods are fixed to a low probability mass (that is, keeping a reduced set size), the samples that are removed from the set are typically the mislabelled ones, probably due to being the ones receiving a low number of



(a) Induced noise figure of 20 %.

(b) Induced noise figure of 40 %.

Figure 4.3: Graphical representation of the results obtained for the experimental comparative of rank methods against conventional Prototype Selection techniques in noisy conditions. Circled symbols remark the non-dominated elements defining the Pareto frontier. The area with the largest number of points has been enlarged for its better comprehension.

votes. This particularity is especially interesting since these rank methods do not incorporate an ENN stage, as opposed to other conventional (most often from the hybrid family) schemes, thus stating their robustness for such scenarios.

Finally, it is also important to mention that, while the probability mass parameter may be seen as a drawback as there is a need to tune it for the particular application at issue, it is actually the opposite. This feature is not typically found in conventional PS schemes. However, it allows the user to look for the proper compromise between accuracy and size for each situation, that is boosting one at the expense of the other depending on the requirements of the system.

4.2 Prototype Selection in large-scale imbalanced binary classification problems

Most standard classification algorithms assume that the classes of the data at issue are equally represented (He & Garcia, 2009). However, this assumption turns out not to be realistic since most data sources do not necessarily exhibit such equilibrium among the different classes. This issue is typically known as the *class imbalance* problem (García, Sánchez, & Mollineda, 2007) and generally results in a bias in the performance of the classifier towards the class representing the majority of the elements (López, Fernández, García, Palade, & Herrera, 2013).

The imbalance issue takes special relevance in the context of this Thesis. As introduced, the estimation of onset points in audio streams may be addressed as an imbalanced classification task in which each analysis frame of the piece may be labelled as either containing an onset (minority class) or its absence (majority class). Since this task requires a temporal resolution of around tens of milliseconds, the analysis of pieces spanning for several minutes entails dealing with thousands of instances. Therefore, it seems interesting to study the issue of classification tasks in imbalanced and large-scale scenarios in the context of the kNN rule.

Prototype Selection (PS) schemes for kNN, as one of the tools for improving the efficiency of this classifier, do not generally consider class-imbalance situations in the set to be reduced and thus the performance of such schemes in large-scale imbalance contexts remains unexplored. Nevertheless, for cases as the one above, there is a need for exploring the performance of PS schemes in such imbalance contexts for large-scale sets that require of a reduction process and compare them to the case in which a preprocessing strategy to deal with imbalance situations is applied.

For this study we shall initially describe the general issue of classification in imbalance scenarios; then, we shall introduce the set of class-balancing strategies considered for this study; after that, the experimental scheme proposed for the assessment of PS techniques in such scenarios is described; afterwards, the results of the experimentation are introduced and analyzed; finally, a brief discussion with the main insights obtained is presented to close the study.

4.2.1 Classification with imbalanced data

Formally, imbalanced classification refers to the cases in which the prior probabilities of the classes at issue significantly differ among them. This particularity generally results in a tendency of the classifier to bias towards the majority class, thus decreasing the overall performance of the system.

Different proposals may be found in the literature to palliate this issue, being typically grouped into three categories (García, Sánchez, & Mollineda, 2012): *(i)* data-level methods that either create artificial data for the minority classes and/or remove elements from the majority one to equilibrate the class representation; *(ii)* algorithmic-level approaches that internally bias the classifier to compensate the skewness in the data; *(iii)* cost-sensitive training methodologies that consider higher penalties for the misclassification of the minority class than for the majority one.

In general, instance-based algorithms such as kNN report a superior tolerance to such imbalance situation as they consider all instances during the classification stage. Nevertheless, when this imbalance effect is combined with class overlapping, performance is severely affected (Fernández, García, & Herrera, 2011).

While PS methods tackle the well-known issues of kNN for large and noisy (overlapped) datasets, these processes have not been devised for class-imbalanced sets. Thus, it seems interesting to explore the performance of PS algorithms in both balanced and imbalanced datasets. To model these two situations we shall consider a collection of imbalanced and overlapped data collections and apply to them a series of data-level balancing methods to test the performance of PS schemes on both situations. Note that we discard the use of any other class-balancing approach as data-level methods are the only ones that do not require to modify the PS algorithm itself.

4.2.2 Data-level balancing techniques

Data-level balancing methods equilibrate the class distribution by *oversampling* the minority class and/or *undersampling* the majority one. To assess their relevance in the context of this experiment, we considered a set of methods of each of the two paradigms as well as combinations of them. For a clear description of these techniques, let $\mathcal{T}_{\text{MAJ}}$ and $\mathcal{T}_{\text{MIN}}$ be the sets containing all the instances from the majority and minority classes of an initial set $\mathcal{T}$, respectively.

As of oversampling techniques, due to being one of the most conventional and widely considered balancing methods, we have considered the Synthetic Minority Over-sampling Technique (SMOTE) algorithm by Chawla, Bowyer, Hall, and Kegelmeyer (2002). In essence, this technique populates the minority class $\mathcal{T}_{\text{MIN}}$ by creating new instances in the space between pairs of instances from that minority subset. This technique is described in

Algorithm 4.4.

Algorithm 4.4: Description of the Synthetic Minority Over-sampling Technique (SMOTE) algorithm.

```
   Data: Training set $\mathcal{T}$, number of new instances $N$ per old one
   Result: Balanced set $\mathcal{T}_{\text{SMOTE}}$
1  $\mathcal{T}_{\text{SMOTE}} \leftarrow \mathcal{T}$ ;
2  foreach $p \in \mathcal{T}_{MIN}$ do
3      $\mathcal{T}_{\text{kNN}} \leftarrow k\text{NN}(p, \mathcal{T}_{\text{MIN}}, k)$ ;
4      for $i$=1 to $N$ do
5          $nn \leftarrow \text{random-select}(\mathcal{T}_{\text{kNN}})$ ;
6          for $j$=1 to $\#attributes\ in\ p$ do
7              $dif = p[j] - nn[j]$ ;
8              $gap = random(0, 1)$ ;
9              $p'[j] = p[j] + gap \cdot dif$ ;
10         end
11         $\zeta(p') \leftarrow \text{Minority class}$ ;
12         $\mathcal{T}_{\text{SMOTE}} \leftarrow \mathcal{T}_{\text{SMOTE}} \cup p'$ ;
13     end
14 end
```

Han, Wang, and Mao (2005) proposed two extensions to the original SMOTE algorithm. Instead of a general populating approach, these extensions focus on detecting and remarking transition zones between classes. For that, these methods initially obtain a set $\mathcal{T}_{\text{DANGER}}$ following this process: *(i)* a prototype p is extracted from the set of data containing the minority class $\mathcal{T}_{\text{MIN}}$; *(ii)* the kNN rule is applied to p in the entire train set $\mathcal{T}$; *(iii)* if more than a half of the neighbors belong to the set of the majority class $\mathcal{T}_{\text{MAJ}}$, instance p is included in $\mathcal{T}_{\text{DANGER}}$; *(iv)* the process is repeated until all prototypes in $\mathcal{T}_{\text{MIN}}$ have been queried.

Once $\mathcal{T}_{\text{DANGER}}$ has been obtained, the SMOTE extensions can be clearly explained. On the one hand, Borderline 1 (B1) populates the minority class using the same process as SMOTE but exchanging the set $\mathcal{T}_{\text{MIN}}$ in Line 2 of Algorithm 4.4 by set $\mathcal{T}_{\text{DANGER}}$. Borderline 2 (B2), on the other hand, maintains the idea of B1 but additionally generates instances for pairs of prototypes from sets $\mathcal{T}_{\text{DANGER}}$ and $\mathcal{T}_{\text{MAJ}}$, always populating the region closer to the instance representing the minority class.

In terms of undersampling techniques, a straightforward yet effective technique is the CNN algorithm for PS adapted by Kubat and Matwin (1997) to be used as a balancing technique. The idea is totally equivalent to the CNN algorithm in PS but with the particularity of initializing the target set $\mathcal{T}_{\text{CNN}}$ with all the instances of the minority class, that is $\mathcal{T}_{\text{CNN}} \leftarrow \mathcal{T}_{\text{MIN}}$, instead of considering an empty initial set.

Just as with the previous CNN case, the ENN method for PS also has its analogue for class-balancing tasks, which was proposed by Laurikkala (2001) under the name of Neighborhood Cleaning Rule (NCL). For undersampling the set, this method considers the following procedure: *(i)* creates a set $\mathcal{T}_{\text{NCL}}$ equal to the initial set $\mathcal{T}$; *(ii)* a prototype p is extracted from $\mathcal{T}$ and its class is estimated using the kNN rule in set $\mathcal{T}$; *(iii)* if p is misclassified, then its actual class is checked; *(iv)* if p belongs to the majority class, p is removed from $\mathcal{T}_{\text{NCL}}$; *(v)* if p belongs to the minority class, all its neighbors belonging to the majority class are removed from $\mathcal{T}_{\text{NCL}}$; *(vi)* the process is repeated until all prototypes have been queried.

Another method we considered for balancing through undersampling is based on the Tomek Links (TL) proposed by Tomek (1976). A TL is a pair of prototypes that are the closest to each other but with different class labels. Because of that, the set $\mathcal{T}_{\text{TL}}$ of all TL defines the decision frontiers of the different classes in the training set $\mathcal{T}$. Mathematically, two prototypes $p \in \mathcal{T}_{\text{MAJ}}$ and $p' \in \mathcal{T}_{\text{MIN}}$ form a link $(p, p') \in \mathcal{T}_{\text{TL}}$ if and only if:

$$\underset{a \in \mathcal{T}_{\text{MAJ}}}{\arg\min}\; d(a, p') = p \;\wedge\; \underset{b \in \mathcal{T}_{\text{MIN}}}{\arg\min}\; d(b, p) = p' \;. \tag{4.3}$$

Eventually, this can be used for obtaining a balanced set $\mathcal{T}_{\text{TOMEK}}$ by removing only the elements of the majority class in $\mathcal{T}_{\text{TL}}$ from the initial set $\mathcal{T}$. Mathematically, this is done as $\mathcal{T}_{\text{TOMEK}} = \mathcal{T} \setminus \{\mathcal{T}_{\text{MAJ}} \cap \mathcal{T}_{\text{TL}}\}$.

It must be mentioned that, while both oversampling and undersampling techniques aim at balancing the class distributions, the latter ones cannot guarantee that the resulting set has the exact number of instances for all classes. Thus, we have also included combinations of the undersampling and oversampling methods previously described, done in that precise order, to test their influence in the experiment.

4.2.3 Experimentation

Figure 4.4 shows the scheme implemented for the experiments. The basic idea is that the *train set* may undergo a class-balancing process and/or a PS method before getting to the kNN classifier, which are the situations to be compared. For our experiments, we fixed a value of $k = 1$ for the kNN stage as well as the Euclidean distance as dissimilarity measure.

For the experimentation we have considered three datasets from the UJI[2] repository (*scrapie*, *spam*, and *phoneme*) and two from the KEEL[3] collection (*segment0* and *yeast3*). Additionally, we have considered the music dataset *prosemus*[4] meant for onset detection, whose features have been extracted with the methodology in Valero-Mas, Iñesta, and Pérez-Sancho

[2] http://www.vision.uji.es/∼sanchez/Databases/

[3] http://sci2s.ugr.es/keel/datasets.php

[4] http://grfia.dlsi.ua.es/cm/projects/prosemus/database.php

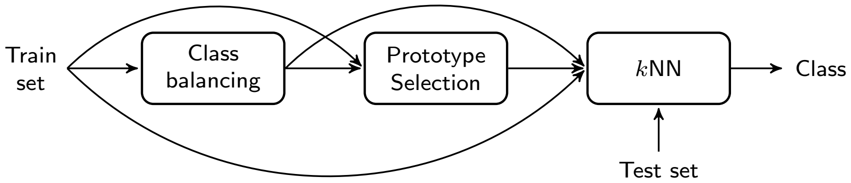


Figure 4.4: Scheme proposed for the assessment of Prototype Selection schemes in imbalanced scenarios.

Table 4.3: Description of the datasets considered for the experimentation of Prototype Selection in imbalanced classification environments in terms of the amount of instances of the majority (Maj.) and minority (Min.) classes.

Dataset	Min.	Maj.	Dataset	Min.	Maj.
prosemus	1,041	4,045	phoneme	3,673	5,170
scrapie	531	2,582	segment0	329	1,979
spam	1,813	2,788	yeast3	163	1,321

(2014)[5]. All these datasets only contain two classes as it constitutes a common practice in studies about imbalanced classification and all of them show a statistical/feature-based representation. Table 4.3 describes them in terms of the number of instances for each class. Also note that, except for *yeast3*, all these sets contain more than 2,000 instances, which constitutes a typical size threshold for which PS is considered to be necessary (García et al., 2012). For all these sets, a 5-fold cross-validation scheme has been considered.

In terms of PS methods, we have contemplated a representative selection of the conventional schemes described in Section 4.1.2, more precisely the condensing-based schemes CNN and FCNN, the ENN algorithm, and the hybrid approaches EFCNN, DROP3, and CHC. Additionally, we have also considered the rank-based strategies 1-NE and 1-FaN. For all these methods we fixed a value of $k = 5$ so that noise present in the set can be managed as well as a low probability mass of 0.1 for the rank methods to obtain a very compact set. Additionally, we contemplate the case in which no PS process is applied with the ALL case.

As of balancing techniques, we have considered the strategies explained in Section 4.2.2. For all of them we fixed a value of $k = 5$ to be robust against the intrinsic noise that may be present in the sets.

Regarding figures of merit, we considered the F-measure (F_1) as it constitutes a typical measure in the context of imbalanced classification. Focusing on the minority class, this metric summarizes the correctly classified elements (True Positive (TP)), the misclassified elements from the majority class as

[5] The methodology for extracting the features is also described in Chapter 5. Nevertheless, consider it for this case to be a generic set of two-class imbalance data.

minority ones (False Positive (FP)), and the misclassified elements from the minority class as majority class (False Negative (FN)) in a single value using Eq. 2.8. Note that for the case of the *prosemus* set this classification task is actually an onset detection process. Thus, for this particular set we have considered the common evaluation procedure for onset detection described in Chapter 2. Finally, for all cases we also consider the Multi-objective Optimization Problem-based evaluation criterion for PS to assess the results in terms of the non-dominance criterion.

4.2.4 Results

The results obtained are shown in Table 4.4. These figures depict the average F_1 score and reduction rate (in percentage) obtained for the considered datasets in terms of the balancing techniques and PS strategy used.

According to the results, the use of PS on the initial imbalance situation implies a decrease in the F_1 measure for all cases. For instance, CHC lowers performance from the average score of $F_1 = 0.69$ in the ALL case to an $F_1 = 0.52$ of the reduced case. In this context of applying PS to an imbalanced set, the results achieved by FCNN are the ones of particular interest since, although there is a decrease in the F_1 value as with the other PS cases, this score is just slightly lower than the original case (0.02 points of difference) but with less than a third of its set size.

When an oversampling technique is considered for artificially balancing the set, the F_1 results show a slight improvement at the expense of an expected increase in the set size. For instance, the SMOTE case improves the result to an $F_1 = 0.70$ but with a set size of, roughly, a 150 % compared to the original. Nevertheless, if a PS stage is added afterwards, some cases retrieve very competitive F_1 results but still with a large reduction rate. For instance, the FCNN and EFCNN selection schemes when considering SMOTE as a balancing technique obtain an F_1 figure similar to the ALL case in the original situation with roughly a third and a fifth of the set size, respectively. Thus, this balancing and PS scheme seems as an appropriate preprocessing stage for large-scale imbalanced sets.

Regarding the undersampling-based balancing schemes, it can be checked that this process generally results in slightly worse scores than when oversampling the set. Particularly, the use of the CNN balancing method implies a general decrease in the F_1 results when PS is applied. However, when this CNN method is used without any PS, results are remarkably good as it achieves the same F_1 as in the initial set but with roughly half of its set size. NCL and TL schemes show better performance when coupled with PS as F_1 results get to improve when compared to their corresponding PS schemes in the initial imbalanced situation.

As of combined balancing strategies, these techniques generally obtain intermediate figures between the solely use of oversampling or undersampling.

Table 4.4: Results obtained for the experimentation of Prototype Selection in imbalanced classification environments. The resulting figures show the F_1 and reduction rate (in percentage referred to the initial case without Prototype Selection) for each combination of Prototype Selection algorithm and balancing method. Bold results remark the elements belonging to the non-dominated set.

Balancing	Metric	Prototype Selection method								
		ALL	CNN	FCNN	ENN	EFCNN	DROP3	CHC	$EN_{0.1}$	$FN_{0.1}$
Original	F_1	0.69	0.64	0.67	0.64	0.63	0.58	**0.52**	0.52	0.58
	Size (%)	100.0	30.6	28.2	85.6	7.2	8.0	**0.7**	1.6	3.3
SMOTE	F_1	0.70	0.64	0.68	0.68	0.67	0.60	**0.62**	0.53	0.59
	Size (%)	150.3	40.4	34.0	132.0	13.3	14.7	**2.1**	2.3	4.9
B1	F_1	**0.70**	0.64	0.68	0.68	**0.67**	0.60	0.59	0.48	0.55
	Size (%)	**150.3**	37.9	32.8	132.2	**12.5**	13.9	2.5	2.8	4.9
B2	F_1	0.69	0.64	0.67	0.67	0.66	0.60	0.58	0.49	0.54
	Size (%)	150.3	40.7	35.9	129.5	12.7	14.9	2.2	3.1	5.0
CNN	F_1	**0.69**	0.62	0.66	0.60	0.60	0.56	**0.47**	0.52	0.54
	Size (%)	**49.2**	27.1	26.4	34.2	4.8	5.3	**0.4**	2.5	2.7
NCL	F_1	0.69	0.64	0.67	0.66	0.65	0.56	**0.59**	0.53	0.59
	Size (%)	78.2	18.6	16.7	71.6	5.9	6.2	**0.8**	1.0	2.5
TL	F_1	0.69	0.63	0.67	0.66	**0.66**	0.58	**0.54**	0.52	0.59
	Size (%)	92.3	25.6	23.2	80.8	**7.1**	7.3	**0.7**	1.4	3.0
CNN-SMOTE	F_1	0.69	0.63	0.66	0.66	0.64	0.60	0.56	0.49	0.52
	Size (%)	65.8	32.4	30.0	49.1	8.4	9.4	1.0	2.9	3.2
CNN-B1	F_1	**0.69**	0.63	0.67	0.66	0.64	0.57	0.54	0.47	0.53
	Size (%)	**65.9**	31.3	29.3	49.1	8.5	9.4	1.1	3.0	3.2
CNN-B2	F_1	0.69	0.62	0.66	0.65	0.63	0.56	0.55	0.47	0.50
	Size (%)	65.9	32.3	31.1	47.9	8.9	9.3	1.0	3.1	3.3
NCL-SMOTE	F_1	0.69	0.65	0.67	0.67	0.67	0.61	0.59	0.52	0.58
	Size (%)	109.7	22.6	18.3	101.6	9.6	10.5	1.7	1.2	3.5
NCL-B1	F_1	0.69	0.65	**0.68**	0.68	**0.67**	0.59	0.58	0.49	0.54
	Size (%)	109.5	21.9	**18.4**	101.7	**9.4**	10.3	2.0	1.7	3.4
NCL-B2	F_1	0.69	0.64	0.67	0.67	0.66	0.60	0.58	0.49	0.53
	Size (%)	109.7	23.5	20.1	100.3	9.8	11.7	1.9	1.8	3.5
TL-SMOTE	F_1	0.69	0.65	0.67	0.68	0.67	0.59	**0.61**	0.52	0.59
	Size (%)	134.9	32.9	27.8	120.6	11.5	11.3	**1.7**	1.9	4.3
TL-B1	F_1	0.69	0.64	**0.68**	0.67	0.66	0.59	0.59	0.48	0.54
	Size (%)	134.9	31.2	**26.7**	120.3	10.9	11.4	2.3	2.3	4.3
TL-B2	F_1	0.69	0.64	0.67	0.67	0.65	0.59	0.58	0.48	0.53
	Size (%)	134.9	33.6	29.5	117.9	11.2	12.6	2.1	2.6	4.4

For instance, focus on the ENN scheme for PS: considering the initial balancing process CNN-B1, an $F_1 = 0.49$ with a 49.1 % of set size is achieved; when simply considering the oversampling scheme B1, an $F_1 = 0.68$ score with a set size of 132.2 % is obtained and if only the CNN undersampling is performed, an $F_1 = 0.60$ is scored with only 34.2 % of the initial prototypes. Thus, these combined balancing solutions together with PS may suit cases with medium reduction requirements, being undersampling techniques the ones indicated for drastic size reductions.

Figure 4.5 shows graphically the results obtained and allows their analysis in terms of the non-dominance criterion. A first point to highlight is that most of the non-dominance set of solutions comprises cases in which some type of balancing stage is considered before applying PS. While all these solutions entail a (sometimes slight) decrease in the F_1 score when compared to the case without any type of PS process, the resulting set is remarkably more compact than the original situation. For instance, the NCL-B1 balancing method coupled with FCNN achieves an $F_1 = 0.68$ with less than a fifth of the total number of prototypes.

As of the case of considering PS without any balancing stage, it may be observed that the case of using the CHC technique constitutes the only case among the non-dominated solutions. Thus, it may be mentioned that, according to the non-dominance criterion, solutions involving PS without a balancing stage may not generally be considered as optimal.

The cases that only consider the balancing scheme and avoid the PS stage are also present among the non-dominant solutions. Particularly, the non-dominated solutions by the CNN and CNN-B1 balancing cases achieve the same F_1 scores than the initial imbalanced case but with a remarkable set reduction.

Finally, it must pointed out the case of the B1 oversampling algorithm. This configuration stands as an interesting option from an accuracy point of view as it achieves the best F_1 score overall, but it entails a remarkable set size increase compared to the the initial one. Thus, while in the particular context of these experiments should be discarded as the premise is to reduce the initial set size, this result evinces a possible path to explore for cases with less retrictive requirements regarding the size of the sets or the cost of the dissimilarity metric.

4.2.5 Discussion

This study of PS in the context of class-imbalance situations points out several insights that are interesting to remark. A first point is that, as expected, the use of PS schemes in imbalanced sets generally lowers the performance of the system as these algorithms are not generally prepared to handle such situations. Thus, a possible alternative is the use of *data-level* balancing methods that aim at artificially equilibrating the classes at issue

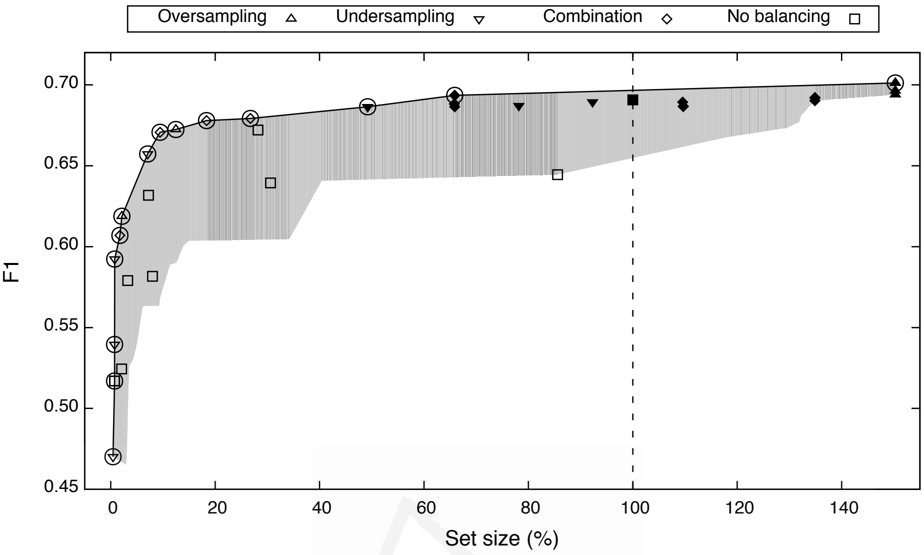


Figure 4.5: Graphical representation of the results obtained for Prototype Selection in imbalanced classification environments. Balancing paradigms are represented by the symbols in the legend. The use or not of Prototype Selection is shown by being these symbols either empty or filled, respectively. Circled symbols remark the elements belonging to the non-dominance set whereas the vertical dashed line refers to the original set size. Symbol (■) in the dashed line depicts exhaustive search without any balancing or selection technique. To avoid graph overload, the grey region depicts the space occupied by all results obtained in this work from the combinations of balancing techniques (oversampling, undersampling, and combination) and PS strategies studied.

for making the data suitable for conventional PS schemes.

The use of oversampling-based methodologies as preprocessing stage stands as a proper option for tackling the imbalance problem. Even though these methods retrieve sets with larger sizes, the fact that the classes are properly balanced allows the correct performance of the PS techniques as it may be checked in the results.

As of undersampling techniques, the first point to remark is that these methods reduce the set size without the need of an additional PS stage. However, when sharper reduction rates are required, the use of a PS stage is usually considered. Nevertheless, in such cases the accuracy figures generally show a drop in performance that may be due to the fact that undersampling techniques do not guarantee a balanced resulting set, thus harming the performance of the PS algorithm.

When considering combinations of oversampling and undersampling techniques for preprocessing the set before PS, the results obtained generally describe an intermediate tendency between the exclusive use of undersampling and oversampling. In general, accuracy figures tend to be higher than the ones achieved with undersampling schemes but also set sizes are more compact than the ones achieves with oversampling methods.

Finally, as a general conclusion we may point out that, while data-level balancing techniques improve the results when applied before a PS stage, the particular type of strategy (oversampling, undersampling, and combined methodologies) is totally dependent on the memory requirements of the eventual application as they achieve different reduction figures and, thus, different accuracy figures.

4.3 General discussion

Prototype Selection (PS) algorithms for the kNN classifier are undoubtedly useful in the context of large and noisy datasets. These algorithms aim at producing a compact and robust subset out of the initial data that maintains, or even improves, the accuracy of an initial set by selecting the most important instances according to a certain heuristic. This chapter studied the use of PS methods for the kNN classifier due to its interest for the rest of the dissertation. A direct application of such study is found in onset detection: when addressed as a classification problem, this task supposes a challenge for the kNN rule as it entails operating with large-scale and imbalanced binary data collections. In this regard, we performed two studies to review the performance of PS schemes in general large-scale sets and to assess their behaviour in the particular context of class-imbalance data collections.

The first of those studies focused on a comparative experimentation of conventional PS algorithms (condensing-based, editing-based, and hybrid approaches) with a rather new methodology known as *rank methods*. Our comprehensive experimentation proved these latter methods to be remarkably competitive compared to conventional PS techniques, especially when considering noisy environments in which they might be considered as a possible alternative to the classic Edited Nearest Neighbor (ENN) by Wilson (1972). Additionally, these rank methods allow the explicit specification of the resulting set size, thus highlighting its adaptability to the requirements of the task at issue by boosting either reduction or accuracy.

The second work studied the issue of PS techniques in datasets that require of a set size reduction (large-scale sets) but that show a class-imbalance issue. Given that general PS techniques are not prepared for such cases, a *data-level* preprocessing stage may be considered for artificially balancing the classes and then apply PS. The study confirmed the initial hypothesis that PS techniques are affected by the imbalanced in the sets and that artificially balancing the class distribution by oversampling, undersampling or combinations of them provides a proper solution to the issue. Specifically, each of the strategies proved to be useful in terms of the resulting set size requirement, being undersampling techniques interesting for cases in which a minimal set size is required, combined balancing methods for the cases in which set size is restrictive but not in such a prohibitive sense and, finally, oversampling methods for the least restrictive set size cases.

Chapter 5

Approaches for Interactive Onset Detection and Correction

"If my calculations are correct, when this baby hits 88 miles per hour... you're gonna see some serious shit"

Dr. Emmett Brown

As pointed out by different authors in the Music Information Retrieval community, current state-of-the-art approaches for Automatic Music Transcription seem to have reached a glass ceiling. New techniques, schemes, and proposals do not report significant advances but simply subtle improvements in the standard bechmark data collections (e.g., the MIREX one).

The proposal of alternative paradigms for Automatic Music Transcription seems appropriate to further develop this field. Given such need, we consider the study of interactive schemes in which the user is not only used as a validation agent but also an active part in the success of the task. While considering the user as an active part of the system may be seen as a drawback, given the limitations found in current transcription systems, the user is indeed always required for post-processing the output of the system. Taking this need into consideration, the study and proposal of schemes for the efficient exploitation of the user effort is a necessary point to address.

In this work we address the task of onset detection from this interactive point of view. Onset information represents the starting points of note events in audio streams and, in spite of the large amount of research carried out in this field, no existing method is error-free. A conceptual description is shown in Fig. 5.1: an onset detection algorithm performs an initial estimation on the signal; the user validates the output and provides feedback to the detection model, which is iteratively improved. Note that in this context the quality in the result is guaranteed by the expertise of the user, and thus the point to assess is whether such interactive provides a workload reduction compared to a non-interactive correction paradigm.

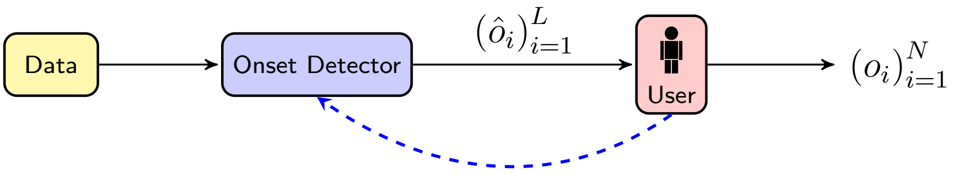


Figure 5.1: Generic scheme for the Interactive Onset Detection and Correction paradigm. Stand-alone onset estimation methods do not consider the feedback provided by the dashed line.

In terms of the precise work presented in this chapter, four different studies related to the aforementioned ideas of interactive schemes and onset detection are presented. The first work studies several aspects of the onset selection stage in stand-alone onset detection systems to gather conclusions that shall be later used in interactive systems; due to the lack of evaluation criteria for interactive onset detection and correction schemes, the second work proposes a set of measures for quantitatively measuring the effort invested by users. The third and fourth studies propose a set of interactive onset correction schemes addressed from signal processing and machine learning perspectives, respectively. Finally, a last section is devoted for the discussion of the main ideas presented in this chapter.

5.1 Analysis of descriptive statistics and adaptive methodologies for Onset Selection Functions

As explained in Chapter 2, the most extended methodology for estimating onsets in audio streams is based on a two-stage approach: the initial Onset Detection Function (ODF) step that processes the target signal computing a time series whose peaks represent the positions of the estimated onsets; and the Onset Selection Function (OSF) stage that filters out the results of the previous stage retrieving only the most promising peaks as onsets.

Given that no ODF process is totally neat, the OSF constitutes a key point in the performance of the system. Nevertheless, except for some particular works explicitly assessing the influence of the OSF process in the overall performance of the system (e.g., the work by Rosão et al. (2012)), there is still a need to further study such processes. Thus, in this work we propose a survey of OSF strategies that both consider the use of different statistical descriptors and sliding window analyses for performing the selection task.

Note that such survey is also relevant for the main aim of the chapter, the case of Interactive Onset Detection systems. As it shall be explained, interactive systems improve their performance by adapting themselves to the particularities of the task through the interactions with the user. In the case of Onset Detection systems, one possibility is to progressively adapt the parameters of either the ODF stage, the OSF one, or even both at the same time. In this regard, studying the influence of the OSF stage in the overall performance of the system shall provide relevant insights about the achievable results when considering an OSF-based interactive correction system, which is the case that we shall later address.

For this study we initially introduce the different OSF schemes considered to be compared; after that, the evaluation methodology comprising the data considered, the assessment figures, and the set of ODF functions for the comparison are presented; then the results obtained are introduced and analyzed to gather the most relevant conclusions; finally, the study ends up with a brief discussion.

5.1.1 Onset Selection Functions

In this work we perform an experimental study comparing different ideas for OSF that, to our best knowledge, has not been previously performed. Particularly, we aim at studying the following premises: *i)* considering other percentile values (i.e., other statistical tendency measures) different to the median, somehow extending the work by Kauppinen (2002) but particularized to onset detection; *ii)* assessing the relation between the window size in adaptive detection methodologies and the overall performance of the system; and *iii)* comparing the difference in performance between static and adaptive methodologies. Mathematically, these criteria can be formalized as:

$$\theta(t_i) = \mu\{O(t_{w_i})\} + \mathcal{P}^{(n)}\{O(t_{w_i})\} \tag{5.1}$$

where $t_{w_i} \in \left[t_i - \frac{W}{2}, t_i + \frac{W}{2}\right]$, W denotes the size of the sliding window, $\mu\{\cdot\}$ and $\mathcal{P}^{(n)}\{\cdot\}$ denote the average and n_{th} percentile value of the sample distribution at issue, respectively, and $O(t)$ stands for the time series resulting from an ODF process.

For assessing the influence of the percentile, 20 values equally spaced in the range $n \in [0, 100]$ were considered. Note that the particular case of $\mathcal{P}^{(50)}$ is equivalent to the median value of the distribution. As of window sizes, 20 values equally space in the range $W \in [0.2, 5]$ s were also considered. The value by West and Cox (2005) of $W = 1.5$ s was included as a reference.

To simulate a case of static OSF strategy, W was set to the length of the $O(t)$. Also, the case of manually imposing the threshold value $\theta(t_i) = \mathcal{T}$ was considered. For that we establish 20 values equally spaced in the range

$\mathcal{T} \in [0, 1]$ [1].

For the rest of the section, consider the following notation for referring to the different OSF configurations: μ and $\mathcal{P}$ denote the exclusive use of the mean or the percentile, respectively, whereas $\mu + \mathcal{P}$ stands for the sum of both descriptors; adaptive approaches are distinguished from the static ones by incorporating the t as a subindex, showing their temporal dependency (i.e., μ_t, $\mathcal{P}_t$, and $\mu_t + \mathcal{P}_t$ in opposition to μ, $\mathcal{P}$, and $\mu + \mathcal{P}$); $\mathcal{T}$ evinces the manually imposition of the threshold value; lastly, $\mathcal{B}$ is used to denote the case in which no threshold is applied and all local maxima are retrieved as onsets: $\mathcal{B} \equiv \theta(t_i) = 0$.

5.1.2 Evaluation methodology

The dataset used for the evaluation is the one introduced in Böck et al. (2012). It comprises a set of 321 monaural real world recordings sampled at 44.1 kHz covering a wide range of timbres and polyphony degrees. The total duration of the set is 1 hour and 42 minutes containing 27,774 onsets with an average duration of 19 seconds per file (the shortest lasts 1 second and the largest one extends up to 3 minutes) and an average figure of 87 onsets per file (minimum of 3 onsets and maximum of 1,132 onsets). A detailed description of the collection in terms of instrumentation and number of onsets is shown in Table 5.1. Note that this partitioning is only for informative purposes; in our experiments, as the idea is to assess the influence of the OSF stage in a generic fashion, we obviate the nature of the data and thus consider a single partition.

Table 5.1: Description of the onset detection dataset by Böck et al. (2012) used for evaluation in terms of instrumentation and number of onsets.

Instrumentation	**Files**	**Onsets**
Complex mixtures	193	21,091
Pitched percussive	60	2,981
Wind instruments	25	822
Bowed strings	23	1,180
Non-pitched percussive	17	1,390
Vocal	3	310
Total	321	27,774

Regarding performance assessment we consider the standard evaluation strategy for onset detection introduced in Section 2.3.1 with a 50 ms tolerance window. Although the results obtained are assessed in terms of the figures

[1] To ensure this range is equally relevant to all resulting $O(t)$ functions, these time series are normalized to range $[0, 1]$ as explained in the Section 5.1.2.

of Precision (P), Recall (R), and F-measure (F_1), in order to be concise we shall only present the last of them as it properly summarizes the initial two metrics.

We selected two different ODF strategies according to their good results reported in the literature for the experiments: the Semitone Filter Bank (SFB) method by Pertusa et al. (2005) and the SuperFlux (SuF) algorithm by Böck and Widmer (2013a, 2013b).

As an energy-based ODF method, SFB analyses the evolution of the magnitude spectrogram with the particular assumption of considering that harmonic sounds are being processed. A semitone filter bank is applied to each frame window of the magnitude spectrogram, being the different filters centered at each of the semitones marked by the well temperament, and the energy of each band (root mean square) is retrieved. After that, the first derivative across time is obtained for each single band. As only energy rises may point out onset information, negative outputs are zeroed. Eventually, all bands are summed and normalized to obtain the $O(t)$ function.

The SuF method bases its performance on the idea of the Spectral Flux (Masri, 1996) signal descriptor and extends it. Spectral Flux obtains positive deviations of the bin-wise difference between the magnitude of two consecutive frames of the magnitude spectrogram for retrieving the $O(t)$ function. SuF substitutes the difference between consecutive analysis windows by a process of tracking spectral trajectories in the spectrum together with a maximum filtering process for suppressing vibrato articulations that tend to increase false detections.

Given that these ODF processes may not span in the same range, we apply a normalization process to time series $O(t)$ to ensure that the resulting function only spans within the range $[0, 1]$.

Finally, the analysis parameters of both algorithms have been configured to a window size of 92.9 ms with a 50 % of overlapping factor. With such configuration we match the resolution of the system with the tolerance window considered for the assessment methodology.

5.1.3 Results

We now introduce and analyze the results obtained for the different experiments proposed. All figures presented depict the weighted average of the results obtained for each of the audio files considering the number of onsets each one contains.

Table 5.2 shows the results obtained for the proposed comparative of static and adaptive threshold methodologies for the two ODF processes considered. In this particular case, the window size for the adaptive methodologies has been fixed to the reference value of $W = 1.5$ s as in West and Cox (2005).

An initial remark to point out according to the results obtained is that, in general, the figures obtained with the sliding window methodology do not

Table 5.2: Results in terms of the F_1 score for the descriptive statistics and adaptive methodologies study when considering a 1.5-second sliding window. Bold elements represent the best figures for each Onset Detection Function considered. Due to space requirements we have selected the threshold and percentile parameters (Th/Pc) showing the general tendency. Also, threshold and percentile parameters are expressed in the range $[0, 1]$.

Th/Pc	**SFB**							**SuF**						
	$\mathcal{T}$	μ	$\mathcal{P}$	$\mu + \mathcal{P}$	μ_t	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{T}$	μ	$\mathcal{P}$	$\mu + \mathcal{P}$	μ_t	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$
0.00	0.65	**0.73**	0.65	**0.73**	0.72	0.65	0.72	0.64	**0.77**	0.64	**0.77**	0.74	0.64	0.76
0.11	0.71	**0.73**	0.65	0.72	0.72	0.65	0.71	0.76	**0.77**	0.64	**0.77**	0.74	0.64	0.76
0.21	**0.75**	0.73	0.66	0.71	0.72	0.66	0.70	0.65	**0.77**	0.65	**0.77**	0.74	0.64	0.76
0.32	**0.73**	**0.73**	0.66	0.70	0.72	0.66	0.69	0.51	**0.77**	0.66	**0.77**	0.74	0.65	0.76
0.42	0.66	**0.73**	0.68	0.68	0.72	0.67	0.67	0.39	**0.77**	0.68	0.76	0.74	0.67	0.75
0.53	0.56	**0.73**	0.69	0.66	0.72	0.69	0.63	0.29	**0.77**	0.70	0.75	0.74	0.69	0.74
0.63	0.44	**0.73**	0.70	0.62	0.72	0.70	0.58	0.21	**0.77**	0.73	0.73	0.74	0.71	0.71
0.74	0.31	**0.73**	0.70	0.53	0.72	0.70	0.50	0.15	**0.77**	0.74	0.68	0.74	0.72	0.65
0.84	0.19	**0.73**	0.65	0.39	0.72	0.63	0.36	0.10	**0.77**	0.68	0.56	0.74	0.65	0.54
0.95	0.10	**0.73**	0.40	0.13	0.72	0.42	0.12	0.06	**0.77**	0.42	0.29	0.74	0.44	0.27
1.00	0.05	**0.73**	0.05	0.00	0.72	0.27	0.00	0.05	**0.77**	0.05	0.00	0.74	0.29	0.00

remarkably differ to the ones obtained with the static one, independently of the ODF process. This can be clearly seen when $\mathcal{P}$ and $\mu + \mathcal{P}$ are respectively compared to $\mathcal{P}_t$ and $\mu_t + \mathcal{P}_t$: all percentile values considered retrieve considerably similar results except for the case when high percentile values are considered, in which the $\mathcal{P}_t$ method shows its superior capabilities. Additionally, while $\mathcal{P}$ approaches show their best performance for percentile values in the range $[60, 75]$, the $\mu + \mathcal{P}$ methods obtain their optimal results in the lower ranges, more precisely around $[0, 30]$.

Regarding the use of the μ methodologies, results obtained for the static and adaptive methodologies did not differ for the SFB function. On the contrary, when considering the SuF function, the adaptive methodology performed slightly worse than the static one. Finally, as these methods do not depend on any external configuration, their performance does not vary with the percentile parameter.

In terms of the $\mathcal{T}$ strategy, its performance matched the static $\mathcal{P}$ and $\mu + \mathcal{P}$ methods in the sense that the performance degrades as the introduced threshold was increased. Nevertheless, this method shows its best performance when the threshold value considered lies in the range $[0.10, 0.30]$.

As a general comparison of the methods considered, the μ methods showed a very steady performance paired with high performance results. Also, while $\mu + \mathcal{P}$ methods generally outperform $\mathcal{P}$ strategies in terms of peak performance, the particular case of $\mathcal{P}_t$ approaches showed less dependency on the percentile configuration value.

Given that all previous experimentation has been done considering a fixed window value of $W = 1.5$ s, we need to study the influence of that parameter in the overall performance of the system. In this regard, Tables 5.3 and 5.4 show the results for the $\mathcal{P}_t$ and $\mu_t + \mathcal{P}_t$ methods considering different window sizes for the SFB and SuF processes, respectively. Additionally, Figures 5.2 and 5.3 graphically show these results for the $\mathcal{P}_t$ and $\mu_t + \mathcal{P}_t$ methods, respectively.

As the figures obtained for both ODF processes qualitatively show the same trends, we shall analyze them jointly. Checking Tables 5.3 and 5.4, the results for the $\mathcal{P}_t$ method show that larger windows tend to increase the overall performance of the detection, at least when not considering an extreme percentile value. For instance, let us focus on the $\mathcal{P}_t^{(74)}$ method for the SuF process (Table 5.4): when considering a window of $W = 0.20$ s, the performance is set on $\mathrm{F}_1 = 0.67$, which progressively improves as the window size is increased, getting to a score of $\mathrm{F}_1 = 0.73$ when $W = 5.00$ s.

As commented, this premise is not accomplished when percentile values are set to its possible extremes. For low percentile values, window size seems to be irrelevant to the system. For instance, when considering the SFB method, $\mathcal{P}_t^{(11)}$, the performance measure was always $\mathrm{F}_1 = 0.65$ independently of the W value considered.

Table 5.3: Results in terms of the F_1 score for the descriptive statistics and adaptive methodologies study for the Semitone Filter-Bank Onset Detection Function. Bold elements represent the best figures obtained for each percentile value considered. Due to space requirements, the most representative percentile parameters (Pc) and window sizes (W) have been selected to show the general tendency.

Pc	$W = 0.20$		$W = 0.71$		$W = 1.50$		$W = 2.22$		$W = 2.98$		$W = 3.99$		$W = 5.00$	
	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$
0	0.65	0.66	0.65	0.71	0.65	0.72	0.65	0.72	0.65	0.72	0.65	0.72	0.65	**0.73**
11	0.65	0.65	0.65	0.70	0.65	0.71	0.65	0.71	0.65	**0.72**	0.65	**0.72**	0.65	**0.72**
21	0.65	0.63	0.65	0.69	0.66	0.70	0.66	0.70	0.66	**0.71**	0.66	**0.71**	0.66	**0.71**
32	0.65	0.60	0.66	0.68	0.66	0.69	0.66	0.69	0.66	0.69	0.66	**0.70**	0.66	**0.70**
42	0.65	0.56	0.67	0.65	0.67	0.67	0.67	0.67	0.67	**0.68**	0.67	**0.68**	**0.68**	**0.68**
53	0.65	0.49	0.68	0.62	**0.69**	0.64	**0.69**	0.64	**0.69**	0.65	**0.69**	0.65	**0.69**	0.65
63	0.65	0.42	0.69	0.56	**0.70**	0.59	**0.70**	0.59	**0.70**	0.60	**0.70**	0.60	**0.70**	0.60
74	0.67	0.32	0.69	0.48	0.69	0.51	**0.70**	0.51	**0.70**	0.51	**0.70**	0.52	**0.70**	0.52
84	**0.67**	0.12	0.66	0.35	0.64	0.36	0.64	0.36	0.64	0.37	0.64	0.37	0.64	0.37
95	**0.67**	0.00	0.47	0.15	0.45	0.12	0.42	0.12	0.42	0.12	0.41	0.12	0.40	0.12
100	**0.67**	0.00	0.47	0.00	0.23	0.00	0.21	0.00	0.17	0.00	0.13	0.00	0.11	0.00

Table 5.4: Results in terms of the F_1 score for the descriptive statistics and adaptive methodologies study for the SuperFlux Onset Detection Function. Bold elements represent the best figures obtained for each percentile value considered. Due to space requirements, the most representative percentile parameters (Pc) and window sizes (W) have been selected to show the general tendency.

Pc	$W = 0.20$		$W = 0.71$		$W = 1.50$		$W = 2.22$		$W = 2.98$		$W = 3.99$		$W = 5.00$	
	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$
0	0.64	0.72	0.64	0.74	0.64	0.76	0.64	0.76	0.64	**0.77**	0.64	**0.77**	0.64	**0.77**
11	0.64	0.72	0.64	0.75	0.64	0.76	0.64	**0.77**	0.64	**0.77**	0.64	**0.77**	0.64	**0.77**
21	0.64	0.72	0.64	0.75	0.65	0.76	0.65	0.76	0.65	**0.77**	0.65	**0.77**	0.65	**0.77**
32	0.64	0.71	0.65	0.75	0.65	**0.76**	0.65	**0.76**	0.65	**0.76**	0.66	**0.76**	0.66	**0.76**
42	0.64	0.70	0.66	0.74	0.67	0.75	0.67	0.75	0.67	**0.76**	0.67	**0.76**	0.68	**0.76**
53	0.64	0.65	0.68	0.73	0.69	0.74	0.69	0.74	0.70	0.74	0.70	0.74	0.70	**0.75**
63	0.64	0.61	0.70	0.69	0.71	0.71	0.71	0.71	**0.72**	**0.72**	**0.72**	**0.72**	**0.72**	**0.72**
74	0.67	0.55	0.70	0.63	0.72	0.66	0.72	0.66	**0.73**	0.66	**0.73**	0.67	**0.73**	0.67
84	0.67	0.38	**0.68**	0.52	0.67	0.54	0.67	0.54	0.67	0.55	0.67	0.55	0.67	0.55
95	**0.67**	0.00	0.48	0.32	0.47	0.27	0.44	0.27	0.44	0.27	0.43	0.28	0.42	0.28
100	**0.67**	0.00	0.48	0.00	0.24	0.00	0.21	0.00	0.18	0.00	0.15	0.00	0.12	0.00

On the other extreme, very high percentile values suffer a performance decrease as larger window sizes are used. As an example, for the SuF configuration, $\mathcal{P}_t^{(100)}$ with $W = 5.00\ s$ achieved an $\mathrm{F}_1 = 0.12$, while when $W = 0.20\ s$ this figure raised up to $\mathrm{F}_1 = 0.67$.

Results for the $\mu_t + \mathcal{P}_t$ method showed similar tendencies since, when not considering extreme percentile values, the overall performance increased as larger windows were used. Nevertheless, in opposition to the $\mathcal{P}_t$ case, this particular configuration showed an improvement tendency as W is progressively increased for low percentile values.

In general it was observed that, for all W window sizes, the best percentile configurations seemed to be in the range $[60, 70]$ for the $\mathcal{P}_t$ approach and in the range $[0, 20]$ for the $\mu_t + \mathcal{P}_t$ case. This fact somehow confirms the hypothesis suggesting that the median value may not always be the best percentile to consider.

Checking now the results considering Figs. 5.2 and 5.3, some additional remarks more difficult to be checked in the aforementioned tables may be pointed out.

The first one is that the selection of the proper parameters of window size and percentile factor is crucial. For both $\mathcal{P}_t$ and $\mu_t + \mathcal{P}_t$ methods, there is a *turning point* in which the performance degrades to values lower than the considered baseline $\mathcal{B}$ (no OSF applied). For the $\mathcal{P}_t$ method there is a clear point for this change in tendency around the 85th percentile for any window size considered. However, for the $\mu_t + \mathcal{P}_t$ approach there was not a unique point but a range, which remarkably varied depending on the type of ODF and window size considered.

Another point to highlight is that the static methodologies ($\mathcal{P}$ and $\mu + \mathcal{P}$) consistently define the upper bound in performance before the so-called turning point. This fact somehow confirms the initial idea of using large windows (in the limit, one single window considering the whole $O(t)$ function) in opposition to small windows.

The results obtained when considering window sizes in the range $[0, 1]$ seconds, which include the performance of the considered reference window of $W = 1.5\ s$, achieved results similar to the obtained upper bound. The other considered window sizes showed a remarkable variability in the performance, ranging from achieving figures similar to the upper bound to figures close to the baseline $\mathcal{B}$.

As a general summary for all the experimentation performed, we may conclude that adaptive OSF methodologies may be avoided as static approaches obtain similar results with less computational cost. Particularly, in our experiments, methods considering percentile ($\mathcal{P}$) or mean and percentile ($\mu + \mathcal{P}$) reported the upper bounds in performance.

Nevertheless, the particular percentile value used remarkably affects the performance. For $\mathcal{P}$, the best results seem to be obtained when the percentile

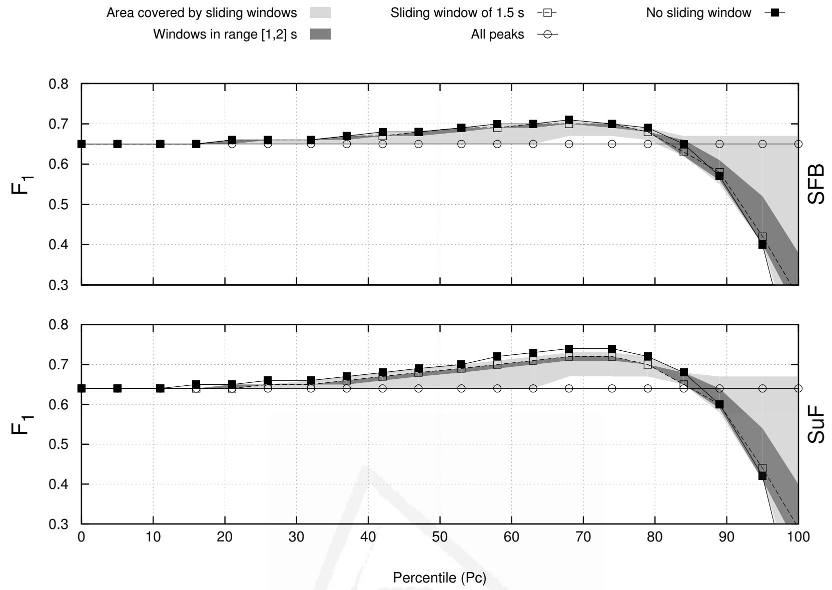


Figure 5.2: Evolution of the F_1 score when considering the $\mathcal{P}_t$ strategy for the Onset Selection Function process.

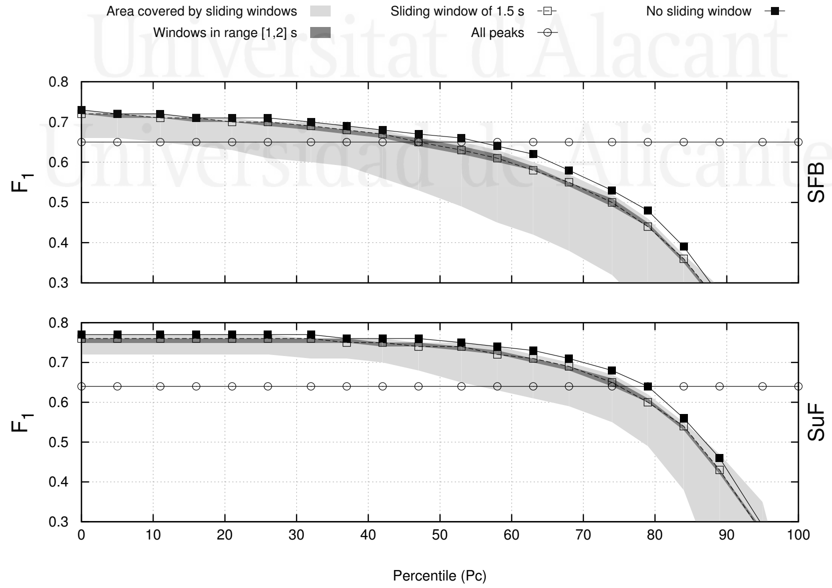


Figure 5.3: Evolution of the F_1 score when considering the $\mu_t + \mathcal{P}_t$ strategy for the Onset Selection Function process.

parameter was set in the range [60,70]. For $\mu + \mathcal{P}$ the best figures were obtained when this parameter was set to a very low value.

Finally, the commonly considered median value for the OSF did not report the best results in our experiments. These results point out that the median statistical descriptor may not always be the most appropriate to be used, being necessary to be tuned for each particular dataset.

Statistical significance analysis

In order to perform a rigorous analysis of the results obtained and derive strong conclusions out of them, we considered a set of statistical tests. Specifically, these analyses were performed with the non-parametric Wilcoxon rank-sum and Friedman tests (Demšar, 2006), which avoid any assumption about the distribution followed by the figures obtained. The former method is helpful for comparisons of different distributions in a pairwise fashion while the latter one generalizes this pairwise comparison to a generic number of distributions.

The Wilcoxon rank-sum test was applied to assess whether there are significant differences among all the methods proposed using the results in Table 5.2. The single scores obtained for each Th/Pc constitute a sample of the distribution for the OSF method to be tested. The results obtained when considering a significance level of $p < 0.05$ can be checked in Table 5.5.

Attending to the results of the Wilcoxon test, an initial point to comment is that method $\mathcal{T}$ was the less robust due to its significantly lower performance in all cases. Oppositely, method μ can be considered the best strategy as it outperformed all other strategies. The adaptive equivalent to this technique, μ_t, achieved similar performances to the static one except for the case of SuF, in which it did not outperform the rest of the methods as consistently as in the static case. Methods $\mathcal{P}$ and $\mathcal{P}_t$ showed an intermediate performance among the previously commented extremes. Their results were not competitive with any of the μ or μ_t strategies. These strategies are generally tied with methods $\mu + \mathcal{P}$ and $\mu_t + \mathcal{P}_t$ as they typically showed significantly similar performances with punctual cases in which one of those methods improved the rest.

We now consider the statistical analysis of the influence of the window size (W) and the percentile values (Pc) on the overall performance. Given that we have two variables to be evaluated, we considered the use of the Friedman test. The idea with this experiment was to assess whether variations in these variables reported statistically significant differences in the F_1 score, so we assume that the null hypothesis to be rejected was that no difference should be appreciated. Note that this type of analysis does not report whether a particular strategy performs better than another one but simply if the differences when using different parameters are statistically significant.

For performing this experiment we considered the data in Tables 5.3 and

Table 5.5: Results for the Wilcoxon rank-sum test of the F$_1$ score for the descriptive statistics and adaptive methodologies study when considering a 1.5-second sliding window size.. Symbols ✓, ✗, and = state that the onset detection capability of the method in the row is significantly higher, lower, or not different to the method in the column. Symbol – depicts that the comparison is obviated. A significance level of $p < 0.05$ has been considered.

Method	SFB							SuF						
	$\mathcal{T}$	μ	$\mathcal{P}$	$\mu + \mathcal{P}$	μ_t	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{T}$	μ	$\mathcal{P}$	$\mu + \mathcal{P}$	μ_t	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$
$\mathcal{T}$	–	✗	✗	✗	✗	✗	✗	–	✗	✗	✗	✗	✗	✗
μ	✓	–	✓	✓	✓	✓	✓	✓	–	✓	✓	✓	✓	✓
$\mathcal{P}$	✓	✗	–	=	✗	=	=	✓	✗	–	=	✗	✓	=
$\mu + \mathcal{P}$	✓	✗	=	–	✗	=	✓	✓	✗	=	–	=	=	✓
μ_t	✓	✗	✓	✓	–	✓	✓	✓	✗	✓	=	–	✓	=
$\mathcal{P}_t$	✓	✗	=	=	✗	–	✓	✓	✗	✗	=	✗	–	=
$\mu_t + \mathcal{P}_t$	✓	✗	=	✗	✗	✗	–	✓	✗	=	✗	=	=	–

5.4 as they contained all the information required for the analysis. In this regard Tables 5.6 and 5.7 showed the p significance values obtained when measuring the influence of W and Pc, respectively.

Attending to the results obtained, we can check the remarkable influence of the W parameter in the overall results as all cases reported very low p scores that reject the null hypothesis. The only exception is the $\mathcal{P}_t$ in the SFB case in which p was not that remarkably low, but still would reject the null hypothesis considering a typical significance threshold of $p < 0.05$.

As of the percentile value (Pc), the p significance values obtained clearly show the importance of this parameter in the design of the experiment. This points out the clear need to considered this as another parameter in the design of onset detection systems.

Finally, it is important to highlight that this statistical analysis confirmed the initial conclusions depicted previously: static approaches are significantly competitive when compared to adaptive methods; window sizes for adaptive methods remarkably influence the performance of the system; when using OSF methods considering statistical descriptions, the percentile should be considered as another design parameter due to its relevance in the results.

5.1.4 Discussion

This comparative study of the influence of the OSF stage in onset detection schemes points out several conclusions that are relevant to highlight. As a first conclusion to remark is that, as suspected, the OSF stage plays a key role in the overall success of the task. Actually, depending on the parameterization of the stage, the results range from completely mistaken estimations to fairly accurate results. This fact supports the premise that shall be later introduced of considering Interactive Onset Detection schemes based on *interacting* with the parameters of the OSF stage to reduce the annotation/correction workload of a user.

Additionally, it has been shown that, somehow in opposition to general intuition, non-adaptive OSF schemes get to improve the results of adaptive methodologies when properly configured. This is a quite relevant point as the latter ones usually imply a much superior computational cost compared to the former ones. Nevertheless it is also important to highlight that, in the context of adaptive schemes, the proper configuration of the window size constitutes a key point in the success of the task.

As a last conclusion to comment, it must be mentioned that the use of percentiles different to the typical median descriptor has also proved its influence in the task. Thus, this element should also be considered another design parameter to be adjusted for the proper consecution of the onset detection task.

Table 5.6: Statistical significance results of the Friedman test when measuring the influence of the window size (W) in the overall performance.

Method	SFB	SuF
$\mathcal{P}_t$	0.03209	$4.917 \cdot 10^{-9}$
$\mu_t + \mathcal{P}_t$	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$

Table 5.7: Statistical significance results of the Friedman test when measuring the influence of the percentile (Pc) in the overall performance.

Method	SFB	SuF
$\mathcal{P}_t$	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$
$\mu_t + \mathcal{P}_t$	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$

5.2 User effort assessment in Interactive Onset Detection

In this second section we introduce the issue of the assessment of interactive methodologies for the particular case of onset detection. As previously introduced, interactive methodologies are progressively being considered in a larger number of fields, among which both MIR and PR are included.

Nevertheless, to our best knowledge, interactive methodologies have not been rigourously applied to onset detection. Thus, there is a lack of not only systems and schemes capable of performing such task but also, and maybe more critical, of evaluation and assessment strategies capable of objectively compare different proposals. We shall now formalize this idea to then propose a set of measures designed for such aim.

Onset detection algorithms rarely retrieve a perfect result in terms of precision. The two types of error that affect this performance are: *i)* the algorithm misses onsets that should be detected – False Negatives (FNs) – and *ii)* the algorithm detects onsets than do not actually exist – False Positives (FPs). In these terms, let $\mathrm{N_{FP}}$ and $\mathrm{N_{FN}}$ be the amount of FP and FN errors committed after processing a given signal with an Onset Detection algorithm. Let also $\mathrm{N_{GT}}$ denote the total number of onsets to be annotated in an audio file (ground truth) and $\mathrm{N_{OK}}$ represent the number of correctly detected onsets.

The amount of onsets obtained by a detection algorithm may be expressed as $\mathrm{N_D} = \mathrm{N_{OK}} + \mathrm{N_{FP}}$ whereas the total number of onsets to be estimated can be expressed as $\mathrm{N_{GT}} = \mathrm{N_{OK}} + \mathrm{N_{FN}}$. Therefore, a user starting from the initial $\mathrm{N_D}$ analysis should manually eliminate the $\mathrm{N_{FP}}$ erroneous estimations and annotate the $\mathrm{N_{FN}}$ missed onsets, thus requiring a total of $\mathrm{C_T} = \mathrm{N_{FP}} + \mathrm{N_{FN}}$ corrections to obtain the appropriate annotation.

User interaction, meaning that the system attempts to adjust its performance based on the corrections performed by the user, is proposed to reduce $\mathrm{C_T}$. The idea is that the total number of corrections performed in an interactive system $\mathrm{C_T}^{int}$ is lower than, or in the worst-case scenario, equal to, the amount required in a complete manual correction $\mathrm{C_T}^{man}$, i.e. $\mathrm{C_T}^{int} \leq \mathrm{C_T}^{man}$.

5.2.1 Measures for Interactive Onset Detection

Based on the theoretical framework presented before, we shall now introduce the two measures we propose. Note that in all cases we assume that the effort in the correction task is represented by the amount of interactions $\mathrm{C_T}$ the user needs to perform.

Total Corrections ratio

The first of the two proposed metrics is the *Total Corrections ratio*, $\mathrm{R_{TC}}$. The idea behind this measure is comparing the amount of corrections a user needs to perform when using an interactive system $\left(\mathrm{C_T}^{int}\right)$ to a manual correction $(\mathrm{C_T}^{man})$. This ratio is obtained as:

$$\mathrm{R_{TC}} = \frac{\mathrm{C_T}^{int}}{\mathrm{C_T}^{man}} = \frac{\mathrm{N_{FP}}^{int} + \mathrm{N_{FN}}^{int}}{\mathrm{N_{FP}}^{man} + \mathrm{N_{FN}}^{man}} \tag{5.2}$$

Depending on the resulting ratio value, it is possible to assert whether the interactive scheme reduces the workload:

$$\mathrm{R_{TC}} \begin{cases} > 1 & \Rightarrow \textbf{Increasing} \text{ workload} \\ = 1 & \Rightarrow \textbf{No} \text{ difference} \\ < 1 & \Rightarrow \textbf{Decreasing} \text{ workload} \end{cases}$$

Corrections to Ground Truth ratio

Although the previous metric is able to assess whether an interactive scheme requires less effort than a manual correction, a certain premise is being assumed: an automatic onset detection stage reduces the annotation workload since it tracks, at least, part of the elements that must be annotated.

However, it is possible that the automatic detection algorithm will not be able to perform this task as expected (for instance, when dealing with a noisy signal). In such cases, the number of correctly tracked onsets $\mathrm{N_{OK}}$ may be negligible, or even non-existing, thus leading to $\mathrm{N_D} = \mathrm{N_{OK}} + \mathrm{N_{FP}} \approx \mathrm{N_{FP}}$. The user would be required to annotate all the onsets $\mathrm{N_{GT}}$ plus eliminating the $\mathrm{N_{FP}}$ errors committed, i.e. $\mathrm{C_T} = \mathrm{N_{GT}} + \mathrm{N_{FP}} = \mathrm{N_{OK}} + \mathrm{N_{FN}} + \mathrm{N_{FP}}$. Under these circumstances, it would be arguable the need for an initial onset detection as the manual annotation of the signal from scratch would imply less workload.

To cope with this issue, the *Corrections to Ground Truth ratio*, $\mathrm{R_{GT}}$, compares the amount of interactions required $\mathrm{C_T}$ in relation to the total amount of ground truth onsets $\mathrm{N_{GT}}$ for both interactive systems (Eq. 5.3) and manual corrections (Eq. 5.4).

$$\mathrm{R}_{\mathrm{GT}}^{int} = \frac{\mathrm{C}_{\mathrm{T}}^{int}}{\mathrm{N}_{\mathrm{GT}}} = \frac{\mathrm{N}_{\mathrm{FP}}^{int} + \mathrm{N}_{\mathrm{FN}}^{int}}{\mathrm{N}_{\mathrm{GT}}} = \frac{\mathrm{N}_{\mathrm{FP}}^{int} + \mathrm{N}_{\mathrm{FN}}^{int}}{\mathrm{N}_{\mathrm{OK}} + \mathrm{N}_{\mathrm{FN}}} \tag{5.3}$$

$$\mathrm{R}_{\mathrm{GT}}^{man} = \frac{\mathrm{C}_{\mathrm{T}}^{man}}{\mathrm{N}_{\mathrm{GT}}} = \frac{\mathrm{N}_{\mathrm{FP}}^{man} + \mathrm{N}_{\mathrm{FN}}^{man}}{\mathrm{N}_{\mathrm{GT}}} = \frac{\mathrm{N}_{\mathrm{FP}}^{man} + \mathrm{N}_{\mathrm{FN}}^{man}}{\mathrm{N}_{\mathrm{OK}} + \mathrm{N}_{\mathrm{FN}}} \tag{5.4}$$

Bearing in mind that a ratio of 1 is equivalent to manually annotating all the onsets, the results depict whether the system forces the user to make more corrections than without any initial detection, thus making the system useless in practice:

$$\mathrm{R}_{\mathrm{GT}} \begin{cases} > 1 & \Rightarrow \textbf{More} \text{ than manual} \\ = 1 & \Rightarrow \textbf{Same} \text{ as manual} \\ < 1 & \Rightarrow \textbf{Less} \text{ than manual} \end{cases}$$

Finally, it must be pointed out the existing relation among measures $\mathrm{R}_{\mathrm{GT}}^{int}$ (Eq. 5.3) and $\mathrm{R}_{\mathrm{GT}}^{man}$ (Eq. 5.4) with measure $\mathrm{R_{TC}}$ (Eq. 5.2) by using the following expression:

$$\mathrm{R}_{\mathrm{TC}} = \frac{\mathrm{R}_{\mathrm{GT}}^{int}}{\mathrm{R}_{\mathrm{GT}}^{man}} = \frac{\mathrm{N}_{\mathrm{FP}}^{int} + \mathrm{N}_{\mathrm{FN}}^{int}}{\mathrm{N}_{\mathrm{FP}}^{man} + \mathrm{N}_{\mathrm{FN}}^{man}} \tag{5.5}$$

5.2.2 Discussion

Onset detection constitutes one of the mostly addressed tasks in the MIR field and thus its evaluation methodology has been largely discussed. Nevertheless, these classic measures are designed for stand-alone onset detection algorithms in which no intervention from a user is expected.

When interactivity is considered, there is a lack of methodology for assessing such strategies. Thus, in this work we proposed a set of figures of merit that shall be considered in the evaluation of the Interactive Onset Detection methodologies to be described in the following sections. More precisely, the two measures proposed assess: *i)* the potential workload reduction when comparing an interactive correction method compared to the case of manually correcting all errors spotted; and *ii)* the potential workload reduction when considering an interactive correction strategy compared to the case of manually annotating all onsets in an audio stream.

While we are aware that these measures simply constitute a first proposal for assessing such type of interactive schemes that may be significantly improved, it must be considered that they provide a first tool that allows to objectively and quantitatively compare the user workload invested in such tasks.

5.3 Signal processing methods for Interactive Onset Detection

In this section we introduce a set of schemes for Interactive Onset Correction. The proposals explained in this section are based on the well-known two-stage Onset Detection approach: an initial Onset Detection Function (ODF) process and a posterior Onset Selection Function (OSF) stage. The idea is that, as the user points out errors in the detection, the system gathers information to modify its performance accordingly. Note that the strategies described in this section are based on a signal processing framework, thus the different user interactions shall, in general, modify the analysis parameters of the scheme.

The study starts with the introduction and explanation of the Interactive Onset Detection strategies proposed; then we introduce the evaluation scheme proposed, data, and figures of merit considered for assessing the proposed interactive strategies; after that, the results are presented and analyzed; finally, a brief discussion is included for summarizing the relevant points observed.

5.3.1 Interactive methodologies proposed

In the context of onset detection, user interaction should *adapt* the performance of the system by changing the parameters involved in the ODF and/or OSF processes. Due to the previously shown influence of the OSF stage in the overall success of the onset detection task (cf. Section 5.1), in this work we assume that the detection errors are exclusively produced by considering an inappropriate configuration of a given OSF. Although we are aware that this constitutes a simplification, there is strong evidence to restrict the work to this hypothesis.

The premise introduced is that the OSF process may not be properly parameterized: a particular OSF configuration may not be suitable for the entire $O(t)$ due to factors as, for instance, changes in instrumentation, dynamics, articulation, and so on. Thus, a given ODF should be examined by an OSF particularly tuned and adjusted for different regions. These regions would be defined by the user as the FP and FN errors are pointed out, and the new local OSF parameters are estimated through the interactions.

As of OSF on which to implement the interactive methods we shall restrict ourselves to variations of the strategy of finding local maxima above or equal to a certain threshold θ in function $O(t)$. The idea is that, while the local maximum condition is kept unaltered, threshold θ now becomes a function $\theta \equiv \theta(t)$ whose value is defined by both the feedback provided by the user and according to one of the interactive policies to be explained.

Note that given that user interactions may not match the actual local maxima in the ODF, the system needs to provide a particular temporal tolerance window. Thus, given an interaction at time point t_{int}, the energy value retrieved from the ODF for the adaptation process is given by:

$$O(t_{int}) \equiv \max\left\{O(t_m)\right\} \quad \text{with} \quad t_m \in [t_{int} - W_T,\ t_{int} + W_T] \tag{5.6}$$

where W_T represents the tolerance window considered. We consider a window of $W_T = 30\ ms$ since, as pointed out by Böck et al. (2012), this time threshold represents a proper tolerance for human beings to perceive onset events.

Exceptionally, Eq. 5.6 may retrieve a value $O(t_{int}) = 0$ in the tolerance time lapse. This issue occurs when the ODF process has not obtained a proper $O(t)$ representation and some onsets are not represented by a peak in this function. In those cases, the correction is performed (the onset is added) but the threshold value is kept unaltered.

Note that, given the time dependency in the output of an onset detection algorithm, we may assume the same premise as in the Interactive Sequential Pattern Recognition (ISPR) framework introduced in Chapter 3: when the user interacts at position t_{int} of the $O(t)$, all information located at time frames $t < t_{int}$ is implicitly validated. Corrections are therefore only required in time frames $t \geq t_{int}$.

After this introduction to the general framework, we shall now explain the two interactive correction policies proposed.

Threshold-based interaction

This first policy, which was initially presented in Iñesta and Pérez-Sancho (2013) bases its performance on directly modifying the threshold value θ of the OSF. In this case, the global threshold is substituted by an initial (*static*) proposal θ_0, and whenever the user interacts with an onset o_{int} (either an FP or an FN) located at a time frame t_{int}, its energy $O(t_{int})$ is retrieved. This figure, once modified by a small value ϵ compared to the variation range in $O(t)$, becomes the new threshold θ_{int} for the new detection process that will be performed for $t \geq t_{int}$:

$$\theta_{int} = \begin{cases} O(t_{int}) - \epsilon & \text{if} \quad o_{int} \notin (\hat{o}_i)_{i=1}^{L} \quad (\text{FN}) \\ O(t_{int}) + \epsilon & \text{if} \quad o_{int} \in (\hat{o}_i)_{i=1}^{L} \quad (\text{FP}) \end{cases} \tag{5.7}$$

where ϵ has been set to 0.001 for this work, as it constitutes a value an order of magnitude lower than the sensibility considered for the $O(t)$ functions.

Figure 5.4 shows an example of the threshold variation as a result of the different interactions performed by the user.

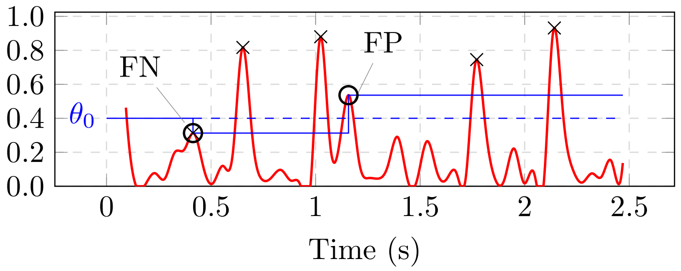


Figure 5.4: Evolution of threshold $\theta(t)$ throughout time as the result of user interaction in the sliding window threshold-based approach: symbol $\otimes$ shows the ground truth onsets while $\bigcirc$ represents the performed interactions. Dashed and solid lines represent the static and interactive thresholds obtained with the sliding window approach, respectively. Initial percentile $\theta(t_i = 0)$ has been set to 50th (median value).

Percentile-based interaction

This second approach is inspired by the idea of using an adaptive threshold for assessing the ODF. As previously introduced, a typical method for doing so consists of using an analysis window around the target point in $O(t)$ and setting as the, now variable, threshold $\theta(t)$ the median value of the window.

In our case, instead of using the median value of the sample distribution, we find useful the use of other percentiles for setting the threshold. The idea is that when the user performs an interaction at time frame t_{int}, its energy $O(t_{int})$ is retrieved for calculating the n_{th} percentile it represents with respect to the elements contained in a W-length window around that point, i.e.:

$$n_{th} \;\Big|\; P^{(n)}\left\{O(t_{w_{int}})\right\} = O(t_{int}) \;\text{ with }\; t_{w_{int}} \in \left[t_{int} - \frac{W}{2},\; t_{int} + \frac{W}{2}\right] \tag{5.8}$$

where $P^{(n)}\{\cdot\}$ retrieves the value representing the n_{th} percentile of the sample distribution.

Then, for calculating threshold $\theta(t_i)$ for time positions $t \geq t_{int}$, the rest of the signal is evaluated with a W-length sliding window using the percentile index n_{th} obtained at the interaction point t_{int} as it follows:

$$\theta(t_i) = P^{(n)}\{O(t_{w_i})\} \text{ with } t_{w_i} \in \left[t_i - \frac{W}{2},\; t_i + \frac{W}{2}\right] \wedge t_i \in t \geq t_{int} \tag{5.9}$$

Conceptually, the premise of using this approach is that, when a correction at t_{int} is made, the particular threshold θ value is not relevant by itself but by its relation with the surrounding values. For example, if $O(t_{int})$ is a low value compared to the elements in the surrounding W-length window, the

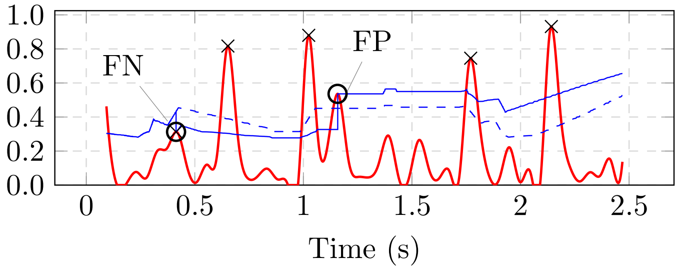


Figure 5.5: Evolution of threshold $\theta(t)$ throughout time as the result of user interaction in the sliding window percentile-based approach: symbol $\otimes$ shows the ground truth onsets while $\bigcirc$ represents the performed interactions. Dashed and solid lines represent the static and interactive thresholds obtained with the sliding window approach, respectively. Initial percentile $\theta(t_i = 0)$ has been set to 50th (median value).

successive analysis windows should use low θ values as well, which can be obtained by using low percentiles. On the other hand, if $O(t_{int})$ is high compared to the surrounding elements, the percentile should be high. Ideally, this approach should adapt the performance of the OSF to the particularities of the ODF.

The duration of the W-length window has been set to cover 1.5 seconds, using as a reference the work by West and Cox (2005) in which windows ranging from 1 to 2 seconds were used.

Figure 5.5 graphically shows the evolution of threshold θ when using this approach.

5.3.2 Experimental configuration

In order to assess the proposed interactive strategies, the scheme shown in Fig. 5.6 has been implemented. First of all, the input data is processed by an *Initial Onset Detection* algorithm (an ODF method that computes an $O(t)$ function and a OSF algorithm that processes it) retrieving a list of estimated onsets $(\hat{o}_i)_{i=1}^{L}$; both the $O(t)$ signal and the estimations $(\hat{o}_i)_{i=1}^{L}$ are the input to the *User Interaction* process. In that last stage, the user validates and interactively corrects those estimations. Note that in our experiments, to avoid the need for a person to manually carry out the corrections, ground truth annotations were used to automate the process as in other works addressing interactive methodologies (Toselli et al., 2011).

Onset Detection and Selection Functions considered

We considered a representative set of ODF methods to cover the different paradigms introduced in Chapter 2 with the aim of exhaustively assessing

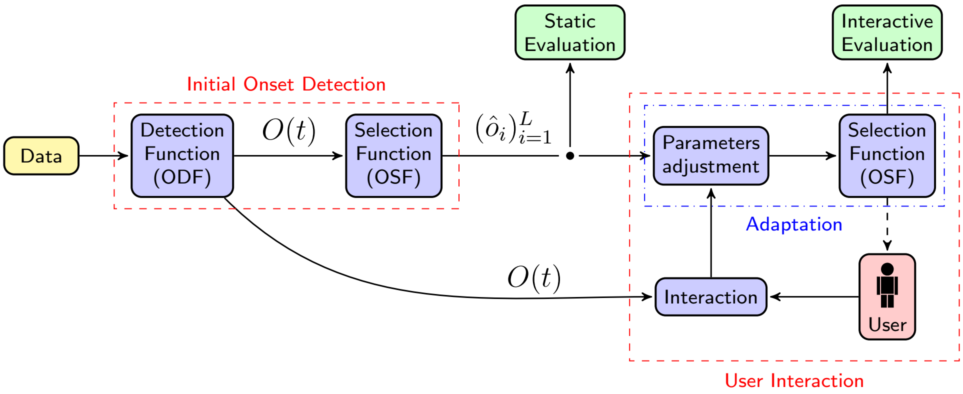


Figure 5.6: Scheme proposed for the evaluation of the signal processing interactive onset detection methods: an initial onset detection is performed on the input signal *(Data)* in the *Initial Onset Detection* block; *Static Evaluation* assesses the performance of the stand-alone algorithm; the *User Interaction* block introduces human verification, interaction and correction; *Interactive Evaluation* assesses the performance of the interactive scheme.

and validating the behaviour of the proposed interactive methodologies with different analysis principles. The precise algorithms studied are:

1. **Sum of Magnitudes (SM)**: This approach bases its performance on measuring changes directly in the energy of the signal. Using the magnitude part of the spectrogram of the signal, this process estimates the energy for each analysis window as the sum of the magnitude component of each frequency bin (Stowell & Plumbey, 2007).

2. **Power Spectrum (PS)**: This approach also bases its performance on measuring changes in energy. The approach is identical to the previous one but performing the sum of the squared value of the magnitude components of the spectrogram (Stowell & Plumbey, 2007).

3. **Semitone Filter Bank (SFB)**: This energy-based algorithm analyses the evolution of the magnitude spectrogram assuming a harmonic sound is being processed. The algorithm applies a harmonic semitone filter bank to each analysis window of the magnitude spectrogram and retrieves the energy of each band (root mean square value); then, consecutive semitone bands in time are subtracted to find energy differences; negative results are filtered out as only energy increases may point out onset information; finally, all bands are summed to finally obtain the detection function (Pertusa et al., 2005).

4. **Phase Deviation (PD)**: This method relies exclusively on phase information. The idea is that discontinuities in the phase component of

the spectrogram may depict onsets. With that premise, this approach basically predicts what the value of the phase component of the current frame should be using the information from previous frames; the deviation between that prediction and the actual value of the phase spectrum models this function (Bello et al., 2004).

5. **Weighted Phase Deviation (WPD)**: A major flaw in the previous phase method is that it considers all frequency bins to have the same relevance in the prediction. This severely distorts the result as low energy components that should have no relevance in the process are considered equal to more relevant elements. In order to avoid that, each phase component is weighted by the correspondent magnitude spectrum value (Dixon, 2006).

6. **Complex Domain Deviation (CDD)**: Extends the principle introduced in the *Phase Deviation* algorithm by estimating both magnitude and phase components for the analysis window at issue using the two preceding frames and assuming steady-state behaviour with a complex domain representation. The difference between the prediction and the actual value of the frame defines the function (Duxbury et al., 2003).

7. **Rectified Complex Domain Deviation (RCDD)**: In the *Complex Domain Deviation* method no distinction in the type of deviation between the predicted spectrum and the one at issue is made. In such case, the algorithm does not distinguish between energy rises, which depict onsets, and energy decreases, which point out offsets. Hence, a slight modification based on half-wave rectification is performed on the method to avoid tracking offsets. The difference between predicted and real values is now carried out when the spectral bins increase their energy along time; in case the energy decreases, a zero is retrieved (Dixon, 2006).

8. **Modified Kullback-Leibler Divergence (MKLD)**: This approach also measures energy changes between consecutive analysis frames in the magnitude spectrum of the signal. The particularity of this approach lies in the use of the Kullback-Leibler divergence for measuring such changes, which allows tracking large energy variations while inhibiting small ones (Brossier, 2006).

9. **Spectral Flux (SF)**: This function depicts the presence of onsets by measuring the temporal evolution of the magnitude spectrogram of the signal. The idea is obtaining the bin-wise difference between the magnitude of two consecutive analysis windows and summing only the positive deviations for retrieving the detection function (Masri, 1996).

10. **SuperFlux (SuF)**: Modifies the *Spectral Flux* method by substituting the difference between consecutive analysis windows by a process of tracking spectral trajectories in the spectrum together with a maximum filtering process. This allows the suppression of vibrato articulations in the signal which generally tend to increase false detections in classic algorithms (Böck & Widmer, 2013a, 2013b).

Given the different principles in which the presented processes are based on, the resulting $O(t)$ functions may not span for the same range. Thus, normalization is applied as a post-process so that all of them lie in the range $O(t) \in [0, 1]$. The analysis parameters of all these algorithms have been configured to a window size of 92.9 ms with a 50 % of overlapping factor.

Regarding OSF processes, based on Eq. 5.1 we selected two methods for obtaining threshold θ. These two methods are:

1. **Global threshold**: Manually setting a constant threshold $\theta = \theta_o$ for analyzing the entire $O(t)$ function.

2. **Sliding window with percentile index**: Using W-length sliding window to analyze $O(t)$ with a time-dependent threshold $\theta \equiv \theta(t)$. More precisely, we use $\theta(t_i) = \mathcal{P}^{(n)}\{O(t_{w_i})\}$, where $t_{w_i} \in \left[t_i - \frac{W}{2},\ t_i + \frac{W}{2}\right]$. Window size W has been set to 1.5 seconds considering the results in West and Cox (2005).

In order to assess the influence of the parameterization of the considered OSF methods, 25 values equally distributed in the range $[0, 1]$ have been used as either threshold or normalized percentile index.

Finally, it must be pointed out that these OSF methods are equivalent to the interactive policies in Section 5.3.1. This has been intentionally done as we want to assess two different configurations in this experimentation: on one hand using the same selection functions for both the static onset detection and the interactive scheme; on the other hand, using different selection functions for both stages.

Dataset and assessment figures

The data collection considered for the evaluation is the one introduced in Böck et al. (2012) and already used in Section 5.1. However, as pointed out and discussed in the same paper, these precise onset annotations (raw onsets) do not necessarily represent the human perception of onsets despite being musically correct. Thus, as this work addresses the human effort in the annotation/correction of onsets, the dataset was processed following the process described in the previous reference: all onsets within 30 ms were combined into one located at the arithmetic mean of their single positions. This process reduced the total number of elements from an initial figure of $27,774$ events to $25,996$ onsets (approximately, 81 onsets per file). For

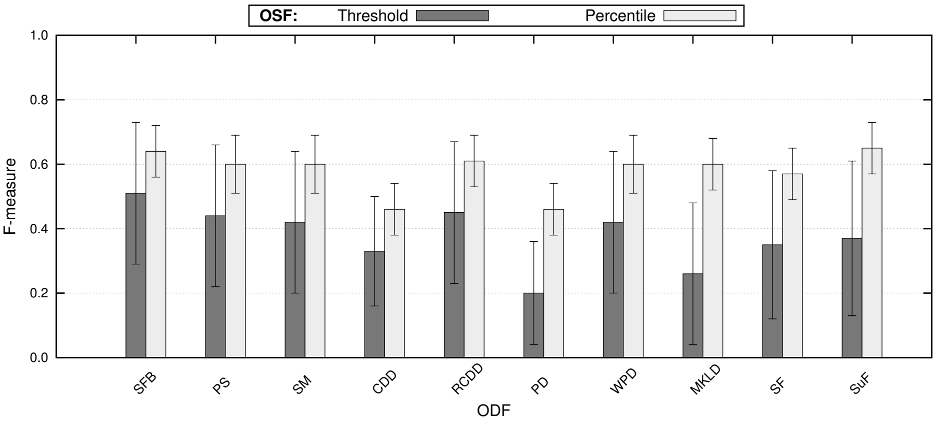


Figure 5.7: Graphical representation of the F_1 results obtained for the static evaluation of the different Onset Detection Functions and Onset Selection Functions considered.

our experiments no partitioning in terms of instrumentation, duration, or polyphony degree was done to the data, as the idea is to check the usefulness of the interactive approach disregarding the nature of the data.

Regarding the evaluation figures, on the one hand we have have considered the classic onset detection evaluation criteria based on the Precision, Recall, and F-measure figures of merit introduced in Chapter 2. The only particularity is that we have reduced the tolerance window to a more restrictive value of 30 ms to match the conditions of the processed dataset.

On the other hand, as the aim of the work is to assess the usefulness of the interactive schemes introduced, thus we have considered the effort-based measures introduced in Section 5.2. As in the previous set of measures, we fixed the tolerance window to 30 ms.

5.3.3 Results

We now present the results obtained when assessing the interactive proposals with the evaluation procedures considered. For each particular pair of ODF and OSF plus either the manual correction or the interactive scheme at issue, the figure of merit shows the average and standard deviation of the 25 OSF initial settings.

Results obtained in the static assessment of the considered ODF algorithms are shown in Table 5.8. Additionally, Fig. 5.7 graphically shows the results obtained but restricted to the F_1 metric.

Figures achieved by the different configurations considered show the intrinsic difficulty of the dataset: focusing on the F_1 score, results are far from being perfect as all the scores are lower than 0.6. In that sense, the PD method showed the lowest accuracy, possibly due to exclusively

Table 5.8: Results obtained in terms of Precision, Recall, and F-measure for the static evaluation of the Onset Detection Functions and Onset Selection Functions considered for the signal processing interactive schemes.

ODF	**OSF**	**Precision**	**Recall**	**F-measure**
SFB	Threshold	0.82 ± 0.12	0.4 ± 0.2	0.5 ± 0.2
	Percentile	0.63 ± 0.10	0.64 ± 0.13	0.59 ± 0.07
PS	Threshold	0.69 ± 0.07	0.4 ± 0.2	0.4 ± 0.2
	Percentile	0.65 ± 0.06	0.57 ± 0.12	0.55 ± 0.08
SM	Threshold	0.66 ± 0.07	0.4 ± 0.2	0.4 ± 0.2
	Percentile	0.64 ± 0.06	0.56 ± 0.12	0.55 ± 0.08
CDD	Threshold	0.36 ± 0.03	0.18 ± 0.11	0.19 ± 0.10
	Percentile	0.33 ± 0.02	0.26 ± 0.06	0.26 ± 0.05
RCDD	Threshold	0.70 ± 0.08	0.4 ± 0.3	0.4 ± 0.2
	Percentile	0.63 ± 0.07	0.61 ± 0.13	0.57 ± 0.08
PD	Threshold	0.29 ± 0.02	0.17 ± 0.15	0.14 ± 0.11
	Percentile	0.35 ± 0.02	0.37 ± 0.10	0.32 ± 0.06
WPD	Threshold	0.66 ± 0.06	0.4 ± 0.2	0.4 ± 0.2
	Percentile	0.64 ± 0.06	0.56 ± 0.12	0.54 ± 0.08
MKLD	Threshold	0.45 ± 0.16	0.3 ± 0.3	0.2 ± 0.2
	Percentile	0.61 ± 0.07	0.63 ± 0.14	0.56 ± 0.08
SF	Threshold	0.53 ± 0.10	0.3 ± 0.2	0.30 ± 0.19
	Percentile	0.51 ± 0.08	0.55 ± 0.12	0.48 ± 0.06
SuF	Threshold	0.93 ± 0.08	0.3 ± 0.3	0.4 ± 0.2
	Percentile	0.67 ± 0.11	0.74 ± 0.15	0.64 ± 0.08

relying on phase information and its reported disadvantage of considering all frequency bins equally relevant. Methods such as SFB or SuF showed good responses as, although mostly relying on an energy description of the signal, the information is processed in very sophisticated ways to avoid estimation errors.

In general terms, the relatively high precision scores achieved suggest that FP may not be the most common type of error in the considered systems. However, recall scores were low, especially when considering the global threshold selection process, thus pointing out a considerable amount of FN errors.

These results also show the clear advantage of adaptive threshold methods in the OSF when compared to a global initial value. In general, the former paradigm achieved better detection figures with lower deviation values than the latter, thus stating its robustness.

Once we have gained a general insight of the performance of the considered ODF and OSF schemes, we shall study them from the interactive point of

Table 5.9: Comparison of the user effort invested in correcting the initial estimation of static onset detectors in terms of the R_{GT} for the signal processing interactive schemes. The F_1 column shows the performance of the static detection method, whereas $\mathrm{R}_{\mathrm{GT}}^{man}$ refers to the effort invested when considering a complete manual correction of the results. $\mathrm{R}_{\mathrm{GT}}^{thres}$ and $\mathrm{R}_{\mathrm{GT}}^{pctl}$ stand for the user effort in the threshold-based and percentile-based correction approaches, respectively. Symbol † denotes the cases in which the deviation is lower than the second significant decimal figure.

ODF	**OSF**	$\mathbf{F}_1$	$\mathbf{R}_{\mathbf{GT}}^{man}$	$\mathbf{R}_{\mathbf{GT}}^{thres}$	$\mathbf{R}_{\mathbf{GT}}^{pctl}$
SFB	Threshold	0.5 ± 0.2	0.41 ± 0.05	$0.34 \pm 0.01^{\dagger}$	0.43 ± 0.02
	Percentile	0.59 ± 0.07	0.45 ± 0.03	$0.34 \pm 0.01^{\dagger}$	0.44 ± 0.01
PS	Threshold	0.4 ± 0.2	0.44 ± 0.03	$0.37 \pm 0.01^{\dagger}$	0.43 ± 0.01
	Percentile	0.55 ± 0.08	0.44 ± 0.02	$0.37 \pm 0.01^{\dagger}$	$0.43 \pm 0.01^{\dagger}$
SM	Threshold	0.4 ± 0.2	0.45 ± 0.03	$0.38 \pm 0.01^{\dagger}$	$0.43 \pm 0.01^{\dagger}$
	Percentile	0.55 ± 0.08	0.45 ± 0.01	$0.38 \pm 0.01^{\dagger}$	$0.44 \pm 0.01^{\dagger}$
CDD	Threshold	0.19 ± 0.10	0.54 ± 0.04	$0.51 \pm 0.01^{\dagger}$	$0.57 \pm 0.01^{\dagger}$
	Percentile	0.26 ± 0.05	0.57 ± 0.02	$0.52 \pm 0.01^{\dagger}$	$0.57 \pm 0.01^{\dagger}$
RCDD	Threshold	0.4 ± 0.2	0.44 ± 0.04	$0.35 \pm 0.01^{\dagger}$	0.44 ± 0.02
	Percentile	0.57 ± 0.08	0.45 ± 0.02	$0.36 \pm 0.01^{\dagger}$	0.44 ± 0.01
PD	Threshold	0.14 ± 0.11	0.54 ± 0.04	$0.52 \pm 0.01^{\dagger}$	$0.59 \pm 0.01^{\dagger}$
	Percentile	0.32 ± 0.06	0.59 ± 0.03	$0.52 \pm 0.01^{\dagger}$	$0.59 \pm 0.01^{\dagger}$
WPD	Threshold	0.4 ± 0.2	0.45 ± 0.03	$0.38 \pm 0.01^{\dagger}$	$0.43 \pm 0.01^{\dagger}$
	Percentile	0.54 ± 0.08	0.45 ± 0.01	$0.38 \pm 0.01^{\dagger}$	$0.44 \pm 0.01^{\dagger}$
MKLD	Threshold	0.2 ± 0.2	0.48 ± 0.02	$0.36 \pm 0.01^{\dagger}$	$0.46 \pm 0.01^{\dagger}$
	Percentile	0.56 ± 0.08	0.46 ± 0.02	$0.36 \pm 0.01^{\dagger}$	$0.46 \pm 0.01^{\dagger}$
SF	Threshold	0.30 ± 0.19	0.48 ± 0.02	$0.44 \pm 0.01^{\dagger}$	$0.46 \pm 0.01^{\dagger}$
	Percentile	0.48 ± 0.06	0.52 ± 0.03	$0.44 \pm 0.01^{\dagger}$	$0.47 \pm 0.01^{\dagger}$
SuF	Threshold	0.4 ± 0.2	0.42 ± 0.07	$0.26 \pm 0.01^{\dagger}$	0.40 ± 0.03
	Percentile	0.64 ± 0.08	0.42 ± 0.07	$0.26 \pm 0.01^{\dagger}$	0.40 ± 0.02

view. Table 5.9 and Fig. 5.8 introduce the effort results in terms of the *Corrections to Ground Truth ratio* (R_{GT}) measure when considering the manual and interactive corrections of the errors.

As an initial remark, it can be seen that the workload figures for manual correction ($\mathrm{R}_{\mathrm{GT}}^{man}$) are close to a value of 0.5 for all the ODF and OSF considered. These results suggest that an initial onset estimation process is indeed beneficial for lowering the manual annotation since such figures depict that half of the total number of onsets are properly handled by the autonomous detection system. The reported low deviation values also suggest that only for some particular cases in which the OSF parameters are not properly selected, the required effort may be higher.

In terms of the threshold-based interaction scheme, there is a consistent workload reduction when compared to the manual procedure. Figures obtained are almost always under the 0.5 value, getting to the point of 0.26 for the SuF algorithm (which broadly means annotating just a fourth of the

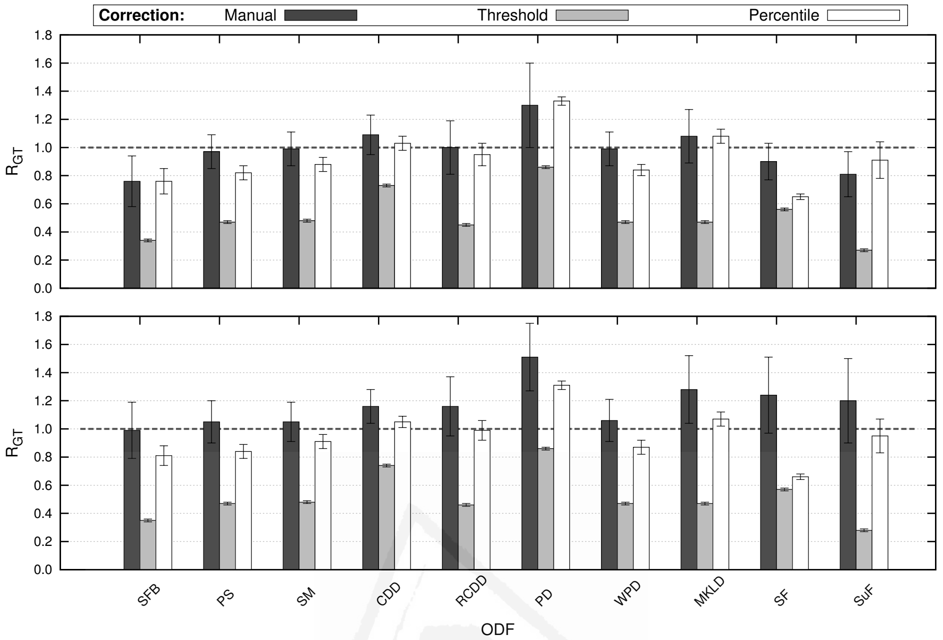


Figure 5.8: Graphical representation of the user effort results obtained in terms of the R_{GT} measure for the manual correction and the threshold-based and percentile-based interactive schemes. Top and bottom figures represent the results obtained when considering either threshold-based or percentile-based Onset Selection Functions, respectively.

total number of onsets), showing the workload reduction capabilities of the scheme. Additionally, the very low standard deviation values obtained point out the robustness of the method: independently of the initial performance of the ODF and OSF at issue, the threshold-based interaction scheme consistently solves the task within a fixed figure of effort. This fact could suggest that, when considering this scheme, the performance of the initial onset estimation by the autonomous algorithm may not be completely relevant as the interactive scheme is able to solve the task within the same figure of effort.

Regarding the percentile-based scheme, the effort figures obtained are clearly worse than in the case of the threshold-based scheme, with up to 0.14 points of difference between the two schemes for this measure, and are qualitatively similar to the figures by the manual correction. This premise can be also seen in the deviation values obtained: in spite of being quite low, in some cases these figures show less consistency than in the threshold-based approach (e.g., SuF or RCDD); nevertheless, it should be noted that when compared to the manual correction, percentile-based interaction shows a superior robustness since for this scheme the deviation figures are consistently

Table 5.10: Results in terms of the R_{TC} measure for the signal processing interactive methodologies for the different onset detectors considered. R_{xy} represents each R_{TC} score, where x refers to the Onset Selection Function used and y to the interactive approach.

ODF	$\mathbf{R}_{\mathbf{TT}}$	$\mathbf{R}_{\mathbf{PT}}$	$\mathbf{R}_{\mathbf{TP}}$	$\mathbf{R}_{\mathbf{PP}}$
SFB	0.75 ± 0.11	0.69 ± 0.07	1.3 ± 0.2	1.00 ± 0.14
PS	0.77 ± 0.06	0.80 ± 0.03	1.10 ± 0.11	0.98 ± 0.05
SM	0.78 ± 0.06	0.80 ± 0.03	1.11 ± 0.12	0.99 ± 0.06
CDD	0.92 ± 0.11	0.83 ± 0.06	1.2 ± 0.2	1.00 ± 0.09
RCDD	0.73 ± 0.07	0.73 ± 0.04	1.3 ± 0.3	1.01 ± 0.11
PD	0.96 ± 0.14	0.79 ± 0.09	1.4 ± 0.3	1.0 ± 0.2
WPD	0.77 ± 0.05	0.80 ± 0.03	1.10 ± 0.11	0.97 ± 0.06
MKLD	0.69 ± 0.04	0.72 ± 0.05	1.2 ± 0.2	1.01 ± 0.13
SF	0.87 ± 0.06	0.78 ± 0.08	0.99 ± 0.07	0.86 ± 0.10
SuF	0.54 ± 0.12	0.56 ± 0.07	1.61 ± 0.2	1.0 ± 0.2

lower than those obtained when considering the manual approach.

Finally, the results obtained in terms of the *Total Corrections* ratio are shown in Table 5.10 and Fig. 5.9. This figure of merit helps us to compare the different interactive configurations among them to gain some insights about their differences in behavior.

Checking the figures obtained, and disregarding the initial selection function, the threshold-based interaction scheme (R_{xT}) clearly outperforms the percentile-based one (R_{xP}) as the R_{TC} results are always lower in the former one. In the same sense, threshold-based figures always achieved values under the unit whereas the other scheme was clearly not capable of doing so. Deviation figures also proved threshold-based interaction as more robust, given that in general they were lower than the ones obtained in the percentile-based scheme.

Focusing on the threshold-based schemes, it can be seen that scores (both in terms of average and deviation) were quite similar independently of the initial selection methods (OSF). This fact suggests that this straightforward modification of the threshold value could be considered a rather robust method capable of achieving good effort figures independently of the estimation given by the initial selection process (OSF).

On the contrary, attending to the difference in the results among the percentile-based interaction schemes, the initial estimation has a clear influence for this type of interaction. As observed, using an initial selection process (OSF) based on either threshold or percentile, results in terms of the R_{TC} get to diverge in 0.3 points (case of SFB) or even 0.6 points (as in SuF). Thus, given the dependency of this interaction scheme with the initial

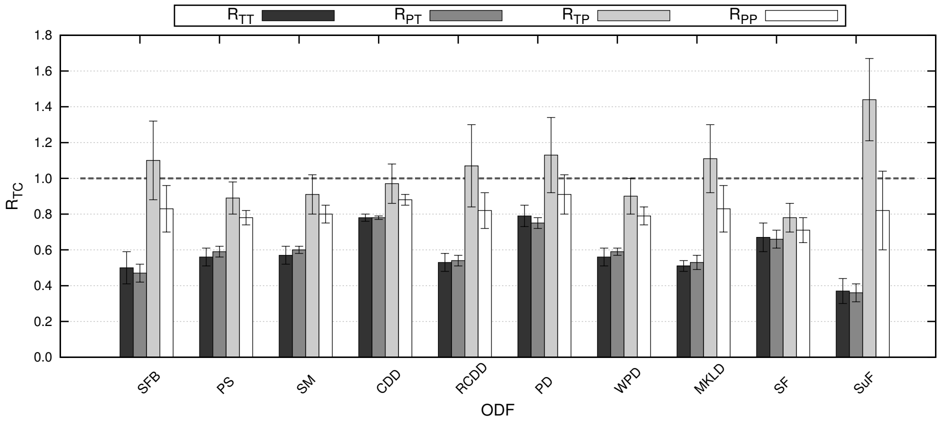


Figure 5.9: Graphical representation of the user effort measure R_{TC} for all the combinations of Onset Selection Function and signal processing interactive schemes. $\mathrm{R}_{\mathrm{TC}} = 1$ is highlighted.

selection process (OSF), results suggest that the best particular configuration for this percentile-based interaction approach is the case in which the initial static selection is based on percentile as well, i.e. R_{PP}.

Statistical significance analysis

To statistically assess the reduction of the user effort, a Wilcoxon rank-sum test (Demšar, 2006) has been performed comparing each interactive method proposed against manual correction. This comparison has been performed considering the *Corrections to Ground Truth ratio* (R_{GT}) values. Table 5.11 shows the results when considering a significance $p < 0.05$.

Figures obtained show that threshold-based interaction significantly reduced the correction workload when compared to the manual correction. It is especially remarkable the fact that this approach consistently reduced the user effort for all the combinations of ODF and OSF methods considered.

Results for the percentile-based interaction also show that for most of the cases there was a significant reduction in terms of workload. However, this statistical evaluation also proves that, for some particular configurations as for instance CDD with the percentile-based OSF or the SuF with the global threshold OSF, this interactive scheme may not be useful if percentiles are used for adapting the system from the user corrections, as the resulting workload does not significantly differ from the manual correction. In addition, a particular mention must be done to the SFB, CDD, and PD algorithms with the global threshold OSF as they constitute the particular cases in which the interactive algorithm implies more user effort than the manual correction.

Finally, figures obtained with this statistical analysis state the robustness

Table 5.11: Statistical significance results of the user effort invested in the correction of the detected onsets using the signal processing interactive schemes. Manual correction ($\mathrm{R}_{\mathrm{GT}}^{man}$) is compared against the threshold-based ($\mathrm{R}_{\mathrm{GT}}^{thres}$) and percentile-based ($\mathrm{R}_{\mathrm{GT}}^{pctl}$) interactive correction methods. Symbols ✓, ✗, and = state that effort invested with the interactive methodologies is significantly lower, higher or not different to the results by the manual correction. Significance has been set to $p < 0.05$.

ODF	**OSF**	$\mathbf{R}_{\mathbf{GT}}^{thres}$ *vs* $\mathbf{R}_{\mathbf{GT}}^{man}$	$\mathbf{R}_{\mathbf{GT}}^{pctl}$ *vs* $\mathbf{R}_{\mathbf{GT}}^{man}$
SFB	Threshold	✓	✗
	Percentile	✓	✓
PS	Threshold	✓	✓
	Percentile	✓	✓
SM	Threshold	✓	✓
	Percentile	✓	✓
CDD	Threshold	✓	✗
	Percentile	✓	=
RCDD	Threshold	✓	=
	Percentile	✓	✓
PD	Threshold	✓	✗
	Percentile	✓	=
WPD	Threshold	✓	✓
	Percentile	✓	✓
MKLD	Threshold	✓	✓
	Percentile	✓	✓
SF	Threshold	✓	✓
	Percentile	✓	✓
SuF	Threshold	✓	=
	Percentile	✓	✓

of the threshold-based interaction when compared to the percentile-based scheme: while results for the former method consistently presented a reduction in workload, the latter one did not show such steady behavior.

5.3.4 Discussion

This section introduced a set of strategies based on signal processing for the interactive correction of automatically detected onset events in audio streams. As a general conclusion it can be pointed out that the use of interactive and adaptive systems for the annotation and correction of detected onset events clearly entails a reduction in the effort and workload invested by the user in the task. This is even more remarkable given that the strategies presented are simply based on modifying the threshold that the OSF stage applies to the result of an ODF process.

In particular, we proposed and assessed two strategies for interactively modifying this threshold value: a first one in which a user interaction directly changes this level by setting it to the energy level of the point at which the interaction is performed; and a second one in which the new threshold value is set according to the percentile that the energy of the interaction point represents in a temporal window around that specific point.

Experiments show that, in general, both strategies imply a remarkable decrease in terms of workload to the user to correct the results of an initial onset detection process. Moreover, this assertion is consistently certain for the first of the interactive schemes whereas for the second one this does not happen as some of the configurations equal, or even outrange, the amount of interactions to perform if compared to manually annotating the entire set.

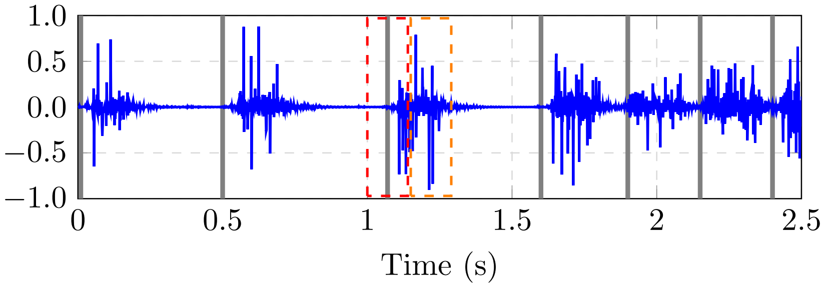
5.4 Interactive Pattern Recognition for Onset Detection

A clear limitation of the previous interactive model is that the OSF progressively adapts it performance to fit the signal at issue, but the underlying onset estimation model is neither modified nor improved. Hence, each time a new piece has to be annotated or corrected, the model obviates the cases and particularities *learned* from pieces processed previously (i.e., errors pointed out by the user in other music pieces) and always starts the task considering the same initial onset estimation model.

In such context it seems interesting to further extend this interactive scheme to methods capable of modifying the base onset estimation model to some extent. In this new approach, the annotation/correction process of a piece is not only affected by local interactions done to analysis parameters of the signal at issue but also by the historical of corrections done in pieces previously evaluated. For that we shall consider Pattern Recognition (PR) models, and more precisely the Interactive Sequential Pattern Recognition (ISPR) framework as described in Chapter 3 due to the time-wise nature of the data, since they are capable of modifying its performance by simply changing the elements in the training set.

Onset detection may be modeled as a binary classification task: the signal is evaluated in a frame-wise fashion in which each frame represents an instance to be classified as either containing an onset event or not. Each of those frames is described in terms of a set of low-level features derived, in principle, from both its temporal and spectral representations. Figure 5.10 graphically shows this idea.

Having introduced the general context, the rest of this section further



$$\mathbf{x}_t = [x_0, x_1, ..., x_N, \text{‘onset’}]$$

$$\mathbf{x}_{t+\Delta T} = [x_0, x_1, ..., x_N, \text{‘non-onset’}]$$

Figure 5.10: Conceptual description of onset detection as a classification task. Onset events are represented as grey vertical lines whereas two analysis windows are depicted by dashed lines. The text below shows the description of the aforementioned analysis windows in terms of their features and class.

explores the idea of onset detection with PR models with a focus on the interactive onset estimation/correction paradigm. More precisely, the rest of the section comprises two different parts: *(i)* an initial one devoted to compare different PR models for addressing the task of onset detection as a stand-alone system (no human-computer interaction); and *(ii)* a second part that extends the model derived from the previous point to the task of interactive onset estimation by proposing and assessing a set of methodologies for properly updating the PR model.

5.4.1 Static approach

The aim of this first part is to derive a PR-based model for stand-alone onset detection by performing a comparative study of different features and classifiers. Nevertheless, note that the main overall point of the section is to obtain a PR-based model that suits the interactive onset estimation/correction paradigm. Thus, this study shall also consider other evaluation parameters beyond classification performance (or, in this case, goodness of the onset estimation) such as the cost of (re)training the model and the time efficiency of the task.

As an initial point to comment, we shall introduce the data considered for the forthcoming experiments. We have considered five data corpora chosen for their availability and to provide a high degree of timbral diversity. Table 5.12 provides a summary of their most relevant points while a thorough description is now provided:

1. **Saarland Music Data**: Collection of 50 piano pieces (audio and MIDI aligned) recorded with a Disklavier and gathered by Müller,

Konz, Bogler, and Arifi-Müller (2011).

2. **RWC-jazz**: Set of 15 monotimbral performances extracted from the RWC jazz database (Goto, Hashiguchi, Nishimura, & Oka, 2002) with the manual annotation of their onset events (Box, 2013). Out of the 15 pieces, 5 of them correspond to piano recordings from the Jazz Music collection (RWC-MDB-J-2001-M01) and the remaining 10 to synthesized MIDI sequences (piano and guitar timbres) from the same collection.

3. **Prosemus**: Data collection consisting of 19 files that result from a mixture of some elements from the RWC database and other real-life recordings, covering a wide range of instruments and genres. It can be freely downloaded from `http://grfia.dlsi.ua.es/cm/projects/prosemus/index.php`.

4. **Trios**: Dataset created by Fritsch (2012) which contains five multi-track recordings of short musical extracts from trio pieces. The mixed version of each trio as well as the isolated instruments are supplied. A MIDI version of the manually aligned audio file and its synthesized version are also provided. In our experiments we have only considered the single instruments.

5. **Leveau**: This corpus gathered by Leveau, Daudet, and Richard (2004) consists of mixture between certain elements of the RWC set and some recordings made in an anechoic chamber by the authors of the dataset. It comprises solo performances of monophonic and polyphonic instruments, and was part of the MIREX 2005 onset detection task.

Note that not all experiments to perform shall consider all the datasets presented as in some cases we aim at obtaining general and qualitative conclusions rather than quantitative figures.

Table 5.12: Description of the onset description datasets for studying the features considered for the classification-based onset estimator.

Collection	Files	Duration	Onsets
Saarland	50	4 h. 43 m. 16 s.	151,207
RWC-jazz	15	56 m. 34 s.	11,553
Prosemus	19	8 m. 43 s.	2,155
Trios	17	11 m. 4 s.	1,813
Leveau	11	2 m. 39 s.	428
Total	112	6 h. 2 m. 16 s.	167,156

Focusing now on the characterization of the signal (i.e., the obtention of the individual instances), we perform a frame-based analysis using Hann windows with a size of 92.8 *ms* and a 50 % of overlapping, which results in a temporal resolution of 46.4 *ms*. We characterize each of the instances with the set of 10 low-level signal descriptors listed in Table 5.13. These features are obtained with the MID.EDU Vamp plugin by Salamon and Gómez (2014) and, while a basic description of the features is provided, the reader is referred to the work by Peeters, Giordano, Susini, Misdariis, and McAdams (2011) for their comprehensive description. In addition, we consider the first-order derivatives of those descriptors as their temporal evolution may provide supplementary information for the onset estimation.

Table 5.13: Low-level signal descriptors considered for the classification-based onset estimation model grouped in either temporal or spectral analysis. Note that their first-order derivatives are included as additional descriptors.

	Feature	**Description**
Time	Zero Crossing Rate (ZCR)	Number of zero-axis crosses
	Root Mean Square (RMS)	Energy of the signal
Frequency	Spectral Centroid	Center of gravity
	Spectral Spread	Spread around mean
	Spectral Skewness	Asymmetry around mean
	Spectral Kurtosis	Flatness around mean
	Spectral Flux	Magnitude variation in time
	Spectral Flatness	Geometric mean over arithmetic mean
	Spectral Crest	Maximum value over arithmetical mean
	Spectral Rolloff	Frequency below which 95 % of the energy is contained

Once we have introduced the descriptors considered, we shall assess their performance when using a set of different classification models in the context of onset detection. For that we consider four different strategies commonly used in PR tasks:

1. **Decision Tree:** Non-parametric classifier that performs the separation of the classes by iteratively partitioning the search space with simple decisions over the features in an individual fashion. The resulting model may be represented as a tree in which the nodes represent the individual decisions to be evaluated and the nodes contain the classes to assign. In our case we consider the J48 algorithm based on the C4.5 implementation by Quinlan (2014).

2. **Multilayer Perceptron (MLP):** Particular topology of an artificial neural network parametric classifier. This topology implements a feed-forward network in which each neuron in a given layer is fully-connected to all neurons of the following layer. The configuration in this case is a single-layer network comprising 100 neurons and a softmax layer for the eventual decision (Duda et al., 2001).

3. ***k*-Nearest Neighbor (*k*NN):** Standard *k*-Nearest Neighbor algorithm with the particularity of being implemented as a *k*-dimensional tree structure (Bentley, 1975). For this classifier we consider values of $k = 1, 3,$ and 5 neighbors with Euclidean distance.

4. **Approximate *k*NN:** Implementation of an approximate *k*NN algorithm with the Fast Library for Approximate Nearest Neighbors (FLANN) by Muja and Lowe (2014). This algorithm constructs a series of random *k*-dimensional trees and performs the classification process by searching through them. In this case we consider values of $k = 1, 3,$ and 5 neighbors, 8 randomized trees, and set 64 leafs to be checked for each search.

Having introduced the classification strategies to compare and the set of features considered, we shall now assess them. For that we consider the RWC-jazz set and compare the different classifiers in three situations: using the set of 10 features in Table 5.13, considering the same set of 10 features but normalized to their respective global maximum respectively, and the case considering the normalized features plus their temporal derivatives (20 features). The results obtained in terms of the F-measure considering a 10-fold cross validation scheme are shown in Fig. 5.11.

As it may be checked, the performance achieved by the different classifiers does not remarkably differ among them. Given this equality in the results, and as the main aim is to obtain a model easy to be updated and retrained for the forthcoming interactive methodologies, we shall restrict ourselves to the use of *lazy learning* techniques and more precisely to the approximate *k*NN search algorithm studied FLANN.

Regarding the different sets of features assessed, it can be checked that the set depicting the highest performance is the one with the 20 descriptors (the initial ones after the normalization process plus their temporal derivatives); the use of the initial 10 features shows a clear performance decrease when compared to the former set of features, possibly due to the lack of the temporal information provided by the derivatives; finally, when considering the normalized version of the initial 10 features, this performance decrease is further accused.

While the performance obtained by the best feature set may be considered sufficient, we shall further study and select the most relevant ones out of it. Thus, to study the discrimination capabilities of each feature, we consider

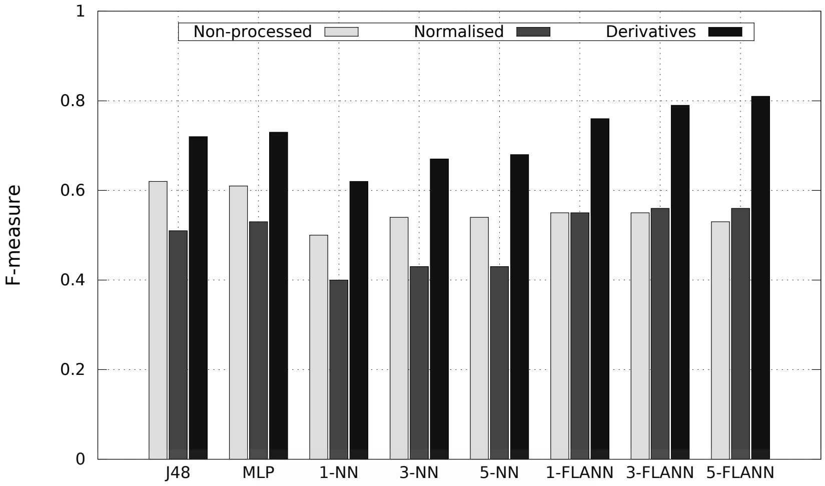


Figure 5.11: Comparison of different classification models for onset estimation. For each classifier, the results show the case of using the initial descriptors, their normalized version, and their normalized version plus their first-order derivatives.

a method based on the statistical significance Z-score test as in Ponce de León and Iñesta (2007). This method scores the discrimination capabilities of each feature according to their means and variances assuming a binary classification scenario. Mathematically, z_n represents the separation score of feature f_n obtained using the following equation:

$$z_n = \frac{\left|\bar{f}_n^{(1)} - \bar{f}_n^{(0)}\right|}{\sqrt{\frac{\sigma^2\left\{f_n^{(1)}\right\}}{N_1} - \frac{\sigma^2\left\{f_n^{(0)}\right\}}{N_0}}} \tag{5.10}$$

where $\bar{f}_n^{(x)}$ represents the mean value of feature f_n for the elements with class x, $\sigma^2\left\{f_n^{(x)}\right\}$ depicts the variance of feature f_n for the elements with class x, and N_x represents the total number of instances with class x.

Table 5.14 shows the results obtained by the 20 descriptors considered ranked according to their Z score obtained for each of the five datasets previously introduced. Each Z value represents the average of the individual Z scores obtained by each file in the set. In addition, the Mean Reciprocal Rank (MRR)[2] is included for summarizing the information of all datasets.

The results of this Z-score test may report some insights which may be helpful in the context of onset detection: for instance, it may be checked

[2] For each feature f_n, the Mean Reciprocal Rank (MRR) is obtained as $\text{MRR}(f_n) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{r_i(f_n)}$, where N stands for the number of different ranks to average and $r_i(x)$ for the position of element x in rank i.

Table 5.14: Onset estimation results obtained by the 20 descriptors considered ordered according to their Z score obtained for each of the 5 datasets. The *(Der)* tag stands for the first-order derivative of the descriptor. Each Z value represents the average of the individual Z scores obtained by each file in the set. The Mean Reciprocal Rank (MRR) is used for summarizing the information of all datasets.

Leveau		**Saarland**		**Prosemus**		**RWC-jazz**		**Trios**		**MRR**
Descriptor	**Z-score**	**Descriptor**	**Z-score**	**Descriptor**	**Z-score**	**Descriptor**	**Z-score**	**Descriptor**	**Z-score**	**Descriptor**
Crest	6.18	Spread	17.88	Flux	8.91	Crest (Der)	25.29	Flatness	17.14	Flux
Crest (Der)	5.59	Flux	16.91	Flux (Der)	8.14	Flux (Der)	23.93	Centroid	14.09	Crest (Der)
Flux (Der)	5.12	RMS	16.47	Skewness	5.82	RMS (Der)	21.69	Spread	13.59	Spread
Flux	4.78	Flatness	15.55	Spread	5.81	Flux	21.66	Rolloff	12.48	Crest
Kurtosis	4.58	Crest (Der)	15.39	Centroid	5.70	Crest	21.51	ZCR	11.76	Flux (Der)
Spread	4.28	RMS (Der)	15.01	Kurtosis	5.70	Skewness (Der)	21.37	RMS	10.12	Flatness
Kurtosis (Der)	4.22	Flux (Der)	14.62	Crest	5.69	Kurtosis (Der)	19.91	Crest (Der)	9.82	Centroid
Skewness (Der)	4.06	Crest	14.10	Flatness (Der)	5.38	Centroid (Der)	17.59	Crest	8.96	RMS
Skewness	4.03	ZCR	10.63	Spread (Der)	5.15	RMS	16.55	Skewness	8.91	RMS (Der)
Flatness	4.00	Rolloff	9.88	Flatness	5.03	Spread (Der)	16.39	Flux (Der)	7.65	Skewness
RMS	3.98	Skewness (Der)	8.14	Skewness (Der)	4.94	Kurtosis	15.65	Flux	6.86	Kurtosis
RMS (Der)	3.80	Spread (Der)	8.07	Crest (Der)	4.84	Skewness	15.32	Kurtosis	6.59	Skewness (Der)
Centroid	3.74	Kurtosis (Der)	6.69	RMS	4.78	Spread	14.64	RMS (Der)	5.86	Rolloff
Flatness (Der)	3.35	Centroid	6.30	Centroid (Der)	4.36	Centroid	14.34	Spread (Der)	5.41	Kurtosis (Der)
Spread (Der)	3.33	Skewness	6.02	Kurtosis (Der)	4.13	ZCR	12.23	Kurtosis (Der)	4.85	ZCR
Centroid (Der)	2.65	Flatness (Der)	5.99	Rolloff	3.59	Rolloff	10.53	Flatness (Der)	4.71	Spread (Der)
ZCR	2.21	Centroid (Der)	4.98	Rolloff (Der)	3.22	Rolloff (Der)	10.01	Skewness (Der)	4.53	Flatness (Der)
Rolloff	1.92	ZCR (Der)	4.68	RMS (Der)	3.11	ZCR (Der)	9.78	ZCR (Der)	3.84	Centroid (Der)
Rolloff (Der)	1.51	Kurtosis	4.40	ZCR	2.95	Flatness (Der)	9.41	Centroid (Der)	3.31	Rolloff (Der)
ZCR (Der)	1.46	Rolloff (Der)	4.26	ZCR (Der)	1.78	Flatness	7.47	Rolloff (Der)	2.69	ZCR (Der)

that some descriptors like the Spectral Flux or the Spectral Crest (which are ranked in the first positions of the MRR) may be more useful than the Spectral Rolloff or the Zero Crossing Rate. Nevertheless, this Z-score test does not report the actual influence of selecting a particular subset out of the general on the performance of the system. In this sense, Fig. 5.12 shows the results of considering the kNN classifier for the task of onset detection using different groups of features from the MRR rank: taking the top-ranked four features and then increasing the set size in groups of four features following the rank order. A 10-fold cross validation scheme over the five aforementioned datasets has been considered.

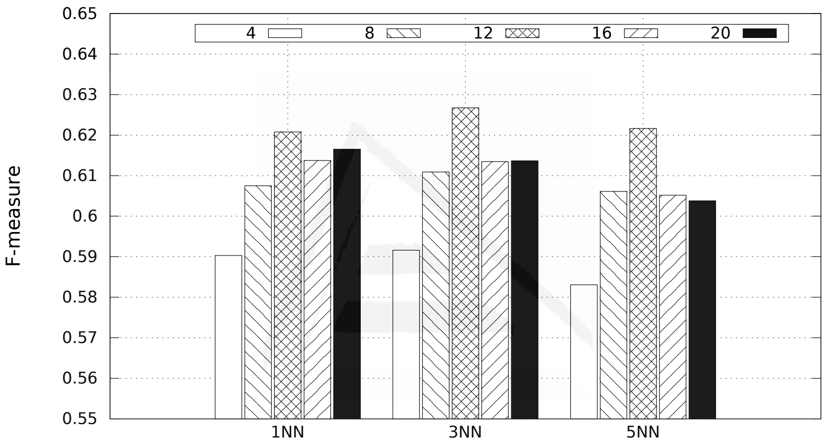


Figure 5.12: Difference in performance when considering different feature sets for onset estimation based on k-Nearest Neighbor. The label in the abscissa axis represent the different number of neighbors considered.

According to the results obtained, the performance of the onset estimation system is maximized when considering the first 12 descriptors of the rank. Thus, we shall consider this particular subset for the rest of the section instead of the space of 20 features initially considered.

Instance optimization

As commented in Chapter 3, the kNN classifier shows very low efficiency figures as no model is derived out of the initial data. Thus, when the training set is excessively large, the classification process generally becomes quite time consuming and not suitable for interactive tasks. In this regard we shall explore the use of different strategies for reducing the number of instances in the training set and thus speed up the process with, in principle, no loss in the detection figure of merit.

In the context of kNN, the most typical framework consists in applying Prototype Selection (PS) processes as the ones introduced in Chapters 3 and 4. The point with such techniques is that they aim at optimizing a particular figure of merit (generally, training set size) disregarding the nature of the data (that is, prior information of the data domain is not considered). Nevertheless, as in this section we are dealing with the precise case of onset detection, we may take advantage of the particular characteristics of the domain at issue, time series representing music pieces, and propose a set of techniques explicitly designed for selecting the proper instances for training the classifier in the context of onset detection.

Before proposing and discussing this domain-based instance selection techniques, we shall introduce some notation. Let us assume that $\mathcal{T} = \left[\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_{|\mathcal{T}|}\right]$ represents the ordered vector of instances with length $|\mathcal{T}|$ obtained from the time-frequency analysis of an audio file from which we shall select the proper instances to be included in the training set of the classifier. Let also $\mathcal{P}_{on}$ be the ordered vector that represents the elements in $\mathcal{T}$ labeled as onset, and $\mathcal{P}_{non}$ the ordered vector of elements labeled as non-onset.

The idea is to obtain an alternative training set $\mathcal{T}'$ that comprises all elements from the entire $\mathcal{P}_{on}$ vector, and a reduced version of the non-onset vector $\mathcal{P}'_{non} \subset \mathcal{P}_{non}$ that allows a faster computation without a remarkable decrease in the detection performance of the system. With this premise we propose following four different strategies for selecting the elements of $\mathcal{P}'_{non}$:

1. **Random**: Randomly selecting a number of prototypes from set $\mathcal{P}_{non}$.

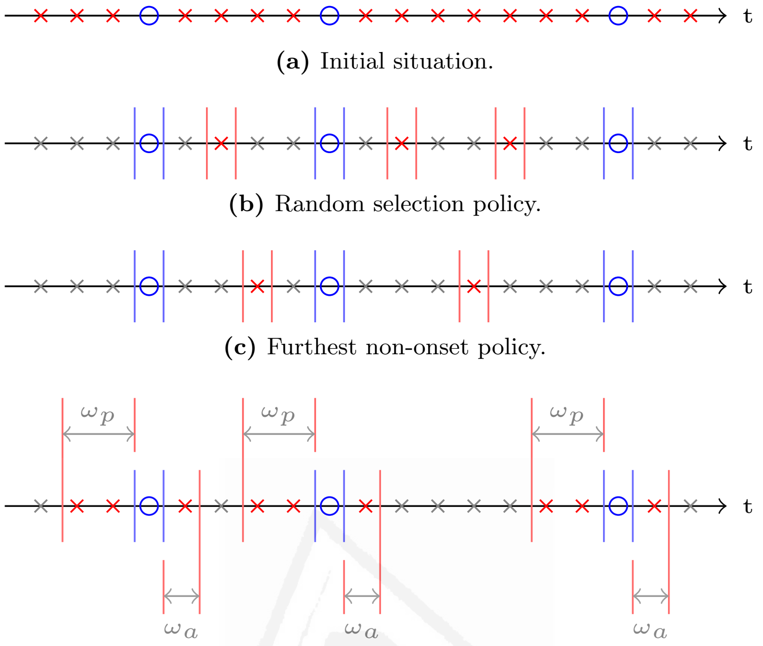
2. **Furthest non-onset**: Selecting, if exists, the most distant non-onset element between pairs of onset elements. That is, for two consecutive onset instances $\mathcal{T}_m \in \mathcal{P}_{on}$ and $\mathcal{T}_n \in \mathcal{P}_{on}$, the element to include should be $\mathcal{T}_{\lfloor \frac{m+n}{2} \rfloor} \in \mathcal{P}_{non}$.

3. **Window around onset**: Selecting non-onsets prototypes whose distance to an onset instance is less than a distance threshold w.

4. **Further non-onset plus window around onset**: Gathering the two previous criteria into one.

For a better comprehension of the techniques proposed, a graphical representation of these policies is shown in Fig. 5.13.

Having presented the different policies proposed for the domain-based instance selection stage, we shall now assess their performance experimentally. For that, we consider the RWC-jazz, Leveau, Trios, and Prosemus collections and implement a 10-fold cross validation scheme and values of $k = 1, 3, 5, 7, 9, 11, \text{and } 13$ for the kNN classifier. The results in terms of the onset detection accuracy and the obtained reduced set size are shown in Table 5.15. Bold elements represent the non-dominated solutions. These figures constitute the average of the individual figures obtained for each fold



(a) Initial situation.

(b) Random selection policy.

(c) Furthest non-onset policy.

(d) Windows around onset policy. The size of previous and posterior windows (ω_p and ω_a, respectively) may be different.

Figure 5.13: Graphical representation of the domain-based instance selection policies. Symbols ○ and × represent onset and non-onset points, respectively. For the different policies, the elements surrounded by bars represent the selected instances.

and dataset. Note that the presented results for the window-based strategies represent the configurations that maximize the onset detection accuracy for each dataset. Also note that the *Random* method has been configured to reduce the number of non-onset instances to match the amount of the onset ones.

The results obtained prove that, if properly configured, some of the proposed policies are able to remarkably reduce the set size with similar estimation capabilities to the initial training set. For instance, the *Random* policy for low k values (e.g., $k = 1, 3$) retrieves estimation figures considerably lower than the baseline; however, when parameterized as $k = 7, 9$, the results are totally equivalent to the baseline with roughly a third of the set size. The *Furthest non-onset* policy is the one achieving the highest reduction, but also shows the worst estimation performance as all the obtained results are always below the baseline for all k configurations. Note that the window-based approaches (*Windows around onset* and *Windows and furthest non-onset*) achieve the best overall estimation performance as, except for the $k = 1$ configuration, they always improve the baseline with roughly 50 % of the total

Table 5.15: Results of the domain-based instance selection policies for the *k*-Nearest Neighbor onset estimation method proposed. For each classification scheme and reduction policy, the F-measure and standard deviation of a 10-fold cross validation is provided. The baseline column represents the initial case without instance selection. The average resulting set size (in percentage with respect to the baseline case) as well as its standard deviation is also shown in the last row. Bold elements represent the non-dominated configurations.

	Classifier	**Baseline**	**Random**	**Furthest non-onset**	**Windows around onset**	**Windows and furthest non-onset**
F-measure	1NN	0.63 ± 0.09	0.56 ± 0.09	0.53 ± 0.07	0.62 ± 0.09	0.62 ± 0.09
	3NN	0.65 ± 0.10	0.60 ± 0.09	0.57 ± 0.07	0.68 ± 0.08	0.68 ± 0.09
	5NN	0.64 ± 0.11	0.62 ± 0.09	0.58 ± 0.07	0.70 ± 0.08	0.70 ± 0.08
	7NN	0.64 ± 0.12	0.64 ± 0.09	0.59 ± 0.07	0.72 ± 0.08	0.71 ± 0.08
	9NN	0.63 ± 0.12	$\mathbf{0.65 \pm 0.09}$	0.60 ± 0.07	0.72 ± 0.08	0.71 ± 0.08
	11NN	0.63 ± 0.13	0.65 ± 0.09	$\mathbf{0.61 \pm 0.07}$	$\mathbf{0.73 \pm 0.07}$	$\mathbf{0.72 \pm 0.08}$
	13NN	0.62 ± 0.13	0.65 ± 0.09	0.61 ± 0.08	0.73 ± 0.07	0.71 ± 0.08
Reduced size (%)		100.0 ± 0.0	28.6 ± 8.3	26.5 ± 7.4	54.2 ± 19.5	49.1 ± 32.3

set size. Nevertheless, note that the figures shown for these window-based approaches constitute the figures obtained for the most suitable configurations per dataset, thus requiring a prior stage of configuration and exhaustive search of parameters not required in the other policies.

The non-dominance analysis of the results reinforces the aforementioned points: the *Windows around onset* and *Windows and furthest non-onset* strategies stand out due to the remarkable performance obtained ($\text{F}_1 = 0.73$ and $\text{F}_1 = 0.72$, respectively); the *Furthest non-onset* policy is part of the non-dominated front since, in spite of not reaching the performance baseline, it achieves the highest reduction rate; finally, the *Random* strategy is clearly a compromise solution between the high performance and large set size of the window-based policies and the low performance and small set size of the *Furthest non-onset* method.

Once we have studied the proposed domain-based instance selection policies, it is necessary to also assess the performance of typical PS schemes for kNN. For that, we shall now replicate the previous study considering PS algorithms as the ones introduced and discussed in Chapters 3 and 4 instead of the previous domain-based reduction policies.

For this study we shall consider the following PS strategies: the Condensed Nearest Neighbor (CNN) and its fast and order-independent version Fast Condensed Nearest Neighbor (FCNN); the Edited Nearest Neighbor (ENN) and its combination with the condensing-based methods, Edited Condensed Nearest Neighbor (ECNN) and Edited Fast Condensed Nearest Neighbor (EFCNN); more recent techniques as the Iterative Case Filtering (ICF) and the Cross-generational elitist selection, Heterogeneous recombination and Cataclysmic mutation (CHC) methods; lastly, the rank-based algorithms Nearest to Enemy (NE) and Farthest Neighbor (FaN), configured with $\alpha = 0.10, 0.20, \text{and } 0.30$ as possible values of probability mass. For all cases we consider a kNN classifier with values of $k = 1, 3, 5, \text{and } 7$.

The results in terms of the onset detection accuracy and the obtained reduced set size are shown in Table 5.15 for the different PS methods with the same datasets as in the previous experiment. This figures constitute the average of the individual figures obtained for each fold and dataset. Bold elements represent the non-dominated solutions.

The obtained results show that, as expected, the set reduction of the PS methods generally entails a decrease in the onset estimation performance. While ENN generally maintains the performance as it aims at removing noisy prototypes, condesing-based strategies as CNN or FCNN suffer a considerable decrease in the F_1 figure of merit (around 0.1 points with respect to the ALL case). This effect is somehow palliated with the use of high k values (for instance, performance improves from a figure of $\text{F}_1 = 0.56$ witk $k = 1$ to $\text{F}_1 = 0.62$ with $k = 7$ for the CNN method) or with the use of ENN method as preprocess also improves the performance as it removes noisy prototypes (when considering the 1NN rule, performance improves from $\text{F}_1 = 0.56$ of

Table 5.16: Results of the Prototype Selection methods for the k-Nearest Neighbor onset estimation method proposed. The ALL case represents the figures obtained with the non-reduced set. For each reduction scheme, the average and standard deviation of the F-measure and obtained set size (in percentage with respect to the baseline case) resulting from of a 10-fold cross validation are provided. Bold elements represent the non-dominated configurations.

PS method	**k = 1**		**k = 3**		**k = 5**		**k = 7**	
	$\mathbf{F_1}$	**Size (%)**	$\mathbf{F_1}$	**Size (%)**	$\mathbf{F_1}$	**Size (%)**	$\mathbf{F_1}$	**Size (%)**
ALL	0.63 ± 0.09	100 ± 0.0	0.65 ± 0.10	100 ± 0.0	0.64 ± 0.11	100 ± 0.0	0.64 ± 0.12	100 ± 0.0
ENN	0.63 ± 0.11	91.7 ± 3.3	0.62 ± 0.14	92.0 ± 3.2	0.61 ± 0.13	92.0 ± 3.1	0.59 ± 0.15	92.0 ± 2.9
CNN	0.56 ± 0.08	20.8 ± 7.0	0.58 ± 0.08	20.4 ± 7.0	0.61 ± 0.07	20.0 ± 7.2	0.62 ± 0.08	19.8 ± 6.9
FCNN	0.56 ± 0.07	18.7 ± 6.5	0.60 ± 0.08	18.7 ± 6.5	0.62 ± 0.09	18.7 ± 6.5	0.62 ± 0.10	18.7 ± 6.5
ECNN	0.61 ± 0.07	6.9 ± 2.3	0.63 ± 0.08	6.2 ± 1.8	0.64 ± 0.08	5.6 ± 1.4	0.65 ± 0.09	5.3 ± 1.8
EFCNN	0.60 ± 0.07	6.0 ± 2.1	0.63 ± 0.09	4.7 ± 1.5	0.62 ± 0.10	4.2 ± 1.5	0.63 ± 0.12	4.1 ± 1.3
1-NE$_{0.10}$	0.55 ± 0.14	1.1 ± 0.5	0.58 ± 0.14	1.1 ± 0.5	0.59 ± 0.14	1.1 ± 0.5	0.59 ± 0.13	1.1 ± 0.5
1-NE$_{0.20}$	0.56 ± 0.12	2.6 ± 1.1	0.61 ± 0.11	2.6 ± 1.1	0.61 ± 0.12	2.6 ± 1.1	0.62 ± 0.12	2.6 ± 1.1
1-NE$_{0.30}$	0.58 ± 0.12	4.8 ± 2.0	0.61 ± 0.11	4.8 ± 2.0	0.62 ± 0.11	4.8 ± 2.0	0.65 ± 0.09	4.8 ± 2.0
1-FaN$_{0.10}$	0.62 ± 0.12	3.3 ± 1.0	0.66 ± 0.12	3.3 ± 1.0	0.65 ± 0.14	3.3 ± 1.0	0.65 ± 0.15	3.3 ± 1.0
1-FaN$_{0.20}$	0.64 ± 0.10	8.0 ± 1.7	0.67 ± 0.10	8.0 ± 1.7	0.67 ± 0.10	8.0 ± 1.7	0.67 ± 0.11	8.0 ± 1.7
1-FaN$_{0.30}$	0.64 ± 0.10	14.0 ± 2.4	0.66 ± 0.12	14.0 ± 2.4	0.67 ± 0.11	14.0 ± 2.4	0.66 ± 0.13	14.0 ± 2.4
ICF	0.51 ± 0.06	14.0 ± 5.4	0.58 ± 0.04	13.4 ± 5.4	0.59 ± 0.05	13.3 ± 5.5	0.60 ± 0.09	13.2 ± 5.5
CHC	0.64 ± 0.07	1.1 ± 0.4	$\mathbf{0.67 \pm 0.08}$	$\mathbf{1.1 \pm 0.5}$	$\mathbf{0.69 \pm 0.07}$	$\mathbf{1.2 \pm 0.6}$	0.68 ± 0.09	1.2 ± 0.6

the FCNN to $F_1 = 0.60$ with the EFCNN).

Rank-based methods show a remarkable competitiveness with respect to other strategies. For instance, the 1-$\text{FaN}_{0.10}$ method obtains an $F_1 = 0.62$ for the 1NN classifier with just a 3 % of the total set size and, if considering a higher probability mass of $\alpha = 0.30$ and $k = 5$, the performance improves up to $F_1 = 0.67$ with just 14 % of the set size, which actually improves the ALL case that reports an $F_1 = 0.64$ for the same 5NN classifier.

As in the condensing-based methods, the ICF scheme remarkably reduces the set size, but there is a considerable performance loss, also around 0.1 in the F_1 figure of merit. Nevertheless, the CHC method achieves one the best overall performances with one of the sharpest set size reductions: $F_1 = 0.69$ with a just 1.2 % of the set size when considering the 5NN classifier.

The non-dominance analysis of the results points out the CHC method for $k = 3$ and $k = 5$ as the cases showing the best compromise between detection performance and set size reduction (F_1 figures of 0.67 and 0.69 with slightly more than 1 % of the total set). Nevertheless, it must be mentioned that the CHC method is a computationally-expensive solution if compared to other lighter strategies such as rank-based methods which also report remarkably good performance figures with very compact set sizes.

In this first part of the section we have studied the possibility of addressing the onset detection task within a classification framework. For that we have compared a group of classifiers with different sets of descriptors and instance reduction techniques, both proposed by us and from the PR literature. From the experiments carried out we may consider some of the conclusions for the forthcoming experimentation in the context of interactive systems the following system: *i)* using the kNN as classifier as it allows an easy model updating by simply adding/removing prototypes from the training set; *ii)* considering the set of 12 descriptors selected with the Z-score test out of the initial set of 20 descriptors as it maximizes the performance for our data collections; *iii)* from the different instance reduction techniques studied, we shall consider the use of stand-alone PS methods as they require less parameterization than the domain-based techniques; as representative examples of such paradigm we shall consider the rank-based method FaN and the genetic CHC approach as they report good compromise figures in terms of onset estimation performance and set size reduction.

5.4.2 Interactive approach

In this second part of the section we take as starting point the classification-based model for onset detection previously obtained to propose and assess a set of methods for the interactive onset detection/correction task. As discussed, the advantage of considering a *lazy learning* classifier as kNN is that the model may be easily updated by simply adding and/or removing elements from its training set without the further need for a re-training process. The research question hence resides in studying which information is needed by the classifier to improve the model.

To perform such study we propose the scheme and workflow shown in Fig. 5.14: *i)* initially, the classification-based system previously proposed retrieves a list of onsets $(\hat{o}_i)_{i=1}^{L}$; *ii)* as the user corrects the errors, information is added to the *Training Set* of the onset detector, whose performance is modified; *iii)* then the detection algorithm recalculates the output; *iv)* after a number of iterations, the correct list of onsets $(o_i)_{i=1}^{N}$ is retrieved.

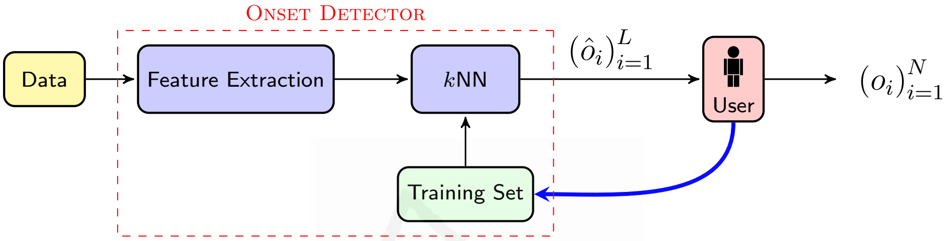


Figure 5.14: Interactive k-Nearest Neighbor (kNN) scheme for onset correction.

As aforementioned in this Chapter, the key point in the interactive framework is the sequential sense of the output: given the time dependency of the onset detection process, when the user points an error at position t_{int}, all information located at time frames $t < t_{int}$ is implicitly validated and corrections are therefore only required in time frames $t > t_{int}$. This fact is remarkably important as the user is not only pointing out an error committed by the algorithm (thus clearly stating the need for *learning* that particular case), but is also stating that all previous frames were correctly classified. In this sense, while it seems unarguable the need for including the error point with the correct label as part of the *Training Set*, the question arises with the rest of the information: should the rest of the instances update the *Training Set*? Only some of them? If only some of them are relevant, which ones? The rest of the section addresses these research questions by proposing and assessing a set of policies for updating the kNN-based onset estimation model.

Updating policies

Let us introduce some notation for the proper description of the updating policies. Assume that $\mathcal{T} = \left[\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_{|\mathcal{T}|}\right]$ represents the vector of instances with length $|\mathcal{T}|$ of the audio file being analyzed. Be t_i a user interaction at frame $\mathcal{T}[t = t_i]$ and let us assume there was a previous interaction t_{i-1} at frame $\mathcal{T}[t = t_{i-1}]$. $\mathcal{M} = [\mathcal{T}[t_{i-1}+1], ..., \mathcal{T}[t_i - 1]]$ represents the set of instances between the two interactions with length $|\mathcal{M}|$.

As aforementioned, the instance $\mathcal{T}[t_i]$ representing the interaction point

t_i is always added to the *Training Set* as it is constitutes an error committed by the system and corrected by the user, and thus it should *learn from that*. Taking that into consideration, we propose four different policies for the elements in $\mathcal{M}$:

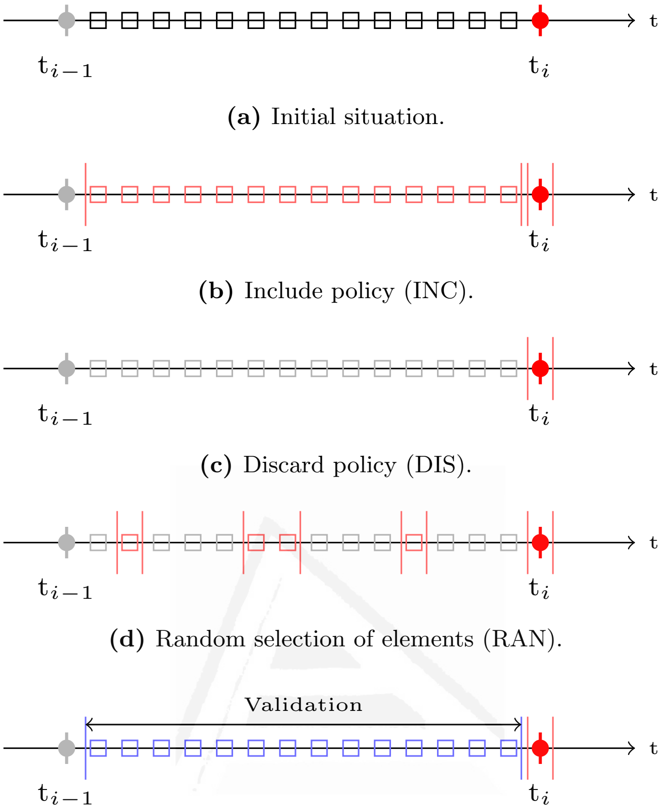
1. **Include (INC)**: All elements in $\mathcal{M}$ are included in the training set.
2. **Discard (DIS)**: None of the elements in $\mathcal{M}$ is included in the training set.
3. **Random selection (RAN)**: $\frac{|\mathcal{M}|}{2}$ elements are randomly selected from $\mathcal{M}$ and included in the training set.
4. **Validation (VAL)**: Point $\mathcal{T}[t = t_i]$ is temporary included in the set and $\mathcal{M}$ is used as a *validation set,* $\mathcal{V}$. If the prediction over $\mathcal{V}$ when including instance $\mathcal{T}[t = t_i]$ in the *Training Set* is different to the one previously obtained, point $\mathcal{T}[t = t_i]$ is eventually discarded; if the prediction remains the same, the point is maintained in the training set.

For a better comprehension of the techniques proposed, a graphical representation of these policies is shown in Fig. 5.15.

Once we have introduced the different proposals for updating the onset estimation model, we shall assess them. For that, the methodology we shall consider is the following one: *i)* for a given dataset we keep a certain percentage of the files for training and the rest for test; *ii)* only for the training files, if the precise experiment considers it, we apply a PS process and the resulting instances define the initial training set of the system; *iii)* each file of the test set is processed with one of the interactive methodologies; *iv)* the process ends when all files in the test set have been processed.

As it can be checked, this assessment scheme implies the definition of a percentage for splitting between train and test files. This parameter shall allow us to model different situations as, for instance, the case in which the training set is smaller than the test set or, just the opposite, the case in which the training data is larger than the test set. While this fact is not typically relevant in PR schemes, in the case of interactive approaches it seems important to analyze the difference in the performance when comparing a model initially trained with a significant amount of data against a model trained with less data which is progressively updated as the user performs the corrections.

Taking this into consideration, we perform an initial experiment in which we exclusively analyze the performance of the interactive policies disregarding the influence of this difference in the sizes of the sets. For that we set the aforementioned percentage to 60 % for training and 40 % for test (files are randomly selected for each set). We consider the RWC-jazz, Prosemus, Trios, and Leveau collections for the experimentation and implement a 5-fold



(a) Initial situation.

(b) Include policy (INC).

(c) Discard policy (DIS).

(d) Random selection of elements (RAN).

(e) Validation set policy (VAL).

Figure 5.15: Graphical representation of the model updating policies proposed for the classification-based onset estimation/correction paradigm. Points labeled as t_i and t_{i-1} represent the current and previous user interactions, respectively. For the different policies, the elements in red and surrounded by bars represent the selected instances.

cross-validation scheme. Table 5.17 shows the average of the results obtained for each of the considered collections in terms of the effort-based assessment figures of merit proposed in Section 5.2 with a tolerance window of 30 *ms*. The number of neighbors considered has been fixed to $k = 1, 3, 5,$ and 7, using the same k value for both the initial PS techniques and the kNN classifier. Note that interaction is again simulated by considering the ground-truth onset events as in the previous cases to avoid the need for a user in the experimentation.

An initial remark to point out is that, on average, the use of the initial static onset estimation stage reduces the annotation workload if compared to the complete manual annotation from scratch. This fact can be observed in the MAN schemes (i.e., manual correction after the initial correction) as for all cases the R_{GT} measure is lower than the unit. Nevertheless, for some

Table 5.17: Results in terms of user effort for the different interactive model-updating policies for the k-Nearest Neighbor onset estimation scheme. The *Initial model* column shows the instance selection process applied to the initial model and the *Updating policy* one shows the interactive methodology considered.

Initial model	**Updating policy**	**k = 1**		**k = 3**		**k = 5**		**k = 7**	
		R_{GT}	R_{TC}	R_{GT}	R_{TC}	R_{GT}	R_{TC}	R_{GT}	R_{TC}
ALL	MAN	0.87	1.00	0.62	1.00	0.59	1.00	0.58	1.00
	INC	0.57	0.67	0.47	0.77	0.50	0.85	0.50	0.85
	DIS	0.65	0.79	0.59	0.97	0.58	0.99	0.55	0.94
	RAN	0.65	0.77	0.51	0.83	0.51	0.87	0.48	0.83
	VAL	0.78	0.90	0.63	1.02	0.57	0.96	0.53	0.91
CHC	MAN	0.67	1.00	0.73	1.00	0.70	1.00	0.67	1.00
	INC	0.53	0.79	0.50	0.73	0.45	0.66	0.48	0.77
	DIS	0.74	1.10	0.70	1.02	0.73	1.05	0.69	1.10
	RAN	0.57	0.85	0.54	0.78	0.50	0.73	0.49	0.78
	VAL	0.71	1.05	0.73	1.06	0.66	0.96	0.68	1.07
1-FaN$_{0.30}$	MAN	0.82	1.00	0.70	1.00	0.68	1.00	0.58	1.00
	INC	0.55	0.67	0.46	0.71	0.47	0.73	0.46	0.82
	DIS	0.73	0.88	0.65	0.99	0.57	0.88	0.61	1.08
	RAN	0.66	0.79	0.51	0.77	0.49	0.76	0.49	0.86
	VAL	0.76	0.92	0.62	0.97	0.56	0.86	0.61	1.07

particular cases the average effort invested in the correction is relatively high (e.g., the ALL and 1-FaN$_{0.30}$ cases for $k = 1$ with $\mathrm{R}_{\mathrm{GT}} = 0.87$ and $\mathrm{R}_{\mathrm{GT}} = 0.82$, respectively), and thus needs to be reduced.

In general, the different policies for updating the training set report a workload reduction in this R_{GT}. The first of the policies considered, the one of including all the validated points as new elements of the training set (the INC one) reports a remarkable workload reduction with respect to the MAN one. For instance, for the case of CHC and $k = 5$ there is a reduction of almost 0.2 points in this effort measure with respect to the ALL case with the CHC strategy. While this reduction is less prominent in other cases (e.g., the 1-FaN$_{0.30}$ scheme with $k = 7$ in which there is roughly a reduction of 0.10 points), note that this workload decrease always takes place.

When discarding all the implicitly validated information (the DIS policy), the results are not that conclusive. In general, it can be checked that the use of this particular policy does not always imply a reduction in the workload with respect to the MAN case: for instance, when considering the 1-FaN$_{0.30}$ scheme with $k = 7$, the effort figure is $\mathrm{R}_{\mathrm{GT}} = 0.58$ for the MAN case and $\mathrm{R}_{\mathrm{GT}} = 0.61$ for the DIS case. This fact suggests that some of the information not being included in the training set due to being considered redundant (i.e., the implicitly validated information) is actually necessary for the system to improve its performance.

The RAN policy, which stands as a compromise solution between the two previous policies, retrieves results that somehow reflect this fact of being an *intermediate* approach. In general, the effort is generally lower than the one achieved with the DIS policy but also higher than with the INC method. For instance, this may be checked in the ALL case for $k = 5$: the INC and DIS policies report effort figures of $\mathrm{R}_{\mathrm{GT}} = 0.50$ and $\mathrm{R}_{\mathrm{GT}} = 0.58$, respectively, while the RAN one achieves a figure of $\mathrm{R}_{\mathrm{GT}} = 0.51$, which constitutes and intermediate effort figure.

In terms of the VAL policy, the effort figures obtained resemble the ones achieved by the DIS strategy. This constitutes a somehow expected behavior as the only difference between both policies is that the instance representing the interaction point may be discarded instead of being always included as a new element of the training set. While it may be checked that this policy does not generally imply a decrease in the user effort as (e.g., the INC strategy), there are some particular points to highlight. For instance, for the ALL case with $k = 7$, the VAL policy reports a sharper workload reduction when compared to the DIS one of 0.05 points. This somehow suggests that in some particular situations it may not be beneficial to include the error pointed out by the user as part of the training set. However, a totally opposite situation may be observed for the ALL case with $k = 3$ since the VAL method increases the effort in 0.04 points with respect to the DIS policy.

Finally, take the R_{TC} measure as a reference. As it can be checked most of the methods report a value in this figure of merit lower than the unit, thus reporting that most of these interactive strategies are capable of reducing the workload compared to the case of manually post-processing an initial stand-alone detection. In general, the only strategies which depict $\mathrm{R_{TC}}$ value higher than the unit are the DIS and VAL policies. This somehow confirms the previous comments on the need for including as part of the training set part of the implicitly validated information when performing the corrections.

As a last point to experiment and assess in the context of interactive onset detection task, we shall examine the influence of the train/test partitioning in the performance of the system. For that, we repeat exactly the same assessment scheme as before but using three additional splitting configurations to the one already considered: *i)* 20 % for training and 80 % for test; *ii)* 40 % for training and 60 % for test; and *iii)* 80 % for training and 20 % for test. The results obtained for this experiment in terms of the $\mathrm{R_{GT}}$ and $\mathrm{R_{TC}}$ effort figures are respectively shown in Figs. 5.16 and 5.17. For a better comprehension, for each particular scheme we only report the value achieved by the particular *k*NN configuration that minimizes each effort measure.

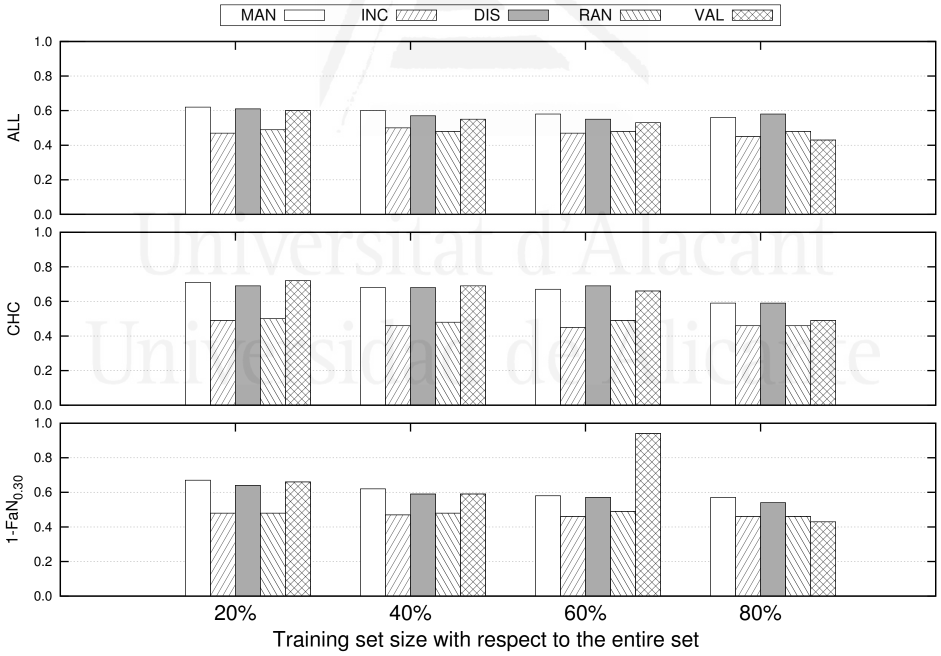


Figure 5.16: Results of the influence of the train/test partitioning in terms of the $\mathrm{R_{GT}}$ effort assessment measure. The different graphs represent each of the possible instance selection methods initially applied to the stand-alone onset detection algorithm. The legend provided depicts the interactive model updating policy followed.

Attending to Fig. 5.16, the one showing the results for the $\mathrm{R_{GT}}$ measure,

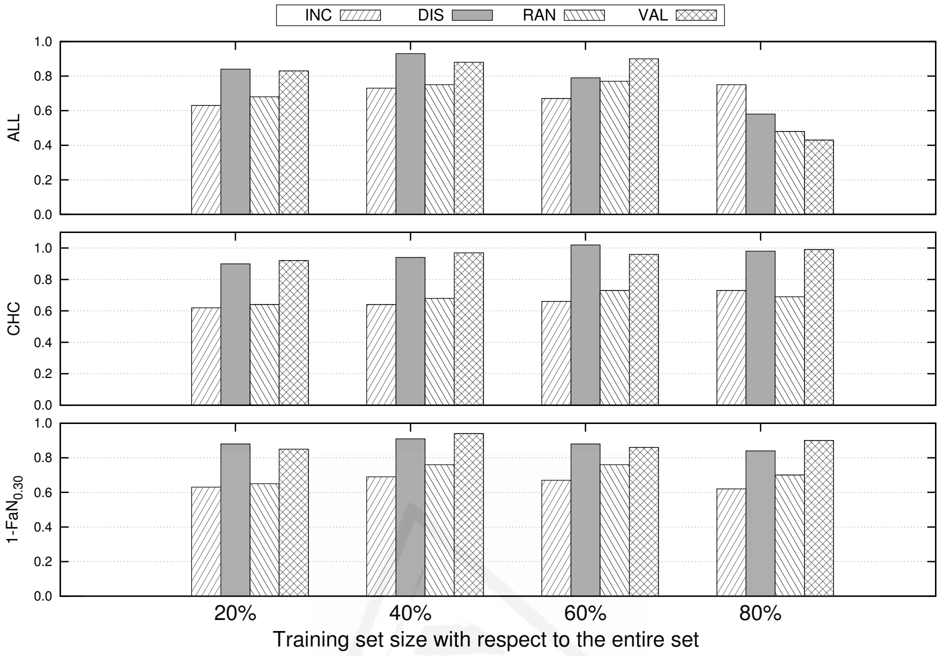


Figure 5.17: Results of the influence of the train/test partitioning in terms of the R_{TC} effort assessment measure. The different graphs represent each of the possible instance selection methods initially applied to the stand-alone onset detection algorithm. The legend provided depicts the interactive model updating policy followed. The MAN policy is omitted since, by definition, it always retrieves $R_{TC} = 1$.

several points may be highlighted. A first one is that, in general, the results suggest that the use of larger training sets implies a reduction in the workload. This point is remarkably observable in the MAN, DIS, and VAL policies as they are the policies which modify the least the initial training set, and thus remarkably depend on the initial model. Just in opposition to this tendency, the INC and RAN policies are less influenced by this initial model condition since, for each user correction, the training set is significantly modified.

As of the influence of the initial PS process, conclusions are similar to the ones already mentioned. The PS process reduces the size of the initial training set, which in some cases implies a loss in the performance and an increase in the user effort. Thus, as it can be observed, when considering any of the PS methods, the workload is increased with respect to the ALL case. This effect is especially noticeable with the CHC algorithm if compared to the 1-FaN$_{0.30}$ as the workload in the former case, which generally reports sharper reduction figure, is superior to the latter one. Nevertheless, as already commented, the INC and RAN policies stand as the most robust of all since, independently of the PS method considered, they obtained very similar R_{GT} effort figures for a fixed train/test percentage split.

Finally, results when considering the R_{TC} measure confirm the previously commented points. The initial training set remarkably influences the performance of the system since, in general, the use of a PS method for reducing the set implies an increase in the user workload. While some particular policies are able to cope with that (more precisely, the INC and RAN strategies), even obtaining similar effort values independently of the initial training set, the rest of the proposed policies are remarkably dependent on this. As aforementioned, this may be due to the fact that the INC and RAN methods remarkably modify the initial training set while the rest of the policies perform slighter modifications. In any case, note that, if properly configured, these interactive schemes imply a decrease in the user workload, which makes them interesting for real-world tasks as, for instance, corpora annotation among others.

5.4.3 Discussion

This section studied the possibility of addressing the interactive onset detection problem as an Interactive Sequential Pattern Recognition (ISPR) task, being a set of conclusions and insights important to be highlighted. A first point to comment is that the experiments carried out proved that onset detection may be addressed as a classification task in which each analysis frame is tagged as either containing or not an onset event using a set of time-based and frequency-based low-level signal descriptors. In addition, a comparative study among different classifiers showed that *lazy learning* schemes are competitive against more sophisticated techniques such as neural networks with the additional advantage of not requiring a re-training stage for updating the model, which is of particular interest in the context of interactive systems.

Another point to highlight is the proposal of a set of techniques for instance selection particularly designed for this classification-based onset detection task. The idea of the proposed techniques is based on the concept of generic Prototype Selection (PS) for the kNN classifier but particularly designed for time-series data and onset detection. The experiments carried out proved the effectiveness of this signal-based selection techniques and their competitive performance when compared to classic Prototype Selection (PS) schemes.

Finally, a set of techinques for updating the classification model have been proposed and assessed. All the proposed techniques start from the idea of always adding as a new instance of the training set the error (once corrected) pointed out by the user and consider different policies for adding or discarding the information implicitly validated by the user. Results show that the inclusion in the training set of the instances implicitly validated during the user correction stage reports a superior robustness of the scheme (i.e., not that dependent on the initial training set) and remarkably reduces the user effort when compared to the complete manual annotation.

5.5 General discussion

Onset information describes music signals in terms of the starting points of the notes events present in the audio stream. Such description of the signal constitutes a considerably relevant piece of information for a number of tasks in the Music Information Retrieval field as music transcription, rhythm description or audio signal transformations, among many others. In this context the present chapter addressed the issue of onset detection and correction from the perspective of interactive schemes: instead of the typical workflow in which the user corrects the estimation given by an onset detection algorithm, this paradigm considers the inclusion of the user as an active part of the detection with the aim of reducing the workload that the manual correction of the initial detection implies.

In terms of precise experimentation, the first section of the chapter studied the influence of the Onset Selection Function in the performance of two-stage onset detection schemes. The idea was to assess the relation between the parameterization of the selection function and the general performance of the onset detector. Such study is of particular interest as the conclusion gathered are of interest for interactive onset estimation schemes in which adaption is achieved by mapping the user corrections to the parameters of the selection function.

As a second contribution of the chapter we may highlight the proposal of a set of measures for assessing the effort invested by the users in the annotation and correction of onset events in audio streams. More precisely, two different measures have been proposed: *i)* a first one that compares the amount of corrections a user needs to perform when using an interactive system to the manual correction of the output; and *ii)* a second one that compares the amount of corrections performed by the user in relation to the total amount of ground-truth onsets. This set of measures allows the quantitave assessment of the effort invested in the process of annotating onset events in audio pieces and thus the formal comparison among future proposal that may be developed in the future.

Finally, the third contribution included in the chapter is related to the proposal and assessment of actual interactive onset annotation schemes. The set of strategies proposed may be divided in two different families: *i)* a first set based on signal processing techniques in which the methods take as starting point a two-stage onset estimation algorithm and progressively modify the onset selection stage according to the user corrections; and *ii)* a second collection of techniques based on classification schemes which incorporate the user corrections by means of modifying the training set of the method. Experiments show that both approaches are able to achieve a remarkable reduction in the user effort compared to the case of manually correcting the events estimated by a stand-alone onset detection algorithm. Nevertheless, when comparing both schemes, the strategies based on signal processing techniques generally achieve sharper workload reductions than the classification-based ones, possibly due to the representation limitations of the low-levels descriptors considered for the latter strategies.

Chapter 6

On the use of Onset Information for Note Tracking

"Science is what we understand well enough to explain to a computer. Art is everything else we do."

Donald Knuth

So far in this dissertation we have addressed the issue of interactivity for the particular case of onset estimation and correction. While this may be seen as a very particular case to consider, one of the motivations is its direct application as an additional source of information for Automatic Music Transcription systems (multimodal transcription systems).

Onset information focuses on the temporal description of the signal, and there are many example in the literature that consider it for post-processing an initial frame-level transcription for correcting timing issues found in Multi-pitch Estimation methods. Nevertheless, we find that there is a lack in formally studying several aspects, as for instance the importance of the quality of the onset information or the influence of the post-processing policy.

This chapter presents two studies in the context of onset information for improving note-level transcription. The first work addresses the issue of quantitatively assessing the improvement that is achieved when considering onset information for performing note tracking on an initial frame-level transcription as well as analyzing the relation between the goodness of the stand-alone onset detection and the quality of the note-level transcription. Then the second study proposes a novel approach for note tracking based on supervised classification and assesses it and compares the results to other benchmark proposals in the context of piano music. Finally, a last section is included to discuss the main ideas gathered from the studies.

6.1 Assessing the relevance of Onset Information for Polyphonic Note Tracking

In this first section of the chapter we study the potential of onset information for improving note tracking performance for the particular case of polyphonic piano music transcription. While onset information has previously been incorporated to Automatic Music Transcription (AMT) systems, the aim of this study is to thoroughly assess how the goodness in the estimation of onset information influences the quality of the note tracking process when used for post-processing an initial frame-based estimation obtained with an Multi-pitch Estimation (MPE) algorithm.

To develop this study we compare two different situations: on the one hand, we consider the use of ground truth onset information (oracle approach) to study a possible upper bound in the performance of the transcription system; on the other hand, we consider onset events obtained with state-of-the-art onset detection algorithms (practical approach) and compare those results with the oracle ones to point out the limitations found.

For that, we model the note tracking task as a sequence-to-sequence transduction problem (raw estimation to onset-based corrected estimation) and thus consider the use of Finite State Transducers (FSTs) for performing it. To our best knowledge the use of FSTs for note tracking constitutes a paradigm not previously considered by any author.

To carry out the proposed study, we have implemented the scheme shown in Fig. 6.1. Audio signals undergo an MPE process which outputs frame-level transcription $T_F(p,t)$, that is a binary representation depicting whether pitch p at time frame t is active. Simultaneously, onset events $(o_i)_{i=1}^{L}$ are estimated with an onset detection algorithm. Eventually both analyses are merged in a note tracking stage obtaining the note-level abstraction $T_N(p,t)$.

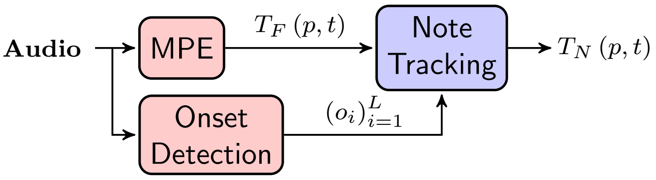


Figure 6.1: Proposed set-up for the assessment of the relevance of onset detection in note tracking.

The details concerning each of the processes shall be explained in the following subsections.

6.1.1 Multipitch estimation

We consider two MPE approaches for comparative purposes: the system by Vincent et al. (2010) based on adaptive Non-negative Matrix Factorisation (NMF) and the one by Benetos, Cherla, and Weyde (2013) based on dictionary-based Probabilistic Latent Component Analysis (PLCA). Both models output a pitch activation probability $P(p,t)$, where p stands for pitch in the MIDI scale and t for time instant. We set a temporal resolution of 10 ms for the input time-frequency representation and output pitch activation.

Vincent et al. (2010) decompose a spectrogram with an NMF-like method by modeling each template spectrum as a weighted sum of narrowband spectra that represents a group of adjacent harmonic partials. This enforces harmonicity and spectral smoothness while it allows adapting the spectral envelope to the instruments in the piece.

Benetos, Cherla, and Weyde (2013) take as input a constant-Q transform (CQT) spectrogram with a resolution of 60 bins per octave and decompose it into a series of pre-extracted log-spectral templates per pitch, instrument source, and tuning deviation from ideal tuning. Model parameters are estimated using the Expectation-Maximization (EM) method by Dempster, Laird, and Rubin (1977).

In both cases, $P(p,t)$ is further processed to obtain the $T_F(p,t)$ binary representation: $P(p,t)$ is normalized to its global maximum so that $P(p,t) \in [0,1]$ and a 7-element median filter is applied over time to smooth it. Then, the function is binarised using a threshold value $\theta = 0.1$, which is obtained taking the work in Vincent et al. (2010) as a reference and refining it for the data used in this work. Finally, a pruning stage with a minimum-length filter of 50 ms is applied to remove spurious note detections. These values were obtained by performing initial exploratory experiments to optimize the parameters for the data considered.

6.1.2 Onset detection algorithms

As mentioned, our aim is to study the influence of the onset information accuracy when considered for note tracking. Thus, we distinguish two situations: a first one considering ground-truth onset events and a second one with estimated onset information.

For the latter case we have selected three different algorithms given their good results reported in literature: Semitone Filter-Bank (SFB) by Pertusa et al. (2005), SuperFlux (SF), and ComplexFlux (CF) by Böck and Widmer (2013b, 2013a). These processes output a list $(o_i)_{i=1}^{L}$ whose elements represent the time positions of the L onsets detected. Reader is referred to Section 5.3 in this manuscript for the explanation of these methods.

The time-frequency analysis parameters of the algorithms have been set to their default values[1]. As all of them comprise a final thresholding stage (i.e., an Onset Selection Function stage), 25 different values equally spaced in the range $(0, 1)$ have been tested to check the influence of that parameter. Onset lists $(o_i)_{i=1}^{L}$ have been filtered with an averaging 30 ms to avoid overestimation issues by the algorithms following Böck et al. (2012).

6.1.3 Note tracking

$T_F(p, t)$ can be considered a set of $|\mathcal{P}|$ binary sequences of $|t|$ symbols. Hence, elements $(o_i)_{i=1}^{L}$ may be used as delimiters for segmenting each sequence $p_i \in \mathcal{P}$ in $L+1$ subsequences, resulting in a frame-level abstraction quantised by the onset information:

$$T_F(p_i, t) = T_F(p_i, 0 : o_1) \;||...||\; T_F(p_i, o_L : |t| - 1) \tag{6.1}$$

where $||$ represents the concatenation operator, p_i the pitch band at issue and L the total number of onsets.

Once onset information has been included in $T_F(p, t)$ we can process each subsequence for each pitch value $p_i \in \mathcal{P}$ separately for correcting the errors committed. For that, we have considered the use of Finite State Transducers (FSTs), a type of automaton which transforms a sequence of symbols $x_0, x_1, ..., x_N$ into another sequence $y_0, y_1, ..., y_N$ (Mohri, Pereira, & Riley, 2002). The input to the FST is each single onset-based subsequence whereas the output is another sequence in which some of the elements have been changed following a particular policy.

Given that each subsequence is a series of ones and zeros representing pitch activations and silences respectively, the two possible actions to model are either activating or deactivating sections. We focus on the former case, i.e. assuming that the MPE process misses active areas. Thus, this note tracking approach tackles the MPE issues of missing onset events in attack phases and the breaking of notes. The main reason for only tackling one of the two types of errors is to assess how beneficial can be the use of onset information for post-processing an MPE estimation when considering a very simplistic note tracking approach. This may somehow depict a lower limit in the note tracking figures that may be surpassed if more sophisticated approaches are considered.

Let the 6-tuple $\Pi = (Q, \Sigma, \Lambda, \delta, \lambda, q_1)$ define our transducer. As we are dealing with binary sequences, the input alphabet is $\Sigma = \{0, 1\}$. Its possible states are $Q = \{q_1, q_2\}$ connected with transitions $\delta(q_1, 0) = q_1$, $\delta(q_1, 1) = q_2$ and $\delta(q_2, a) = q_2$ where $a \in \Sigma$. The output alphabet $\Lambda = \{1, v_1, v_2\}$ is a non-binary representation which is parsed once the subsequence has been processed to model different FST behaviors. The outputs are given by

[1] SFB considers windows of 92.8 ms with a temporal resolution of 46.4 ms; SF and CF consider smaller windows of 46.4 ms every 5.8 ms

$\lambda(q_1, 0) = v_1$, $\lambda(q_2, 0) = v_2$ and $\lambda(b, 1) = 1$ where $b \in Q$. Finally, q_1 represents the initial state of the process. This transducer Π is graphically shown in Fig. 6.2.

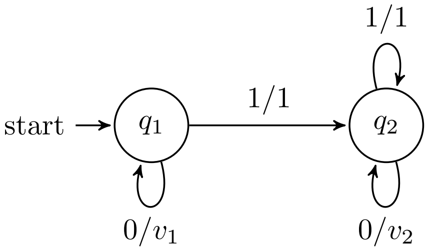
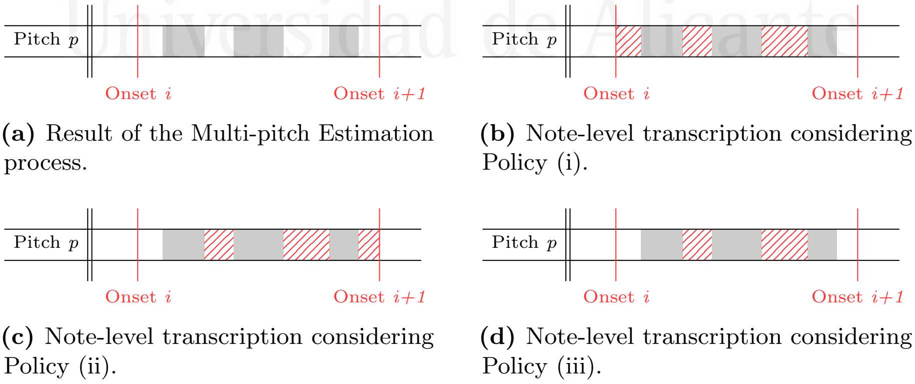


Figure 6.2: Graphical representation of the Finite State Transducer proposed for note tracking.

To model different performances of the FST we parse symbols v_1 and v_2 to values of the input alphabet Σ following three different policies. For clarity, let $\zeta(v_x) \in \Sigma$ be the first element after v_x which is different to it and let # represent the end-of-string character. All policies fill the gaps in-between active areas and two of them additionally fill other gaps which may be present. Policy (i) fixes $v_1 = 1$ and $v_2 = 1$ if $\zeta(v_2) = 1$ or, alternatively, $v_2 = 0$ if $\zeta(v_2) = \#$, which fills the possible gap between the onset and the first active area. Policy (ii) fixes $v_1 = 0$ and $v_2 = 1$, thus filling the possible gap between the last active area and the end of the sequence. Policy (iii) is equivalent to (i) but setting $v_1 = 0$, thus not filling any other type of gap. Figure 6.3 graphically shows their behavior.



(a) Result of the Multi-pitch Estimation process.

(b) Note-level transcription considering Policy (i).

(c) Note-level transcription considering Policy (ii).

(d) Note-level transcription considering Policy (iii).

Figure 6.3: Comparison of the behavior of the different Finite State Transducer configurations proposed for the note tracking process. Solid blocks represent time frames estimated as active by the Multi-pitch Estimation method whereas striped regions represent the areas filled by the Finite State Transducer.

Finally, before the FST processes the subsequences, they undergo a pruning stage of 50 ms for removing spurious detections.

6.1.4 Evaluation methodology

For assessing the proposed experience we consider the use of the MAPS database (Emiya et al., 2010) containing audio piano performances (both from real and synthesized pianos) synchronized with MIDI annotations. From that we have taken the pieces of the MUS set recorded with the Disklavier piano in both "ambient" and "close" configurations (i.e., recording microphones near and far from the source, respectively). We have also used the Saarland Music Data (SMD) collection (Müller et al., 2011) that comprises 50 piano pieces (audio and MIDI aligned) also recorded with a Disklavier. As in other AMT works (e.g., the work by Sigtia et al. (2016)), we only considered the first 30 seconds of each piece. Table 6.1 provides a summary of these sets.

Table 6.1: Description of the datasets for the study of the relevance of onset detection for note tracking in terms of the number of pieces and notes.

Collection	**Pieces**	**Notes**
MAPS-Close	30	7,353
MAPS-Ambient	30	8,764
Saarland	50	12,231

Regarding the evaluation, as we aim at assessing the relevance of using proper onset information for note tracking, we shall evaluate both tasks. For that, we consider the evaluation methodologies introduced in Chapter 2: on the one hand, we consider the standard onset assessment evaluation with a tolerance window of 50 ms; on the other hand, in terms of note tracking we shall restrict ourselves to the onset-based figure of merit as we are not considering note offsets, also with a tolerance window of 50 ms.

6.1.5 Results

Table 6.2 shows the results obtained for the onset detection process, which constitute the average and deviation of the figures obtained when evaluating each dataset using the 25 threshold values considered, in terms of Precision (P), Recall (R), and F-measure (F_1).

The high precision figures obtained state the robustness of these algorithms against false alarm detections in these data. Recall figures, though, are not that consistent: SFB commits a number of false positive errors while SF and CF seem to properly deal with them. The F_1 figures obtained show the performance limitations of these methods. For instance, the best-case scenarios are the SF and CF algorithms when tackling the MAPS-Close set

Table 6.2: Onset detection results in terms of average and standard deviation for the datasets considered for the study of the relevance of onset detection in note tracking.

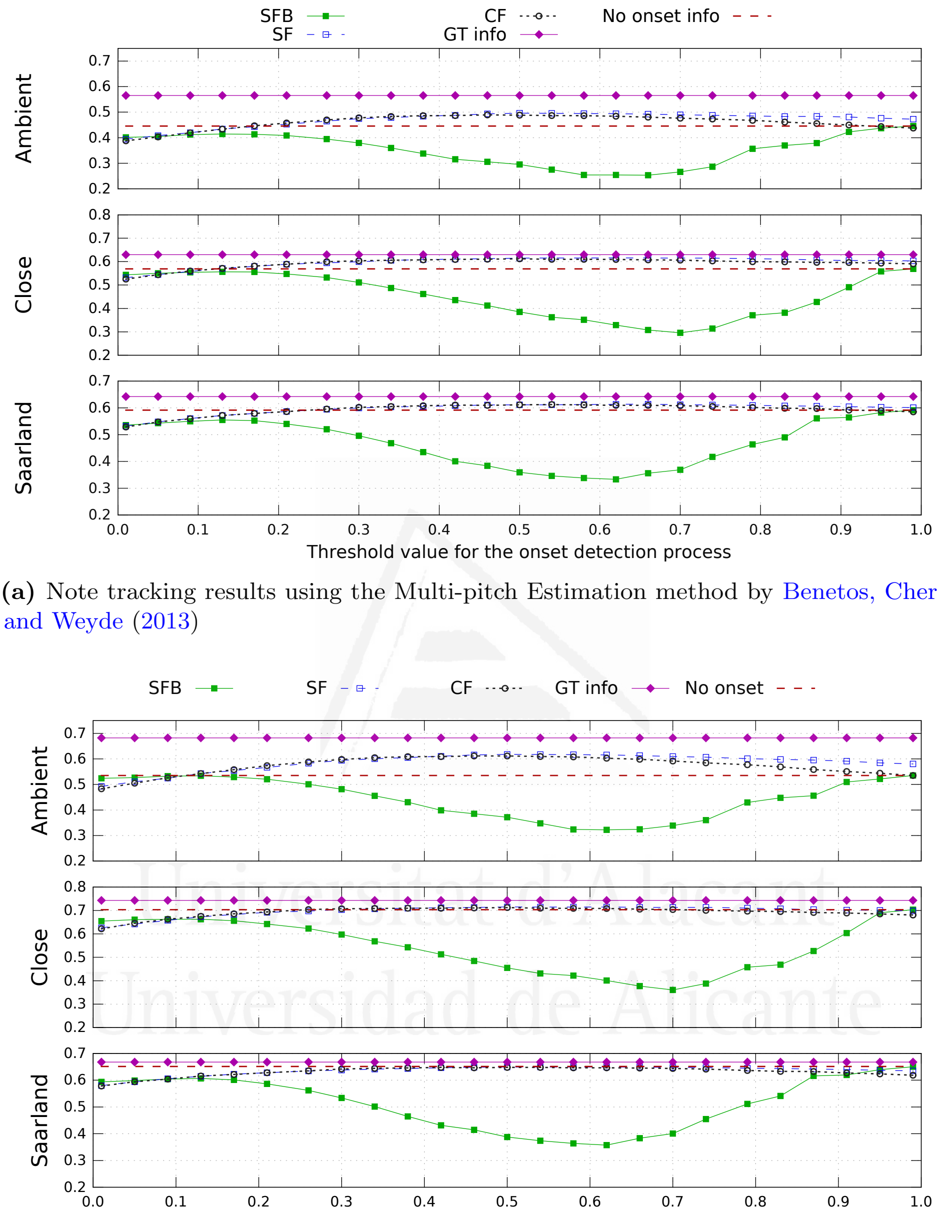
	Onset detector	**Ambient**	**Close**	**Saarland**
P	**SF**	0.78 ± 0.14	0.82 ± 0.13	0.86 ± 0.13
	CF	0.80 ± 0.14	0.84 ± 0.13	0.87 ± 0.13
	SFB	0.8 ± 0.2	0.9 ± 0.2	0.8 ± 0.2
R	**SF**	0.79 ± 0.07	0.87 ± 0.04	0.78 ± 0.05
	CF	0.76 ± 0.10	0.85 ± 0.05	0.77 ± 0.06
	SFB	0.3 ± 0.3	0.4 ± 0.3	0.3 ± 0.3
$\mathbf{F_1}$	**SF**	0.76 ± 0.07	0.82 ± 0.08	0.80 ± 0.06
	CF	0.75 ± 0.06	0.82 ± 0.07	0.79 ± 0.06
	SFB	0.4 ± 0.3	0.4 ± 0.3	0.4 ± 0.3

(average $F_1 = 0.82$, possibly due to being the dataset recorded in the most favorable conditions, i.e. close to the source) which are far from a score of 1. We shall check how this limitation affects the note tracking stage.

Figures 6.4, 6.5 and 6.6 show the note tracking results obtained for the proposed FST with Policies (i), (ii), and (iii) respectively for the two MPE schemes considered. For simplicity in the analysis, figures have been limited to the F_1 score.

Results for Policy (i) of the FST (Fig. 6.4) show that, for both MPE processes, the use of onset information for note tracking benefits the process: onsets estimated with SF and CF improve results compared to the case in which no additional information is considered. In contrast, onset information from SFB implies a decrease in performance, possibly due to the reported tendency of this algorithm to miss onset events, which may be providing inaccurate subsequences to the FST.

The performance boost observed when ground-truth onset information is provided suggests the usefulness of onset information for note tracking. Nevertheless, the actual point here is the need for accurate onset information. As shown, SF and CF improve results when compared to a simple pruning stage (e.g., an improvement around 5 % to 10 % in F_1 may be achieved in the MAPS-Ambient set depending on the MPE method with respect to the single pruning stage), but these figures are far from results achieved with ground-truth onset information (e.g., ground-truth onset information implies a further improvement of up to 5 % in F_1 on top of the improvement achieved by SF and CF in the Saarland set). Furthermore, there seems to be more room for improvement in the MAPS-Ambient set than in the rest, possibly due to being the set with the most unfavorable recording conditions (far from the source) and thus the one with the lowest figures in both onset estimation

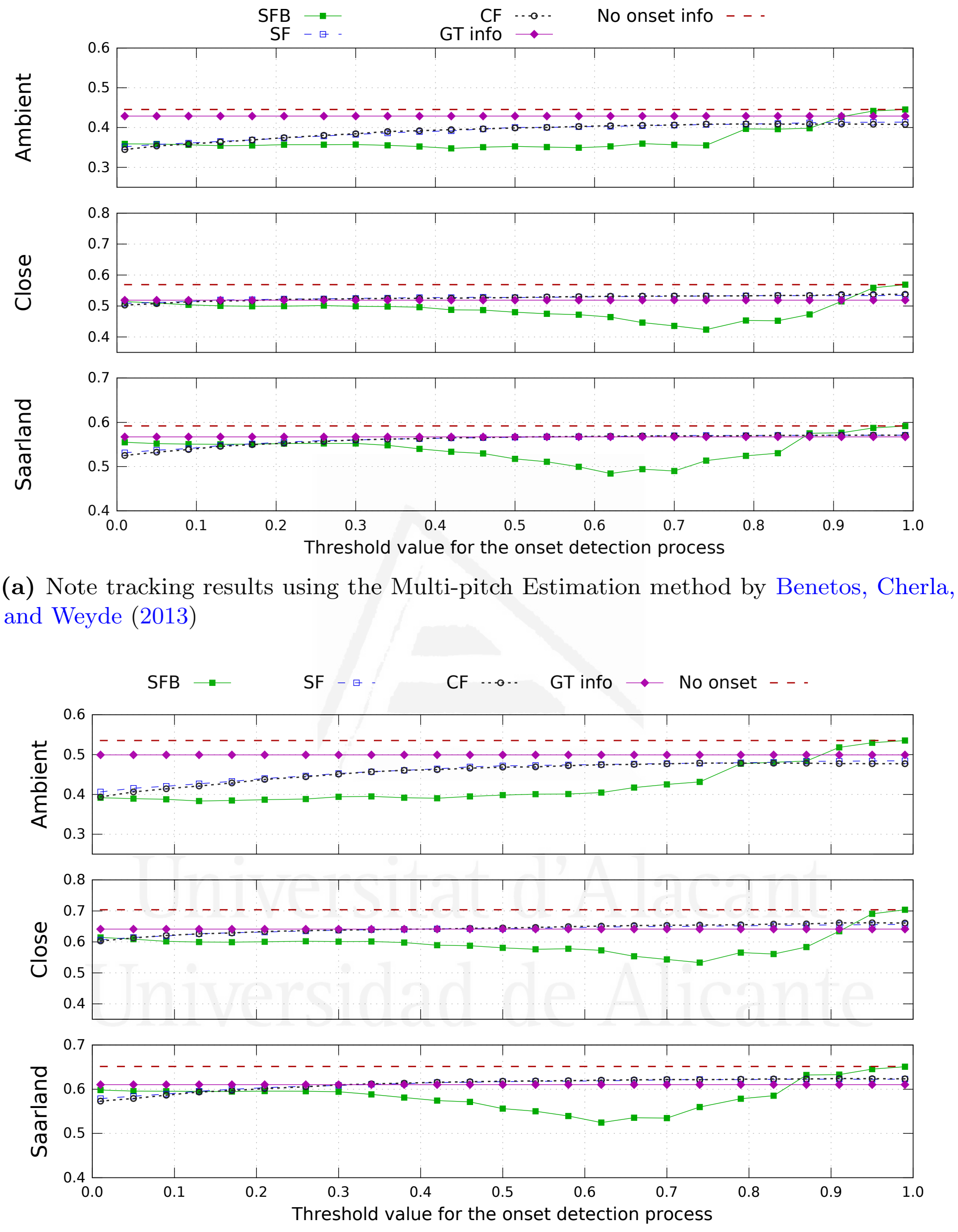


(a) Note tracking results using the Multi-pitch Estimation method by Benetos, Cherla, and Weyde (2013)

(b) Note tracking results using the Multi-pitch Estimation method by Vincent et al. (2010)

Figure 6.4: Note tracking results (F_1 score) obtained when applying Policy (i) in the Finite State Transducer for note tracking for the different Multi-pitch Estimation systems considered.

(cf. Table 6.2) and the MPE process (qualitatively reflected on the note tracking scores when not considering onset information, i.e. $F_1 \approx 0.45$). Additionally, threshold values maximizing onset estimation in the entire

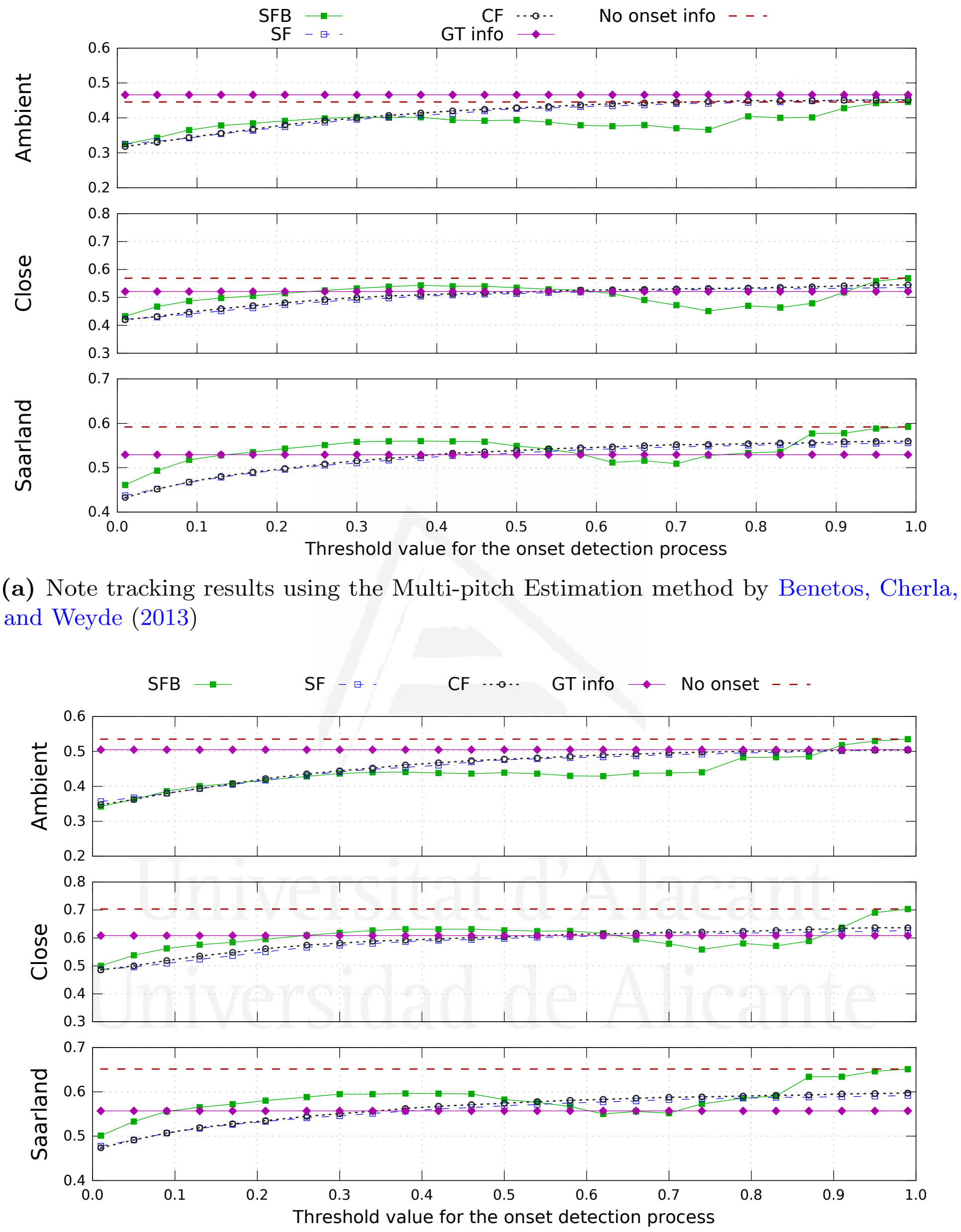


(a) Note tracking results using the Multi-pitch Estimation method by Benetos, Cherla, and Weyde (2013)

(b) Note tracking results using the Multi-pitch Estimation method by Vincent et al. (2010)

Figure 6.5: Note tracking results (F_1 score) obtained when applying Policy (ii) in the Finite State Transducer for note tracking for the different Multi-pitch Estimation systems considered.

collections (for all sets, these threshold values are around 0.5 for SF and CF, reporting $F_1 \approx 0.8$, and 0.15 for SFB, achieving $F_1 \approx 0.7$) also exhibit the maximum for note tracking results. This reveals a relation between the



(a) Note tracking results using the Multi-pitch Estimation method by Benetos, Cherla, and Weyde (2013)

(b) Note tracking results using the Multi-pitch Estimation method by Vincent et al. (2010)

Figure 6.6: Note tracking results (F_1 score) obtained when applying Policy (iii) in the Finite State Transducer for note tracking for the different Multi-pitch Estimation systems considered.

accuracy of onset information and the success of the note tracking process (i.e., the better onset detection, the better note tracking), with the ideal case being the one considering ground-truth onset information.

Figures obtained when considering Policy (ii) (Fig. 6.5) and Policy (iii) (Fig. 6.6) of the FST do not show such improvement for the results in note tracking. For policies (ii) and (iii), results obtained when onset information is not considered outperform all other cases. Clearly, the fact that Policy (i) is able to correct missed attack events by the MPE stage makes it stand as a better alternative for note tracking than the other policies considered. Moreover, this fact states the relevance of the note tracking stage: when providing onset information to the system, a proper strategy has to be followed to correctly incorporate that knowledge and take advantage of it. Thus, the use of more elaborated tracking processes which may take advantage of the particularities of piano notes should report an improvement.

Additionally, it can be checked that the MPE method by Vincent et al. (2010) consistently improves results with respect to Benetos, Cherla, and Weyde (2013): the figures obtained by the former method outperform the latter in around 5 % to 10 % in F_1, which suggests that the former method is more precise in terms of timing than the latter one. Finally, the improvement in the note tracking results of both MPE methods when onset information is considered states the robustness of onset-based tracking when compared to a basic pruning stage.

6.1.6 Discussion

This work studied the potential improvement that can be achieved when using onset information for post-processing an initial frame-level transcription obtained with a Multi-pitch Estimation system in the context of piano music. For performing such study, we compare the cases in which this onset description is either in the form of estimated onset events using state-of-the-art algorithms or in the form of ground-truth onset events as they represent the most accurate onset information. For all cases, the frame-level transcription is combined with the onset information using Finite State Transducers which, to our best knowledge, no author has previously considered.

The comparison of the results obtained when considering the estimated and ground-truth onset events points out an intrinsic relation between the accuracy of the onset information and the overall quality of the note tracking process. In general, this may be observed since improvements in the results of the onset estimation match the improvements in the note tracking figures.

Also it is shown that the performance of current existing state-of-the-art onset estimators limits the performance of onset-based note tracking systems. This may be checked as note tracking results obtained when considering ground-truth onset information generally outperform the ones achieved with estimated onset events for each particular note tracking policy.

These experiments also state the importance of the combination policy for onset and pitch information on the success of the task. The method in which the onset information is used for correcting the attack phase of the note is the one reporting the best overall results, possibly due to the fact that Multi-pitch Estimation methods tend to miss such attack stages. Finally, these experiments also point out the influence of the recording conditions of the piece as well as the relevance of the Multi-pitch Estimation algorithm on the performance of the note tracking stage.

6.2 Supervised Classification for Note Tracking

The second section of the chapter is devoted to the proposal and assessment of a novel method for note tracking in AMT based on supervised classification. As commented in Chapter 2, a great deal of note tracking approaches typically are based in hand-crafted policies, and thus the idea of this method is to somehow let the computer automatically infer those policies.

Figure 6.7 shows the general workflow for the AMT system, being the area labeled as *Note tracking* the one devoted to the proposed note tracking method. In this system, the audio signal to transcribe undergoes a series of concurrent processes: an MPE stage to retrieve the pitch-time posteriorgram $P(p,t)$ which is binarized and post-processed to obtain frame-level transcription $T_F\left(p,t\right)$ (binary representation depicting whether pitch p at time frame t is active), and an onset estimation stage that estimates a list of onset events $(o_i)_{i=1}^{L}$. These three pieces of information are provided to the note tracking method which post-processes the initial frame-level transcription $T_F\left(p,t\right)$ using the onset events to retrieve the note-level transcription $T_N(p,t)$. Note that this process is carried out in two different stages: an first one that considers the onset events $(o_i)_{i=1}^{L}$ for segmenting frame-level representation $T_F\left(p,t\right)$ into a set of instances and a second stage which classifies these instances as being active or inactive elements in the eventual note-level transcription $T_N(p,t)$.

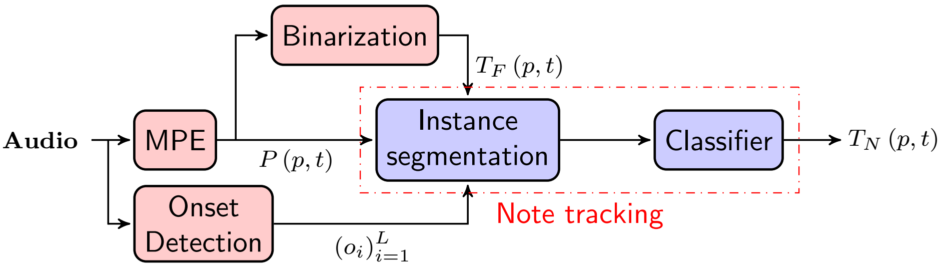


Figure 6.7: Proposed set-up for the assessment of the classification-based note tracking method proposed.

We shall now introduce the gist of the classification-based note tracking method proposed. Also, the rest of the section describes the experimental methodology considered for assessing the performance of the method.

6.2.1 Classification-based note tracking

This part of the work details the core idea of the note tracking approach proposed. It must be mentioned that the main contribution of this approach resides in how an initial frame-level transcription $T_F\,(p,t)$ is mapped into a set of instances to be classified, and not on the definition or proposal of a new (or, at least, specialized) supervised classification algorithm. Thus, this part is entirely devoted to the explanation of this segmentation process while the particular behaviour when considering different classification algorithms shall be later studied in the experimental assessment.

As introduced, the proposed note tracking strategy requires three sources of information: the pitch-time posteriorgram $P(p,t)$, where p and t correspond to the pitch and time indexes respectively, retrieved from a MPE analysis of an audio piece; a base frame-level transcription $T_F(p,t)$ obtained from the binarisation and basic post-processing of $P(p,t)$; and an L-length list $(o_i)_{i=1}^{L}$ of the onset events in the piece. Additionally, let $T_R(p,t)$ be the ground-truth piano-roll representation of the pitch-time activations of the piece, which is required for obtaining the labelled examples of the training set.

The initial binary frame-level transcription $T_F(p,t)$ can be considered a set of $|\mathcal{P}|$ binary sequences of $|t|$ symbols, where $|\mathcal{P}|$ and $|t|$ stand for the total number of pitches and frames in the sequence respectively. In that sense, we may use the elements $(o_i)_{i=1}^{L}$ as delimiters for segmenting each sequence or pitch band $p_j \in \mathcal{P}$ in $L+1$ subsequences. This process results in a frame-level abstraction quantised by the onset events that may be expressed as follows:

$$T_F(p_j,t) = T_F(p_j, 0 : o_1) \;||\; T_F(p_j, o_1 : o_2) \;||\; ... \;||\; T_F(p_j, o_L : |t|-1) \qquad (6.2)$$

where $||$ represents the concatenation operator.

Each of these onset-based $L+1$ subsequences per pitch are further segmented to create the instances for the classifier. The delimiters for these segments are the points in which there is a change in the state of the binary sequence, i.e. when there is a change from 0 to 1 (inactive to active) or from 1 to 0 (active to inactive). Mathematically, for the onset-based subsequence $T_F(p_j, o_i : o_{i+1})$ the $|C|$ state changes are obtained as:

$$C = \{t_m : T_F(p_j, t_m) \neq T_F(p_j, t_{m+1})\}_{t_m=o_i}^{o_{i+1}} \,. \qquad (6.3)$$

Thus, the resulting $|C|+1$ segments, which constitute the instances for the classifier, may be formally enunciated as:

$$T_F(p_j, o_i : o_{i+1}) = T_F(p_j, o_i : C_1) \;||\; ... \;||\; T_F(p_j, C_{|C|} : o_{i+1}) \,. \qquad (6.4)$$

Figure 6.8 illustrates graphically this procedure. In this example, for frame-level transcription $T_F(p,t)$, in the interval given by $[o_i, o_{i+1}]$ and band p_j,

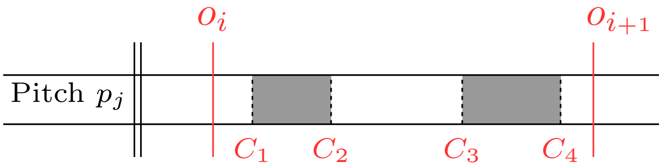


Figure 6.8: Segmentation of the onset-based subsequence $T_F(p_j, o_i : o_{i+1})$ into instances for the classifier. Grey (sequences of 1) and white areas (sequences of 0) depict active and inactive segments in the subsequence, respectively.

there are $|C| = 4$ state changes (i.e., changes from active to inactive of viceversa) and thus we obtain $|C| + 1 = 5$ subsequences.

So far we have performed the segmentation process based only on the information given by $T_F(p,t)$. Thus, at this point we are able to derive a set of instances that may serve as test set since they are not tagged according to the ground-truth piano roll $T_R(p,t)$. However, in order to produce a training set using the labels in $T_R(p,t)$, an additional step must be performed. For that we should *merge* the pieces of information from both $T_F(p,t)$ and $T_R(p,t)$ representations, which we perform by obtaining the C set of delimiters as:

$$C = C_{T_F} \cup \{t_m : T_R(p_j, t_m) \neq T_R(p_j, t_{m+1})\}_{t_m=o_i}^{o_{i+1}} \tag{6.5}$$

where C_{T_F} represents the segmentation points obtained from $T_F(p,t)$. This need for merging these pieces of information in shown in Fig. 6.9: if we only took into consideration the breakpoints in $T_F(p_j,t)$ (i.e., the band labeled as *Detected*), subsequence $T_F(p_j, t_a : t_b)$ would have two labels if checking the figure labeled as *Annotation* – subsequence $T_F(p_j, t_a : t_c)$ should be labeled as non-active and $T_F(p_j, t_c : t_b)$ as active. Thus, we require this additional breakpoints to further segment the subsequences and align them with the ground-truth labels to produce the training set. Again, note that this process is not required for the test set since evaluation is eventually done in terms of note tracking and not as classification accuracy.

Once the process for segmenting into instances has been performed, a set of features is extracted for each of the instances: *i)* descriptors related to the temporal description of the instance, as its duration, its distance to the previous and posterior onsets, and its duration with respect to the inter-onset interval; and *ii)* features related to the posteriorgram $P(p,t)$ as the average energy in the current and adjacent octave-related bands.

To avoid that the considered features may span for different ranges, we opted to normalize them: energy descriptors (E, E_l, and E_h) are already constrained to the range $[0, 1]$ as the input posteriorgram is normalised to its global maximum (cf. Section 6.2.2 in which the experimentation is described); occupation ratio $\mathcal{D}$ is also inherently normalized as it already represents a

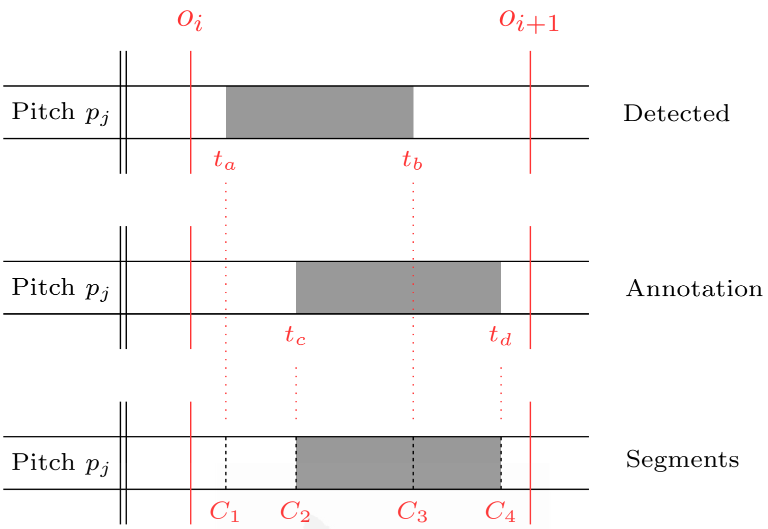


Figure 6.9: Segmentation and labelling process for the training corpus. Breakpoints t_a and t_b from frame-level transcription $T_F(p_j, t)$ – labelled as *Detected* – together with breakpoints t_c and t_d from ground-truth piano roll $T_R(p_j, t)$ – labelled as *Annotation* – are considered for segmenting sequence $p_j \in \mathcal{P}$. Labels are retrieved directly from $T_R(p, t)$. For each case, grey and white areas depict sequences of 1 and 0, respectively.

Table 6.3: Summary of the features considered for the classification-based note tracking approach proposed. Operator $\langle \cdot \rangle$ retrieves the average value of the elements considered.

Feature	**Definition**	**Description**
Δt	$C_{m+1} - C_m$	Duration of the block
Δo_i	$C_m - o_i$	Distance between previous onset and the starting point of the block
Δo_{i+1}	$o_{i+1} - C_{m+1}$	Distance between end of the block and the posterior onset
$\mathcal{D}$	$\frac{\Delta t}{o_{i+1} - o_i}$	Occupation ratio of the block in the inter-onset interval
E	$\langle P(p_j, C_m : C_{m+1}) \rangle$	Mean energy of the multipitch estimation in current band
E_l	$\langle P(p_j - 12, C_m : C_{m+1}) \rangle$	Mean energy of the multipitch estimation in previous octave
E_h	$\langle P(p_j + 12, C_m : C_{m+1}) \rangle$	Mean energy of the multipitch estimation in next octave

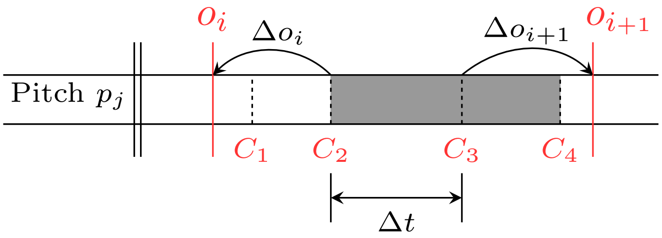


Figure 6.10: Graphical representation of the set of features considered. In this case, the instance being characterized is $T_F(p_j, C_2 : C_3)$.

ratio between two magnitudes; absolute duration Δt and distance features Δo_i and Δo_{i+1} are manually normalised using the total duration of the sequence $|t|$ as a reference.

Finally, in an attempt to incorporate *temporal knowledge* in the classifier, we include as additional features the descriptors of the instances surrounding the one at issue (previous and/or posterior ones). To exemplify this let us take the case in Fig. 6.10. Also consider a temporal context to include a temporal context that of one previous and one posterior windows to the instance to be defined. To do so, and for the precise case of instance $T_F(p_j, C_2 : C_3)$, we should take into account the features of both instances $T_F(p_j, C_1 : C_2)$ and $T_F(p_i, C_3 : C_4)$.

6.2.2 Experimentation

This part of the work introduces the experimentation carried out to assess the performance of our note tracking proposal and its comparison with other existing methods. For that, we initially introduce the corpora and the figures of merit considered; then we present the MPE strategy used for obtaining the posteriorgram $P(p, t)$ and its post-processing to obtain the frame-level transcription $T_F(p, t)$; after that we introduce the different onset estimation strategies assessed in the work; and finally we list and explained the different supervised classification strategies considered as well as other alternative note tracking strategies for the comparison of the results obtained.

Evaluation methodology

In terms of data, we employ the MAPS database (Emiya et al., 2010) for assessing the proposed approach. This collection comprises several sets of audio piano performances of isolated sound, chords, and complete music pieces from both real and synthesised instruments and synchronised with MIDI annotations.

For comparative purposes we reproduced the evaluation configuration of the MAPS dataset used in Sigtia et al. (2016). In that work the evaluation was restricted to the use of the subset of complete music pieces. Such subset comprises 270 music pieces, out of which 60 are directly recorded with two different Yamaha Disklavier units and the rest are synthesized via software emulating different types of piano sounds. Within their evaluation, the data was organized considering a 4-fold cross validation, being 216 out of the 270 music pieces for used for training and 54 music pieces for test. The precise description of the sets may be found in `http://www.eecs.qmul.ac.uk/~sss31/TASLP/info.html`. Additionally, only the first 30 seconds of each of the pieces were considered for the experimentation as done in other AMT works, which gives up to a corpus with a total number of 72,585 note events. Table 6.4 summarizes the number of note events per train/test fold.

Table 6.4: Summary in terms of note events for each train/test fold considered for the evaluation of the classification-based note tracking approach reproduced from Sigtia et al. (2016).

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**
Train	59,563	59,956	54,589	60,527
Test	13,022	12,629	17,996	12,058

Regarding the evaluation, as in Section 6.1, we shall evaluate the performances of both the onset estimation and the note tracking tasks. For that, we again rely on the evaluation methodologies introduced in Chapter 2: on the one hand, we consider the standard onset assessment evaluation with a tolerance window of 50 ms; on the other hand, in terms of note tracking we shall restrict ourselves to the onset-based figure of merit as we are not considering note offsets, also with a tolerance window of 50 ms.

Multipitch estimation

For the initial multipitch analysis of the audio music pieces we considered the system by Benetos and Weyde (2015) that belongs to the Probabilistic Latent Component Analysis (PLCA) family of methods, which ranked first in the 2015 evaluations of the MIREX Multiple-F0 Estimation and Note Tracking Task[2]. This particular system takes as input representation a variable-Q transform (VQT) and decomposes into a series of pre-extracted log-spectral templates per pitch, instrument source, and tuning deviation from ideal tuning. Outputs of the model include a pitch activation probability $P(p, t)$ (p stands for pitch in MIDI scale), as well as distributions for instrument contributions per pitch and a tuning distribution per pitch over time. The unknown model parameters are iteratively estimated using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977), using 30 iterations

[2] `http://www.music-ir.org/mirex/wiki/MIREX_HOME`

in this implementation. For this particular study we consider a temporal resolution of 10 ms for the input time-frequency representation and output pitch activation and $|\mathcal{P}| = 88$ pitch values.

The retrieved pitch-time posteriorgram $P(p,t)$ is then processed to obtain a frame-level transcription $T_F(p,t)$ with the same postprocessing stage as in the experiments done in Section 6.1: first of all, $P(p,t)$ is normalized to its global maximum so that $P(p,t) \in [0,1]$; then, for each pitch value $p_i \in p$, a median filter of 70 ms of duration is applied over time to smooth the detection; after that, the resulting posteriorgram is binarised using a global threshold value of $\theta = 0.1$ which is obtained taking the work in Vincent et al. (2010) as a reference and refining it for the data used in this work; finally, a minimum-length pruning filter of 50 ms is applied to remove spurious detected notes.

Onset information

Regarding the onset description of the signal, and as done in Section 6.1, we distinguish two different situations: a first one in which we considered ground-truth onset events and a second one in which we automatically estimate onset information. By studying these two situations we can additionally assess the potential improvement that may be achieved with the proposed note tracking approach when considering the most accurate onset information that may be provided (the ground-truth one) and compare it to the improvement achieved with the estimated events.

As of onset estimation algorithms we selected four representative methods found in the literature: a simple Spectral Difference (SD), the Semitone Filter-Bank (SFB) method by Pertusa et al. (2005), the SuperFlux (SF) algorithm by Böck and Widmer (2013a, 2013b), and Complex Domain Deviation (CDD) by Duxbury et al. (2003). All these processes retrieve a list $(o_i)_{i=1}^{L}$ whose elements represent the time positions of the L onsets detected in the signal. Reader is referred to Section 5.3 for the explanation of these methods.

The analysis parameters of the alternatives considered are set to their default values[3]. Additionally, as all of them comprise a final thresholding stage, we test 25 different values equally spaced in the range $(0,1)$ to check the influence of that parameter. From this analysis selected the value that optimizes the estimation for then using it in the note tracking stage. Finally, the onset lists $(o_i)_{i=1}^{L}$ are processed with an averaging 30 ms filter to avoid overestimation issues by the algorithms as commented in Böck et al. (2012).

[3] SFB considers windows of 92.8 ms with a temporal resolution of 46.4 ms; SF considers smaller windows of 46.4 ms with a higher temporal granularity of 5.8 ms; SD and CDD both consider windows of 11.6 ms with also a temporal resolution of 5 ms.

Comparative approaches

Given that the proposed method models the note tracking problem as a classification task, we aim to study the behaviour and performance of several supervised classification algorithms in this particular context. While the considered classification strategies are now introduced, the reader is referred to works by Bishop (2006) and Duda et al. (2001) for a thorough description of the methods:

1. **Nearest Neighbour (NN)**: Non-parametric classifier based on dissimilarity. Given a labelled set of samples $\mathcal{T} = \{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathcal{X},\ y_i \in \mathcal{Y}\}_{i=1}^{|\mathcal{T}|}$, the NN rule assigns to a query x' the class of sample $x \in \mathcal{T}$ that minimizes a dissimilarity measure $d(x, x')$. Generalising, if considering k neighbours for the classification (kNN rule), x' is assigned the mode of the individual labels of the k nearest neighbours. For our experiments we restrict to the use of one single nearest neighbour (i.e., 1NN) with Euclidean distance as dissimilarity measure.

2. **Decision Tree (DT)**: Non-parametric classifier that performs the separation of the classes by iteratively partitioning the search space with simple decisions over the features in an individual fashion. The resulting model may be represented as a tree in which the nodes represent the individual decisions to be evaluated and the nodes contain the classes to assign. In this case we consider the Gini impurity as the measure to perform the splits in the tree and that a leaf must become a node when it contains more than one sample.

3. **AdaBoost (AB)**: Ensemble-based classifier that is based on the linear combination of weak classification schemes. Each weak classifier is trained on different versions of the training set $\mathcal{T}$ that basically differ on the weights (classification relevance or importance) given to the individual instances. In this case, the weak classifiers are based on decision trees as the ones from the previous point.

4. **Random Forest (RaF)**: Ensemble-based scheme that categorizes query x' considering the decisions of one-level decision trees (decision stumps) trained over the same training set $\mathcal{T}$. The class predicted by the ensemble is the mode of the individual decisions by the stumps. For our experiments, the number of decision stumps has been fixed to 10.

5. **Support Vector Machine (SVM)**: Non-parametric binary classifier that seeks for a hyperplane that maximizes the margin between the hyperplane itself and the nearest samples of each class (support vectors) of training set T. For non-linearly separable problems, this classifier

relies on the use of Kernel functions (i.e., mapping the data to higher-dimensional spaces) to improve the separability of the classes. In this work the radial basis function (rbf) Kernel has been considered for performing such mapping.

6. **Multilayer Perceptron (MLP)**: Particular topology of an artificial neural network parametric classifier. This topology implements a feed-forward network in which each neuron in a given layer is fully-connected to all neurons of the following layer. The configuration in this case is a single-layer network comprising 100 neurons with rectified linear unit (ReLU) activations and a softmax layer for the eventual prediction.

Note that the interest of the work lies in the exploration of the classification-based proposal rather than in its optimization. In that sense, the algorithms considered are directly taken from the Scikit-learn Machine Learning library (Pedregosa et al., 2011).

For comparative purposes we also considered the use of pitch-wise two-state Hidden Markov Models (HMMs) in a similar way to the work by Poliner and Ellis (2007). HMMs constitute a particular example of statistical model in which it is assumed that the system at issue can be described as a Markov process (i.e., a model for which the value of a given state is directly influenced by the previous one) with a set of unobservable states. In this work we replicate the scheme proposed in the aforementioned reference: we define a set of 88 HMMs (one per pitch band considered) with two hidden states, active or inactive step; each HMM is trained by simply counting the type of transition between consecutive analysis frames (i.e., all combinations of transitioning from an active/inactive frame to an active/inactive one) of the elements of the training set; decoding is then performed on the test set using the Viterbi algorithm (Viterbi, 1967).

Finally, we also compare the proposed method with the results obtained by Sigtia et al. (2016) as we have both replicated their experimental configuration and considered the same PLCA-based MPE method that this work. This consideration is mainly motivated by the fact that the aforementioned work constitutes a very recent method that tackles note-level transcription by implementing a polyphonic Music Language Model (MLM) based on a hybrid architecture of Recurrent Neural Networks (a particular case of neural networks that model time dependencies) and a Neural Autogressive Distribution Estimation (a distribution estimator for high dimensional binary data).

6.2.3 Results

This section presents the results obtained with the proposed experimental scheme for both the onset detection and note tracking methods. The figures shown in the section depict the average value of the considered figure of merit obtained in each of the cross-validation folds.

First of all, we study the performance of the different onset detection methods considered. The aim is to assess the behaviour of these algorithms on the data considered to later compare the performance of the note tracking method proposed when considering different onset descriptions of the signal. The difference in performance of the onset detectors will, in principle, imply a difference in the performance of the note tracking method, which shall give insights about the robustness of the strategy proposed. For the assessment of the onset detectors we only consider the elements of the training set (test partition is not accessible) and we assume that the conclusions derived from this study shall be applicable to the test set as they represent the same data distribution. In these terms, Fig. 6.11 graphically shows the average F_1 of the folds considered by the different onset estimation algorithms used as the selection threshold varies.

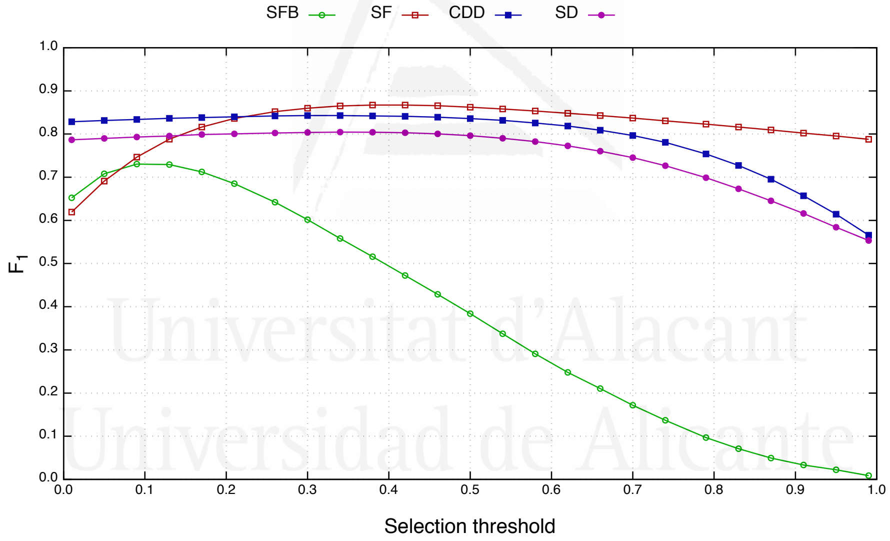


Figure 6.11: Onset detection results in terms of F_1 when varying the selection threshold. Acronyms in the legend stand for each onset estimation method: SFB for Semitone Filter-Bank, SF for SuperFlux, CDD for Complex Domain Deviation, and SD for Spectral Difference.

An initial remark to point out is the clear influence of the threshold parameter of the selection stage in the performance of the onset estimation methods. In these terms, SFB arises as the one whose performance is more affected by this selection stage, retrieving performance values that span from a completely erroneous estimation of $F_1 \approx 0$ to fairly accurate results of $F_1 \approx 0.75$. Attending to its performance, we select a threshold of $\theta = 0.13$ that accomplishes an approximate value of $F_1 = 0.75$ in the detection task.

SD and CDD depict a totally opposite behaviour to the SFB method: these algorithms show a relatively steady performance for the threshold values studied with goodness figures of $F_1 \approx 0.8$ that only decrease to a performance of $F_1 \approx 0.5$ when the selected threshold approaches the unit. It can be seen that the CDD method shows a slightly better performance than the SD one, possibly due to the use of phase information for the estimation. For these two methods we find the local maxima when selecting threshold values of $\theta = 0.34$ of the SD methods and $\theta = 0.30$ for the CDD one, retrieving performances of $F_1 \approx 0.80$ and $F_1 \approx 0.82$ for the SD and CDD algorithms, respectively.

Finally, the SF method also presents a very steady performance for all threshold values studied with the particular difference that the performance of the onset estimation degrades as the threshold value considered for the selection is reduced. Also, it must be pointed out that this algorithm shows the best performance among all studied methods when the selection stage is properly configured. In this case we select $\theta = 0.38$ as the threshold value that maximizes the performance of the algorithm.

Having analysed the performance of the considered onset selection methods, we now assess the performance of the proposed note tracking approach. Table 6.5 shows the results obtained with the proposed note tracking method for the different classification strategies and numbers of adjacent instances for both the *frame-based* and *note-based* assessments. Note that the different onset detection methods use thresholds that optimize their respective performance. These onset estimators are denoted with the same acronyms as above while the particular case when considering ground-truth onset information is denoted as GT.

On a broad analysis of the results obtained, a first point to highlight is that the proposed note tracking strategy achieves its best performance when considering ground-truth onset information (i.e., the one labelled as GT). While this may be seen as the expected behaviour, such results prove the validity of the note tracking method proposed: with the proper configuration (in this case, the most precise onset information that could be achieved for the data) this strategy is capable of retrieving performance values of $F_1 = 0.70$ in the frame-based analysis and $F_1 = 0.73$ in the note-based one. Note that such figures somehow constitute the maximum achievable performance of the proposed note tracking method given that actual onset estimators are not capable of retrieving such accurate onset description of a piece. Nevertheless, these values might be improved by considering the use of other descriptors different to the ones studied, obtained as either hand-crafted descriptors or with the use of *feature learning* approaches to automatically infer the most suitable features for the task.

When considering estimated onset events instead of ground-truth information there is a decrease in the performance of the note tracking system. In general, and as somehow expected, this drop is correlated with the goodness

Table 6.5: Note tracking results for the proposed classification-based note tracking method across several classifiers, using frame-based and note-based metrics. Each figure depicts the average F_1 obtained of the 4-fold cross validation scheme considered. Notation (x, y) stands for the number of previous and posterior additional instances considered. Bold figures remark the best performing configuration per onset estimator and number of surrounding windows considered.

		GT		**SD**		**SFB**		**SF**		**CDD**	
		Frame	Note	Frame	Note	Frame	Note	Frame	Note	Frame	Note
$(0, 0)$	NN	0.64	0.63	0.61	0.56	0.60	0.57	0.64	0.62	0.62	0.56
	DT	0.60	0.56	0.58	0.50	0.57	0.51	0.60	0.55	0.59	0.51
	RaF	0.66	0.65	0.63	0.57	0.62	0.58	0.66	0.64	0.63	0.58
	AB	0.56	0.60	0.53	0.52	0.51	0.54	0.56	0.59	0.54	0.53
	SVM	0.60	0.67	0.58	**0.61**	0.57	**0.62**	0.60	0.66	0.57	**0.62**
	MLP	**0.67**	**0.69**	**0.65**	0.60	**0.63**	0.61	**0.67**	**0.68**	**0.66**	0.61
$(1, 1)$	NN	0.65	0.69	0.61	0.56	0.60	0.59	0.63	0.62	0.61	0.57
	DT	0.62	0.59	0.60	0.50	0.59	0.52	0.62	0.54	0.60	0.50
	RaF	0.68	0.70	0.64	0.58	0.63	0.60	0.66	0.64	0.64	0.59
	AB	0.57	0.61	0.55	0.56	0.52	0.56	0.56	0.59	0.55	0.56
	SVM	0.58	0.69	0.56	0.58	0.54	**0.64**	0.57	0.64	0.56	0.58
	MLP	**0.70**	**0.72**	**0.66**	**0.60**	**0.65**	0.62	**0.68**	**0.66**	**0.66**	**0.61**
$(2, 2)$	NN	0.65	0.70	0.60	0.57	0.59	0.58	0.63	0.63	0.61	0.57
	DT	0.62	0.59	0.59	0.49	0.59	0.51	0.61	0.53	0.60	0.50
	RaF	0.68	0.70	0.63	0.58	0.63	0.59	0.66	0.64	0.64	0.58
	AB	0.59	0.63	0.55	0.55	0.53	0.57	0.57	0.59	**0.66**	0.56
	SVM	0.57	0.70	0.60	**0.62**	0.54	**0.64**	0.55	0.63	0.56	0.59
	MLP	**0.69**	**0.73**	**0.66**	0.61	**0.64**	0.61	**0.69**	**0.66**	**0.66**	**0.60**

of the onset estimation. As a first example, SF achieves the best results among all the onset estimators: its performance is, in general, quite similar to the case when ground-truth onset information is considered and only exhibits particular drops that, in the worst-case scenario, get to a value of 3 % and 10 % for the frame-based and note-based metrics, respectively, lower than the maximum achievable performance. The SD and CDD estimators exhibit a very similar performance between them, being the latter one the algorithm that occasionally overpasses the former one; both estimators show a decrease between 3 % and 6 % for the frame-based metric and between 10 % and 20 % for the note-based figure of merit when compared to the ground-truth case. As of the SFB algorithm, while reported as the one achieving the lowest performance in terms of onset accuracy, it reports very accurate note tracking figures that practically do not differ to the ones achieved by the SD and CDD algorithms.

Regarding the classification schemes, it may be noted that the eventual performance of the system is remarkably dependent on the classifier considered. Attending to figures obtained, the best results are obtained when considering an MLP as classifier, and occasionally an SVM scheme. For instance, in the ground-truth onset information case, MLP reports performance figures of $\mathrm{F}_1 = 0.70$ for the frame-based evaluation and a $\mathrm{F}_1 = 0.73$ in the note-based one, thus outperforming all other classification strategies considered that also employ the same onset information. As accuracy in the onset information degrades, the absolute performance values suffer a drop (for instance, $\mathrm{F}_1 = 0.66$ in the note-based evaluation for the SF estimator or $\mathrm{F}_1 = 0.60$ for the same metric and the CDD estimator), but MLP still obtains the best results. As commented the only strategy outperforming MLP is SVM for the particular cases when onset information is estimated with the SD and SFB methods and assessing with the onset-based metric. Nevertheless, experiments reported that convergence in the training stage for the SVM classifier turned out to be much slower than for the MLP one, thus exhibiting the latter one this additional characteristic.

On the other extreme, AB and DT generally report the lowest performance figures for the frame-based and note-based assessment strategies, respectively. For instance, in the ground-truth onset information case, AB reports a decrease in the frame-based metric close to 16 % with respect to the maximum reported by the MLP. Similarly, when compared to the maximum, DT reports a decrease close to a 20 % in the note-based assessment.

The NN classifier exhibits a particular behaviour to analyse. As it can be checked, this scheme retrieves fairly accurate results for both the frame-based and note-based metrics for the ground-truth onset information (on a broad picture, close to $\mathrm{F}_1 = 0.65$). Nevertheless, when other source of onset description is considered, the note-based metric remarkably degrades while the frame-based metric keeps relatively steady. As the NN rule does not perform any explicit generalisation over the training data, it may be possible that instances with similar feature values may be labelled with different classes and thus confuse the performance of the system.

The RaF ensemble-based scheme, while not reporting the best overall results, achieves scores that span up to values of $F_1 = 0.68$ and $F_1 = 0.70$ for the frame-based and note-based metric, respectively, with ground-truth onset information. While it might be argued that these figures may be improved by considering more complex base classifiers, ensemble methods have been reported to achieve their best performance using simple decision schemes, such as the one-level decision trees used in this work. Besides, given the simplicity of the base classifiers, the convergence of the training model in RaF is remarkably fast, thus exhibiting an additional advantage to other classification schemes with slower training phases.

According to the obtained results, the use of additional features which consider the surrounding instances leads to different conclusions depending on the evaluation scheme considered. Except for the case when considering ground-truth onset information in which such information shows a general improvement in the performance of the system, no clear conclusions can be gathered when considering these additional features for the rest of the cases. For instance, consider the case of the SVM classifier with the SD estimator; in this case, note-based performance decreases from $F_1 = 0.61$ when no additional features are considered to $F_1 = 0.58$ when only the instances directly surrounding the one at issue are considered; however, when the information of two instances per side is included, the performance increases to $F_1 = 0.62$.

With respect to the comparison with existing note tracking methods from the literature, Table 6.6 shows results in terms of F_1 comparing the following approaches: *Base*, which stands for the initial binarisation of the posteriorgram, Poliner and Ellis (2007), using a two-stage HMM for note tracking, and Sigtia et al. (2016) which considers a Music Language Model (MLM) based post-processing scheme. Finally *Classification* shows the best figures obtained with the proposed method for the different onset estimators. These methods are denoted by the same acronyms used previously in the analysis while ground-truth onset information is referred to as GT.

As can be seen from Table 6.6, the proposed classification-based method stands as a competitive alternative to other considered techniques. For both frame-based and onset-based metrics, the proposed method is able to surpass the baseline approach by more than +10 % in terms of F_1 for both metrics considered.

When compared to the HMM-based method by Poliner and Ellis (2007), the proposed approach also demonstrates an improvement of +10 % when considering frame-based metrics and +3 % in terms of note-based metrics, when using the SF onset detector. As expected, the improvement increases further when using ground truth onset information with the proposed method.

The method by Sigtia et al. (2016) achieves similar figures to the HMM-

Table 6.6: Note tracking results on the MAPS dataset in terms of F_1, comparing the proposed classification-based method with the considered benchmark approaches. *Base* stands for the initial binary frame-level transcription obtained; Poliner and Ellis (2007) refers to the HMM-based note tracking method proposed on that paper; Sigtia et al. (2016) represents the MLM-based post-processing technique; *Classification* stands for the proposed method with the different onset detection methods considered.

	Base	Poliner and Ellis (2007)	Sigtia et al. (2016)	Classification				
				GT	SD	SFB	SF	CDD
Frame	0.57	0.59	0.65	0.70	0.66	0.65	0.69	0.66
Note	0.62	0.65	0.66	0.73	0.62	0.64	0.68	0.62

based approach with a particular improvement on the frame-based metric. In this sense, conclusions gathered from the comparison are quite similar: the proposed approach shows an improvement using frame-based metrics while for the note-based ones it is necessary to consider very precise onset information (e.g. the SF method or the ground-truth onset annotations).

Finally, the existing gap between the figures obtained when considering ground-truth onset information and the SF onset detector suggests that there is still room for improvement simply by focusing on improving the performance of onset detection methods.

6.2.4 Discussion

In this work we explored the use of a data-driven approach for note tracking by modelling the task as a supervised classification problem. The proposed method acts as a post-processing stage for an initial frame-level multi-pitch detection: each pitch band of the initial frame-level transcription is segmented into instances using onset events estimated from the piece and a set of features based on the multi-pitch analysis; each instance is classified as being an active or inactive element of the transcription (binary classification) by comparing to a set of labelled instances.

The results obtained when assessing the note tracking method proposed on a collection of piano music provide several insights. A first one is that, checking the results obtained, we can confirm that this alternative proposal is, at least, as competitive as other existing approaches: the proposed classification-based method generally outperforms other note tracking strategies typically considered as hand-crafted rules or the *de-facto* standard approach of a pitch-wise two-state Hidden Markov Model.

When the proposed method is compared to more recent strategies as the Music Language Model by Sigtia et al. (2016), this improvement is not observed, at least in general terms. Nevertheless, the performance of these two methods they may be considered totally equipable as differences in performance are not remarkable. This fact constitutes a point to highligh as our proposal exhibits lower computational requirements than the other proposal.

Finally, somehow confirming the conclusions presented in Section 6.1, there is a clear relation between the goodness of the onset estimation and overall performace of the tracking system. Note that, as in the aforementioned work the best results are obtained when ground-truth onset information (the *most accurate* onset information that would be expectable) is considered for the system, being the state-of-the-art SuperFlux method the one that achieves the closest to this scheme.

6.3 General discussion

Note onset information is undoubtly useful in the context of Automatic Music Transcription as it provides an accurate temporal description of the piece to transcribe in terms of the starting points of the note events. Such information is generally *merged* with other descriptions of the signal (e.g., pitch analysis) to define and shape discrete note events out of its raw frequential analysis. In this context, we presented two works which dealt with the use of onset information for post-processing an initial frame-level transcription for correcting the errors committed in the Multi-pitch Estimation stage.

The first of these works focused on the study of the relevance of onset information in such note tracking systems. More precisely, the idea was to assess the actual relation between the quality in the estimation of onset events and the accuracy of the resulting note-level transcription. The comprehensive experimentation proved the intrinsic relation between these two pieces of information, showing that more accurate onset detectors obtained better results in terms of the eventual note-level description. Also these experiments pointed out the relevance of the note tracking policy in the overall success of the task, showing that policies addressing the correction of note attack phases improve the results as they constitute typical errors committed by Multi-pitch Estimation methods. Finally, this work showed as well the limitations of current state-of-the-art onset estimation methods and its implications on the potential improvement for note-level transcription. These limitations somehow support the pieces of research presented previously in terms of interactive schemes for onset detection.

The second work presented a novel method for note tracking based on supervised classification. A great deal of existing note tracking methods consist in collections of hand-crafted rules adapted to the precise type of data at issue. In this case though we proposed an alternative system that somehow *infers* these note tracking rules modeling the note-level transcription stage as a binary classification task. Results obtained show that the proposed method is competitive against other existing strategies such as the aforementioned set of hand-crafted rules of the *de-facto* standard post-processing strategy by Poliner and Ellis (2007) based on Hidden Markov Models

Chapter 7

Conclusions and future perspectives

"We can only see a short distance ahead, but we can see plenty there that needs to be done."

Alan M. Turing

This dissertation addressed the topic of music transcription from audio considering two different, yet complementary, perspectives: *i)* the *interactive* aspect, which considers the inclusion of the user as an active element of the transcription rather than a simple verification agent; and *ii)* the *multimodal* perspective, which postulates the need for using different descriptions of the music signal to achieve a precise and accurate high-level symbolic transcription.

The consideration of interactive schemes may be seen as a deviation from the *ideal* stand-alone Automatic Music Transcription paradigm. Nevertheless, this dissertation takes as starting point the claims by a number of researchers in the Music Information Retrieval community of having reached a *glass-ceiling* in music transcription methodologies, thus being necessary a change of paradigm. Note that, as no transcription system is error-free, a human agent is generally required to revise and correct the estimation by the system. Thus, assuming this need for human supervision in the transcription process, there is a clear need for developing interactive strategies for efficiently exploiting this human effort.

As of the multimodal perspective, in this dissertation we studied the use of onset events as an additional source of information for the transcription process. While onset information has already been explored for transcription tasks, in this work we have further studied its potential in the particular context of note tracking systems. Moreover, given the relevance of this source of information for the proper consecution of the transcription task, all the studies related to the aforementioned interactive paradigm have considered the issue of interactive annotation and correction of onset events in audio streams.

In addition, this dissertation made extensive use of Pattern Recognition techniques for addressing the two aforementioned perspectives. Thus, as additional contributions of this work not related to the Music Information Retrieval field but to the general Pattern Recognition one, we have studied different strategies for coping with class-imbalance and large-size data collections in the context of the instance-based k-Nearest Neighbor classifier.

7.1 Conclusions and contributions

The main contributions and conclusions gathered from the development of this dissertation are summarized in the following points:

1. A thorough revision of the state of the art in the Automatic Music Transcription field as well as for more general applications of Pattern Recognition in Music Information Retrieval.

2. The assessment of a set of novel Prototype Selection methods for the k-Nearest Neighbor classifier based on ranking principles, namely Nearest to Enemy and Farthest Neighbor strategies. The experiments performed prove the competitiveness of these rank-based methods in terms of both set size reduction and noise elimination capabilities as well as their low computational complexity compared to more sophisticated methods.

3. A comparative study on the use of Prototype Selection methods for the k-Nearest Neighbor classifier in the particular case of imbalanced classification problems which require of an instance selection process (e.g., large size datasets). The experiments carried out show that, in general, it is beneficial in terms of the classification performance to apply data-level balancing techniques, and more precisely combinations of *Oversampling* and *Undersampling* methods, before the classification stage given that the latter methods are generally prepared for class-balanced data collections.

4. The proposal of a set of figures of merit for quantitatively assessing the human effort invested in interactive onset annotation and correction processes in audio music pieces.

5. A collection of interactive onset annotation schemes based on signal processing techniques. This set of schemes is based on the general two-stage onset estimation approach, that is, an initial detection function

process followed by an onset selection stage; the corrections performed by the user modify the parameters of the latter stage in order to adequate this selection curve to the particular audio piece being annotated. Experimental results prove that these interactive methods constitute effective yet simple approaches for remarkably reducing the user workload invested in the correction and annotation of onset events in audio streams.

6. The proposal and study of the interactive onset annotation and correction issue from an Interactive Pattern Recognition point of view. Initially, the onset estimation problem is modeled as a Pattern Recognition task in which each analysis frame of the signal is classified as either containing or not an onset event. In such context, the interactive model updating is achieved by dynamically modifying the training set of the classifier according to the user corrections. The main research question in this framework is the issue of which information is relevant for the model to improve its performance. The conclusions gathered from a thorough experimentation point out that as a larger amount of information is provided to the feedback loop, the model improves its robustness as well as noticeably decreases the user workload required to perform the annotation process.

7. A formal study of the potential improvement that the use of onset information in music transcription systems supposes. Note that the use of onset information in transcription systems does not constitute a novelty by itself as a number of authors have already considered such schemes. In this dissertation the novel contribution resides in the formal evaluation of the potential improvement that may be achieved with the use of onset-based transcription systems when compared to systems which ignore it. Also, an additional contribution in this context is the evaluation and study of the relation between the quality of the onset estimation stage (i.e., performance of the onset estimator) with the overall performance of the transcription system.

8. The introduction of a novel note tracking approach for Automatic Music Transcription designed from a Pattern Recognition perspective. In general, note-level transcriptions are obtained using a set of handcrafted rules that post-process the initial frame-level transcription. In this work a classification model was considered to, at some extent, let the system automatically infer these note-tracking rules rather than manually defining them. Results obtained showed that, while the proposed approach equals the performance of other existing note tracking strategies without outperforming them, this scheme constitutes a change of paradigm in this task. In this sense, the novelty resides in the proposal and exploration of this alternative classification-based

strategy rather than in its actual competitive performance.

7.1.1 Publications

Part of the contents of this dissertation have been published in several journals and conference events. These publications are now listed showing, in brackets, the chapter to which they are related:

- Valero-Mas, J. J., Iñesta, J. M., & Pérez-Sancho, C. (2014, November). Onset detection with the user in the learning loop. In *Proceedings of the 7th International Workshop on Machine Learning and Music (MML)*. Barcelona, Spain. [**Chapter 5**]

- Valero-Mas, J. J., & Iñesta, J. M. (2015, October). Interactive onset detection in audio recordings. In *Late Breaking/Demo extended abstract, 16th International Society for Music Information Retrieval Conference (ISMIR)*. Málaga, Spain. [**Chapter 5**]

- Valero-Mas, J. J., Calvo-Zaragoza, J., Rico-Juan, J. R., & Iñesta, J. M. (2016). An experimental study on rank methods for prototype selection. *Soft Computing*, 1–13. [**Chapter 4**]

- Valero-Mas, J. J., Benetos, E., & Iñesta, J. M. (2016, September). Classification-based Note Tracking for Automatic Music Transcription. In *Proceedings of the 9th International Workshop on Machine Learning and Music (MML)* (pp. 61–65). Riva del Garda, Italy. [**Chapter 6**]

- Valero-Mas, J. J., Calvo-Zaragoza, J., Rico-Juan, J. R., & Iñesta, J. M. (2017, June). A study of prototype selection algorithms for nearest neighbour in class-imbalanced problems. In *Proceedings of the 8th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*. Faro, Portugal. [**Chapter 4**]

- Valero-Mas, J. J., Benetos, E., & Iñesta, J. M. (2017, June). Assessing the Relevance of Onset Information for Note Tracking in Piano Music Transcription. In *Proceedings of the Audio Engineering Society (AES) International Conference on Semantic Audio*. Erlangen, Germany. [**Chapter 6**]

- Valero-Mas, J. J., & Iñesta, J. M. (2017, July). Experimental assessment of descriptive statistics and adaptive methodologies for threshold establishment in onset selection functions. In *Proceedings of the 14th Sound and Music Computing Conference (SMC)*. Espoo, Finland. [**Chapter 5**]

Two additional contributions are currently under review process:

- Valero-Mas, J. J., & Iñesta, J. M. Interactive User Correction of Automatically Detected Onsets: Approach and Evaluation. *EURASIP Journal on Audio, Speech, and Music Processing.* [**Chapter 5**]

- Valero-Mas, J. J., Benetos, E., & Iñesta, J. M. A Supervised Classification Approach for Note Tracking in Polyphonic Piano Transcription. *Journal of New Music Research.* [**Chapter 6**]

In addition to the previous works, the author has also collaborated in a series of works mostly related to the field of Pattern Recognition:

- Valero-Mas, J. J., Salamon, J., & Gómez, E. (2015, July). Analyzing the influence of pitch quantization and note segmentation on singing voice alignment in the context of audio-based Query-by-Humming. In *Proceedings of the 12th Sound and Music Computing Conference (SMC)* (pp. 371–378). Maynooth, Ireland.

- Calvo-Zaragoza, J., Valero-Mas, J. J., & Rico-Juan, J. R. (2015, June). Prototype Generation on Structural Data using Dissimilarity Space Representation: A Case of Study. In *R. Paredes, J. S. Cardoso, & X. M. Pardo (Eds.), Proceedings of the 7th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)* (pp. 72–82). Santiago de Compostela, Spain: Springer.

- Calvo-Zaragoza, J., Valero-Mas, J. J., & Rico-Juan, J. R. (2015). Improving kNN multi-label classification in Prototype Selection scenarios using class proposals. *Pattern Recognition*, 48(5), 1608–1622.

- Valero-Mas, J. J., Calvo-Zaragoza, J., & Rico-Juan, J. R. (2016). On the suitability of Prototype Selection methods for kNN classification with distributed data. *Neurocomputing*, 203, 150–160.

- Calvo-Zaragoza, J., Valero-Mas, J. J., & Rico-Juan, J. R. (2016). Prototype generation on structural data using dissimilarity space representation. *Neural Computing and Applications*, 1–10.

- Calvo-Zaragoza, J., Valero-Mas, J. J., & Rico-Juan, J. R. (2016). Selecting promising classes from generated data for an efficient multi-class nearest neighbor classification. *Soft Computing*, 1–7.

7.2 Future research perspectives

“*Science never solves a problem without creating ten more.*”
George Bernard Shaw

The work in this dissertation constitutes a small contribution to the fields of Automatic Music Transcription and Pattern Recognition which, at least partially, solve the research questions and hypotheses initially proposed. Nevertheless, this work opens different paths and perspectives to be addressed in the future.

A first point is the extension of the proposed interactive paradigms to other tasks within the Music Information Retrieval field. As it has been shown, the interactive schemes proved to remarkably reduce the user effort in the onset annotation process. Thus, it seems promising to extend these interactive methodologies for other tasks such as *chord estimation* or *beat induction* in which the label of a particular analysis window (e.g., the chord name of an audio excerpt or whether certain time instant contains a beat of the piece) influences the labels of the posterior windows.

The second extension that could be addressed is the further exploration of multimodality for improving the performance of the transcription scheme. In this context the use of harmony information, such as chord or key descriptions should, in principle, report improvements in transcription systems as it would narrow the space of possible solutions. While some works have already explored this idea (e.g., the work by Benetos et al. (2014) with key information or the work by Laaksonen (2014) for melody transcription with chord information), the examples in the literature are still relatively scarce, possibly due to the difficulty of properly merging the different sources of information.

Closely related to this multimodal proposal is the recent appearance of Music Language Models for Automatic Music Transcription. Somehow replicating the general workflow in speech recognition systems in which an acoustic model is post-processed with a language model to correct errors in the estimation by applying prior knowledge of the language at issue, some researchers in the MIR field have started exploring this path. While results reported by some researchers are not yet conclusive about the usefulness of Music Language Models in the context of Automatic Music Transcription (cf. to the work by Sigtia et al. (2016)), the general intuition suggests that this piece of information is of complete relevance for the success of the task but that the main issue resides in how to properly combine the acoustic model with the language-based one.

Other recent techniques more related to the low-level processing of the signal are the ones related to the recent development of Deep Learning. The main advantage of these techniques resides in their capability of obtaining proper feature representations for the task at issue (onset detection, chord recognition, pitch estimation...). In this sense, and given the extraordinary results being achieved by this relatively novel paradigm, it is of remarkable interest to further explore in that research direction.

A more practical point to address is the further development of new data collections. Generally, research in Automatic Music Transcription is biased towards piano-related music as a large number of data collections consider those types of instruments. While there exist some datasets considering other timbral spaces, they generally constitute very scarce examples compared to the piano-based ones. Thus, in order to obtain more general conclusions, it is of remarkably interest the development of new datasets which allow gathering more general insights.

Finally, a last point we consider to explore is the possibility of studying and adapting generic Prototype Selection methods for the particular case of class-imbalance situation with no need of an initial data-based balancing process. From our point of view this is a research point of remarkable relevance since, while Prototype Selection schemes generally consider that the representation of the classes in the data is balanced, in real-world data this assumption is hardly ever correct, and thus alternative methods should be proposed.

Appendix A

Resumen

A.1 Introducción

Desde que originalmente fuera propuesto y acuñado por Kassler (1966), el campo de la *Extracción y recuperación de información musical* (del inglés Music Information Retrieval, MIR) ha sido ampliamente estudiado y analizado por la comunidad científica con el objetivo de definir sus líneas y ramas de investigación.

Entre las diversas referencias de este campo, destacan dos definiciones representativas del mismo. La primera es la propuesta por Orio (2006) que define el área como "*el campo de investigación dedicado a satisfacer las necesidades musicales de los usuarios*". La segunda definición representativa de este campo es la facilitada en Serra et al. (2013) en la que se señala como "*un campo que cubre todos los temas de investigación relacionados con el modelado y la comprensión de la música y su relación con las tecnologías de la información*".

Dentro de este contexto, esta disertación se desarrolla dentro de la llamada *transcripcin automática de música por computador* (del inglés Automatic Music Transcription, AMT). Desde un punto de vista de la musicología, la trancripción se entiende como la "*la representación de la ejecución de una pieza musical en algún tipo de notación o cifrado*" (Gallagher, 2009). Por tanto, el campo AMT se puede entender como el equivalente computacional de esta tarea, es decir, la creación y desarrollo de algoritmos capaces de codificar una ejecución musical en una notación simbólica de alto nivel.

Dada la extensiva investigación desarrollada en este campo, podemos encontrar diferentes definiciones representativas del mismo. Por ejemplo, Klapuri (2004b) lo define como:

> *"la transformaciíon de una señal acústica en notación simbólica"*

Por otro lado, Pertusa (2010) enfatiza el hecho de que la abstracción sea comprensible por un posible usuario final:

> *"la extracción de una representación legible e interpretable por un humano, como por ejemplo una partitura musical, a partir de una señal de audio"*

Como último ejemplo, Benetos (2012) también remarca, al menos de manera implícita, la necesidad de obtener una representación legible por un humano:

> *"el proceso de conversión de una grabación sonora en una representación simbólica con algún tipo de notación musical"*

Por lo expuesto, AMT puede ser considerado como el proceso por el cual se obtiene una abstracción simbólica de alto nivel del contenido musical de una señal de audio utilizado las tecnologías de la información. Sin embargo, el punto clave es que la representación sea computable para permitir a otras tareas del campo MIR utilizarla. Además, esta codificación ha de permitir su traducción a cualquier tipo de notación musical. Un ejemplo de codificación que cumple estos requisitos es la creada por el consorcio Music Encoding Initiative (MEI).

Sin embargo, la representación más extendida en sistemas AMT prácticos es la pianola (en inglés, *piano roll*). Esta codificación es básicamente un gráfico bidimensional en la cual el eje de abscisas representa la evolución temporal de la pieza musical mientras que el eje de ordenadas codifica el contenido en altura, típicamente notas musicales. Por tanto, cada coordenada del gráfico muestra qué la actividad o inactividad de cada nota para un instante temporal dado. La Fig. A.1 muestra un ejemplo gráfico de la misma.

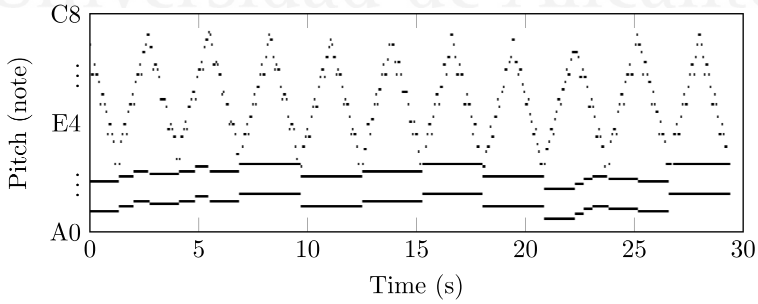


Figure A.1: Ejemplo de una representación de pianola: la evolución temporal y el contenido en notas musicales están se representan en los ejes de abscisas y ordenadas, respectivamente.

La utilidad de AMT en el campo de la música es bastante significativa, habiendo llegado a ser definida como "*el Santo Grial en el campo del análisis* *musical*" (Benetos et al., 2012). Por un lado, en algunos casos como es la preservación de música por medio de partituras (digitales), el resultado de AMT constituye una finalidad por sí misma, ya que el objetivo es la obtención de esa codificación; por otro lado, para otras tareas en MIR (por ejemplo, la búsqueda de música por similitud), sistemas musicales interactivos (por ejemplo, el seguimiento automático de partituras) o análisis computacional de música, AMT constituye un proceso de obtención de una notación intermedia para afrontar el resto del problema. La Fig. A.2 muestra un resumen gráfico de las posibles aplicaciones de AMT.

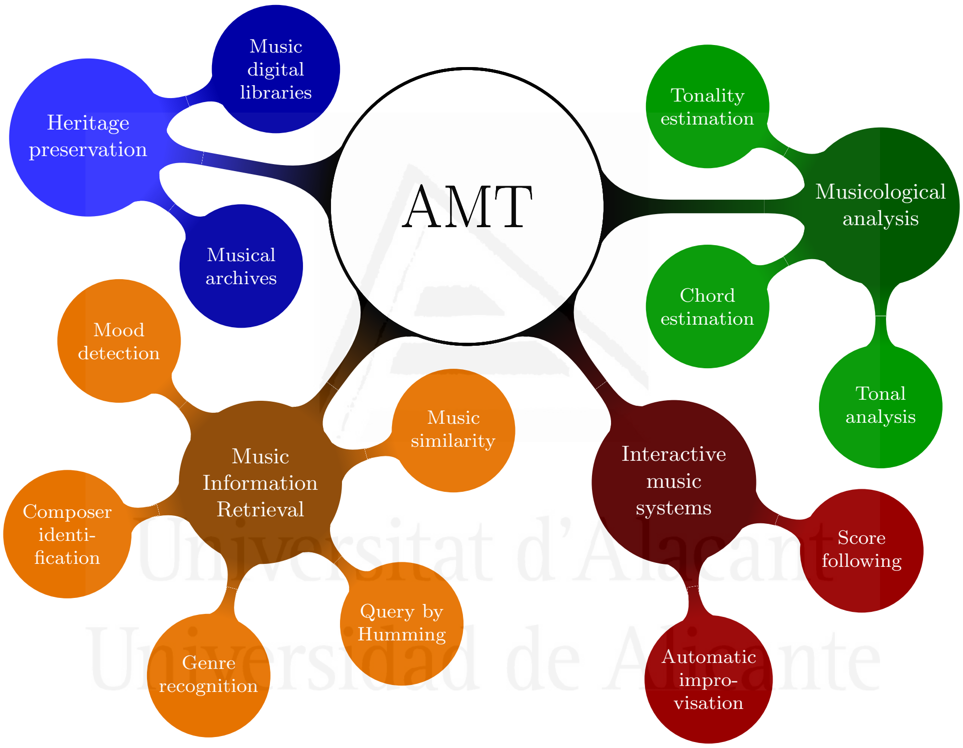


Figure A.2: Ejemplos representativos de aplicación de la transcripción automática de música (AMT).

La mayoría de los sistemas AMT están basados en procesos de dos etapas (Benetos et al., 2012): una primera denominada *estimación de frecuencias fundamentales* (del inglés Multi-pitch Estimation, MPE), típicamente considerada el núcleo de todo sistema AMT, en la cual se estiman las frecuencias fundamentales de los sonidos presentes en la señal; y una segunda etapa, denominada *conformado de notas* (del inglés Note Tracking, NT), que procesa el resultado del método MPE para obtener un evento de nota musical representado por un valor discreto de altura, un inicio de nota (en inglés, *onset*) y un final de nota (del ingles, *offset*). Por tanto, mientras que el primero de los procesos se encarga de obtener una descripción básica del contenido en frecuencias de las señal a transcribir, la segunda actúa como una etapa de corrección y segmentación cuya finalidad es la obtención de una representación musical de alto nivel.

La principal problemática de los sistemas MPE reside en el nivel de polifonía de la pieza a transcribir (Grosche et al., 2012). Por un lado, mientras que la transcripción de piezas monofónicas ha sido ampliamente estudiada, e incluso ha llegado a ser considerada una tarea resuelta por muchos autores, para el caso polifónico todavía constituye un problema de investigación abierto y con grandes limitaciones a salvar (Klapuri, 2004b; Argenti et al., 2011). Además, dentro de los sistemas AMT destaca el hecho de que gran parte de la investigación se ha dedicado a la etapa de MPE en detrimento de la de NT, lo cual puede deberse a la dependencia de los resultados de la segunda con los de la primera (Duan & Temperley, 2014).

Recientemente, una parte notable de autores e investigadores en el campo apuntan a que, a de algún modo, parece estar alcanzándose un límite tecnológico en cuanto a las metodologías típicamente aplicadas (Benetos et al., 2012): por una parte, los resultados cuantitativos obtenidos con conjuntos de datos de referencia para la evaluación parecen estar estancados con pequeñas mejoras marginales ; además, la mayor parte de los sistemas parece estar demasiados adecuados a ciertos tipos de datos por tanto obviando la flexibilidad que se espera de este tipo de herramientas; por último también destaca el hecho de que existan ejemplos muy escasos de sistemas AMT capaces de obtener partituras musicales legibles y comprensibles por humanos. Por tanto parece necesaria, y totalmente justificada, la búsqueda y consideración de paradigmas alternativos a los típicamente considerados (Benetos, Dixon, et al., 2013).

En este sentido, algunos autores han comenzado a incorporar procesos adicionales al esquema mencionado anteriormente. En su mayoría, estos procesos son directamente otras tareas de MIR las cuales facilitan descripciones adicionales de la señal a transcribir como, por ejemplo, información sobre armonía, descripción rítmica, fuentes de sonido, instrumentación, etc. De alguna manera estas informaciones complementarias imponen ciertas restricciones a los métodos de MPE y NT para reducir el espacio de búsqueda, imitando de alguna manera la manera en que el ser humano utiliza información contextual a la hora de afrontar una transcripción musical (transcripción multimodal).

Sin embargo, incluso con el uso de estas descripciones adicionales en los sistemas AMT, uno de los principales problemas yace en que ninguno de los componentes puede ser considerado totalmente libre de error, requiriéndose por tanto una inspección manual por parte de un usuario para encontrar y corregir los errores cometidos por el sistema. Por tanto, dado el hecho de que es necesario mantener un usuario externo para la correcta consecución de la tarea, algunos sistemas están comenzando a considerar al usuario como una parte activa del sistema AMT. Benetos, Dixon, et al. (2013) resumen estas ideas de interacción y multimodalidad para sistemas AMT por medio del esquema que muestra la Fig. A.3.

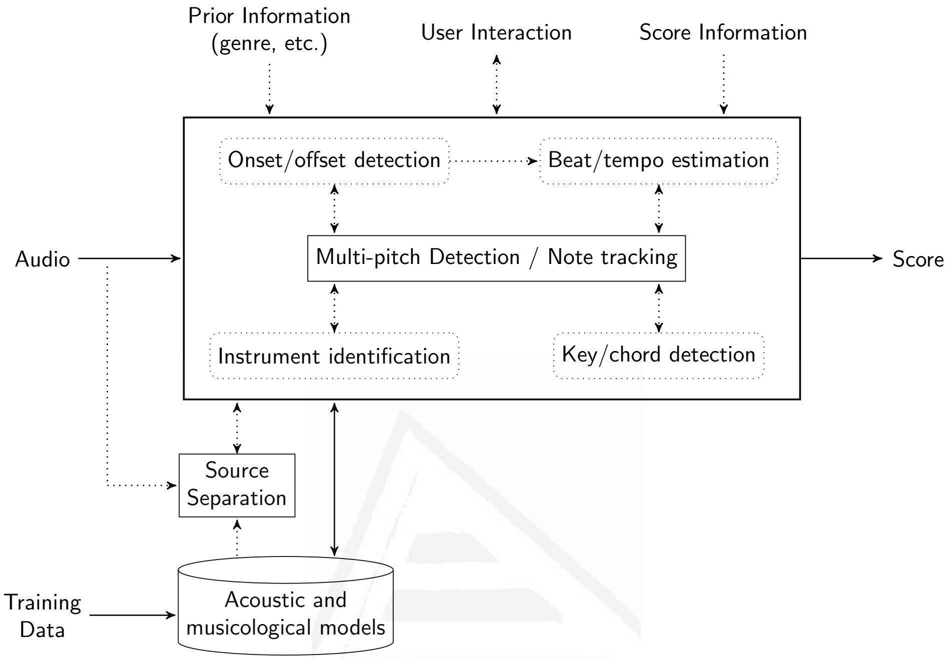


Figure A.3: Esquema genérico del sistema de transcripción de música propuesto por Benetos, Dixon, et al. (2013). Las líneas punteadas representan tareas adicionales a las típicamente consideradas en estos esquemas.

Como se puede observar, en esta propuesta el núcleo del sistema todavía depende de los procesos MPE y NT. Sin embargo, se puede observar que estos procesos reciben ahora información adicional como es la información de onsets y/o offsets, descripciones rítmicas e incluso análisis armónicos para mejorar el rendimiento y precisión del sistema. Además, este esquema conceptual también contempla el hecho de que informaciones totalmente externas a la señal a transcribir en sí puedan ser empleadas para la transcripción, como son los modelos computacionales de teoría musical, principios de organología o las particularidades del género de la música a transcribir. Cabe destacar que, como se ha comentado antes, toda esta información también puede ser facilitada por un usuario en el contexto de sistemas de transcripción interactiva.

Motivación y objetivos

El punto de partida de esta disertación son los conceptos introducidos sobre interacción y multimodalidad aplicados a AMT. Concretamente, exploramos el uso de la información de onsets aplicados a esta finalidad. Los eventos de inicio de nota (onsets) constituyen una importante fuente de informacinón para la segmentación temporal y la descripción rítmica de las señales musicales, siendo además de gran utilidad en las etapas NT para la definición de los eventos de nota musical como fuente de información multimodal (Grosche et al., 2012).

Por otro lado, como se ha comentado, la estimación de esta descripción de la señal no está exenta de errores. Es por ello que esta tarea constituye un ejemplo idóneo para el estudio de metodologías aplicadas a la transcripción de música.

Finalmente cabe destacar que, a pesar de que estos conceptos se pueden estudiar desde mútiples perspectivas, en el caso de esta disertación la mayor parte del trabajo se enfocará desde una perspectiva de la rama del Reconocimiento de Formas (en inglés Pattern Recognition, PR). Es por ello que algunas de las aportaciones que se mostrarán se enfocan desde un punto más generalista y no necesariamente centrados en datos de corte musical.

A.2 Contribuciones

Esta sección describe de una manera breve los aportes de esta disertación. En concreto estas aportaciones se pueden en tres grandes grupos temáticos: *i)* estudios sobre algoritmos de selección de prototipos en entornos de clasificación basados en la regla del vecino más cercano (del inglés k-Nearest Neighbor, kNN) y desarrollados en el Capítulo 4; *ii)* propuestas para modelos interactivos de anotación y corrección de onsets en señales de audio, llevadas a cabo en el Capítulo 5; y *iii)* desarrollo de nuevas propuestas para esquemas de conformado de notas, las cuales se estudian en el Capítulo 6.

Los siguientes apartados resumen de manera breve las propuestas, desarrollos y resultados obtenidos para cada uno de estos grupos temáticos.

Aportes a la selección de prototipos para el clasificador kNN

El clasificador de vecino más cercano (kNN) constituye uno de los ejemplos más representativos del paradigma *lazy learning*, es decir, los clasificadores que no derivan un modelo a partir de los datos de entrenamiento sino que clasifican nuevas muestras directamente comparando con las de referencia.

Una de las grandes ventajas de kNN es precisamente que para cambiar su comportamiento basta con cambiar su conjunto de datos de entrenamiento, convirtiéndolo por tanto en un clasificador ideal para entornos interactivos. Sin embargo, el hecho de que no derive un modelo a partir del conjunto de datos de entrenamiento hace que el proceso de clasificación se ralentice considerablemente. Es por ello que aparecen alternativas que permiten optimizar el tamaño del conjunto de datos para que, con la menor pérdida posible en la tasa de acierto, el modelo sea lo más eficiente posible. De las posibles técnicas existentes nos centramos en los llamados algoritmos de selección de prototipos (del inglés Prototype Selection, PS), los cuales obtienen un subconjunto de entrenamiento a base de eliminar prototipos del conjunto inicial que sólo aportan redundancias y/o ruido que ha de ser inferior en tamaño al original.

El primer aporte en este grupo temático ha sido el estudio comparativo de una serie de algoritmos de PS de reciente aparición propuestos por Rico-Juan and Iñesta (2012). La particularidad de estos algoritmos es su simplicidad conceptual y su gran eficiencia computacional: los prototipos del conjunto de entrenamiento *votan* a los elementos que maximizan la tasa de clasificación y después se establece un umbral por el que sólo los prototipos con cierta cantidad de votos mínima son trasladados al conjunto reducido. El trabajo por tanto ha consistido en la experimentación exhaustiva contra otra serie de algoritmos PS típicamente utilizados. Los resultados permiten inferir una serie de conclusiones de gran importancia: por un lado se puede ver que estos métodos de PS presentan una gran robustez sin la necesidad de una etapa previa de eliminación de ruido; por otro lado también se demuestra que estos algoritmos permiten una gran reducción del conjunto de entrenamiento sin una pérdida significativa de la bondad en la tasa clasificación; por último cabe destacar la gran eficiencia computacional de estos métodos debido a su simplicidad conceptual y al bajo coste de las operaciones requeridas.

La segunda contribución aportada en este capítulo ha sido el estudio de los algoritmos de PS en los llamados entornos no balanceados, es decir, colecciones de datos en las que las clases no están igualmente representadas. Los métodos de PS están diseñados para conjuntos de datos en los que la cantidad de ejemplos por clase es, aproximadamente, la misma; sin embargo, esta condición no se suele dar en datos reales. En este contexto hemos comparado el comportamiento de estos algoritmos de PS en entornos no balanceados contra los casos en los que se han aplicados técnicas de equilibrado de datos. Los resultados obtenidos muestran que los métodos de PS obtienen mejores resultados cuando las diferentes clases del problema están igualmente representadas. Es por ello que los procesos de equilibrado de datos cobran especial relevancia en este contexto como técnicas de preprocesado.

Modelos interactivos para la estimación de onsets en señales de audio

La informacinón de onsets ha demostrado ser de gran utilidad para una gran cantidad de tareas en el campo de MIR, siendo una de ellas la transcripción de música. Dada la relevancia de esta información en este capítulo se propone el paradigma de la anotación y corrección interactiva de onsets.

La primera contribución en este sentido es la propuesta de una serie de métricas para la evaluación cuantitativa del esfuerzo realizado por un usuario en el proceso de anotación de onsets. A pesar de la gran cantidad de literatura en el campo de la estimación de onsets, no es posible encontrar referencia alguna a trabajos que se encarguen de medir, al menos de manera cuantitativa, el esfuerzo de un humano en el proceso de anotación de un corpus de onsets. Es por ello que las métricas propuestas comparan el coste de anotar una pieza en un sistema interactivo contra la anotación manual de todos los eventos del fichero de audio y contra la corrección totalmente manual de una estimación inicial dada por un algoritmo de estimación de onsets autónomo.

Como segundo aporte se proponen una serie de esquemas interactivos para la anotación de onsets en señales de audio desde un punto de vista de técnicas de procesado de señal. La idea es que las anotaciones y correcciones apuntadas por el usuario cambien los parámetros del algoritmo de estimación de onsets para adecuarse paulatinamente a la señal a analizar. Los resultados experimentales con las métricas de esfuerzo previamente citadas demuestran que estos esquemas son capaces de reducir significativamente el esfuerzo requerido por parte del usuario en el proceso de anotación. Sin embargo, la principal limitación de estos esquemas radica en que las correcciones realizadas sobre una pieza sólo afectan a esa pieza en sí y no a futuros casos que puedan venir. Se hace por tanto interesante el explorar paradigmas basados en modelos con una mayor *plasticidad* para su modificación como son, por ejemplo, los sistemas basados en aprendizaje automático.

La tercera y última contribución en este campo es otra serie de esquemas interactivos para la anotación de onsets pero enfocados como sistemas de clasificación. En este contexto de sistemas interactivos de PR toma especial relevancia el clasificador kNN debido a que el añadir nuevos ejemplos a su conjunto de entrenamiento es suficiente para modificar su funcionamiento. La idea en este caso es modelar la detección de onsets como una tarea de clasificación en la que cada ventana de análisis de la señal se cataloga como portadora o no de un evento de onset. Con estas premisas se han estudiado diferentes políticas de interacción en las que cada corrección del usuario aporta más o menos datos al conjunto de entrenamiento. Los experimentos realizados muestran que los mejores resultados (es decir, mayor reducción de esfuerzo por parte del usuario) se obtienen cuanta más información es facilitada por el bucle de realimentación del sistema interactivo.

Estudios sobre conformado de notas en transcripción de música

La información de onsets ha sido ampliamente utilizada para tareas de transcripción de música por paliar significativamente la imprecisión temporal de los algoritmos de estimación de frecuencias fundamentales. Sin embargo, a pesar de su continua utilización en este contexto concreto, no existe ningún estudio formal que analice y compare el impacto que supone la utilización o no de información de onsets en sistemas de transcripción.

Es por ello que el primer aporte de este capítulo consiste precisamente en la realización de este estudio formal. En primer lugar se han seleccionado una serie de algoritmos de estimación de onsets representativos de la literatura además de la utilización de la información de referencia de los onsets de la señal para simular el caso un detector de onsets con funcionamiento perfecto; por otro lado también se han considerado unos sistemas de estimación de frecuencias fundamentales (MPE) que obtienen resultados de calidad respaldados por la literatura del campo de AMT; por último, para la unificación de ambas fuentes de informacinón se ha considerado la utilización de transductores de estados finitos (del inglés Finite State Transducer, FST). Los resultados muestran que el uso de la información de onsets mejora la estimación inicial siempre que la política de fusión sea la adecuada; por otro lado también se demuestra que los resultados obtenidos con los algoritmos que actualmente definen el estado de la cuestión en detección de onsets distan bastante de los conseguidos con la información de onsets manualmente anotada y considerada de referencia, dando así a entender que estas técnicas todavía tienen recorrido para la mejora.

La segunda y última contribución se centra en la tarea de conformado de notas. En general, estos sistemas NT suelen estar formados por una serie de reglas heurísticas que procesan el resultado del proceso de estimación de frecuencias fundamentales (MPE) para conformar un evento de nota musical. En este trabajo se estudia la posibilidad de modelar este problema como una tarea de clasificación y que el sistema, en lugar de recibir impuestas una serie de reglas manualmente definidas, sea capaz de inferirlas de manera automática. Los resultados obtenidos muestran que el esquema propuesto es capaz de igualar a otras estrategias de conformado de notas aún sin llegar, por lo general, a superarlas. Sin embargo, esta propuesta constituye un ejemplo de paradigma alternativo para casos de NT no considerado previamente, constituyendo así una contribución por sí misma con un largo recorrido para ser explorada y mejorada.

A.3 Conclusiones y trabajo futuro

Esta disertación ha tratado el tema de la transcripción de música desde audio considerando dos perspectivas diferentes pero complementarias: *i)* el aspecto *interactivo*, el cual considera la inclusión del usuario como parte activa de la transcripción en lugar de como un simple agente de verificación y corrección; y *ii)* la perspectiva *multimodal*, la cual establece la necesidad de diferentes descripciones de la señal musical para la obtención de transcripciones simbólicas precisas y fiables.

La consideración de esquemas interactivos puede verse como un alejamiento del concepto ideal de los sistemas de transcripcinón totalmente autónomos. Sin embargo, esta disertación toma como punto de partida las conclusiones apuntadas por diferentes investigadores relevantes en el campo sobre un posible límite tecnoógico que los sistemas de transcripción están alcanzando, siendo por tanto necesario un cambio en el paradigma de funcionamiento. Dado que ningún sistema es totalmente fiable y libre de error, normalmente se precisa de un agente humano para revisar y corregir la estimación dada por el sistema. En ese sentido, dada esta limitación, es necesario estudiar estrategias interactivas para la explotación eficiente de este esfuerzo humano.

En lo relativo a la perspectiva multimodal, en esta disertación hemos estudiado la utilización de la información de onsets como una fuente adicional para el proceso de transcripción. Aunque este tipo de información ya ha sido empleado anteriormente en sistemas AMT, en este trabajo hemos extendido su estudio en el contexto de sistemas de conformado de notas o NT. Además, dada la relevancia de la información de onsets para la correcta consecución de la tarea de transcripción, todos los estudios relacionados con el citado paradigma interactivo han considerado el caso concreto de sistemas para el anotado y corrección de eventos de onset en señales musicales de audio.

Por último también cabe destacar la notable presencia y relevancia de los sistemas de Reconocimiento de Formas (PR) en este trabajo. Es por ello que, además de las contribuciones hechas en MIR, también se han realizado aportes al desarrollo de técnicas propias de esta disciplina. Más concretamente, los estudios han tenido como objetivo el análisis sobre cómo tratar las situaciones en las que, siendo necesario un proceso de selección de prototipos para reducir el tamaño del conjunto de datos, las distribuciones de los mismo no están equilibradas sino que existe un sesgo hacia alguna de ellas (conocido en inglés como *class imbalance problem*).

Aportes realizados

Los principales aportes y conclusiones obtenidos con el desarrollo de esta disertación son los que se listan a continuación:

1. Una exhaustiva revisión del estado de la cuestión tanto en los temas de AMT como de PR aplicados al campo de MIR.

2. La evaluación exhaustiva y la comparativa de un conjunto novedoso de técnicas de PS para el clasificador kNN basadas en principios de ordenación y llamadas Nearest to Enemy (NE) y Farthest Neighbor (FaN). Los resultados obtenidos demuestran la competitividad de estas novedosas técnicas tanto en reducción de tamaño del conjunto de datos y eliminación de ruido como en su bajo coste computacional en comparación a otros métodos más sofisticados.

3. Un estudio comparativo sobre el uso de técnicas de PS para el clasificador kNN en el contexto particular de clasificación en entornos de clases no equilibradas pero con suficiente cantidad de prototipos como para requerir un proceso de reducción del conjunto de entrenamiento.

Los experimentos llevados a cabo muestran que, en general, es beneficioso aplicar ténicas para el equilibrado artificial de las clases antes del proceso de selección para mejorar la respuesta de los segundos. Concretamente, las combinaciones de los principios de *Oversampling* y *Undersampling* para el equilibrado de clases antes del proceso PS son los que mejores resultados reportan, tomando como criterio el compromiso entre la reducción de datos y la tasa de clasificación.

4. La propuesta de una serie de figuras de mérito para la evaluación cuantitativa del esfuerzo invertido por el usuario en el proceso de anotación y/o corrección de una estimación inicial de onsets.

5. Una colección de sistemas interactivos para la anotación de onsets basados en técnias de procesado de señal. Este conjunto de esquemas está basado en el esquema clásico de estimación de onsets en dos etapas, es decir, una primera dedicada a obtener la función de detección de onsets (del inglés Onset Detection Function, ODF)seguida de la función de selección de onsets (del inglés Onset Selection Function, OSF); las correcciones llevadas a cabo por el usuario modifican los parámetros de la segunda etapa para adecuar la forma de la curva de selección a la pieza en cuestión a ser anotada. Los resultados experimentales muestran que las metodologías de interacción propuestas constituyen un acercamiento simple a la par que efectivo para la reducción de la carga de trabajo de anotación y corrección por parte del usuario.

6. La propuesta y estudio de una serie de metodologías de corrección y anotación de onsets desde el punto de vista del Reconocimiento de Formas Interactivo (del inglés Interactive Pattern Recognition, IPR). En primera instancia, el problema de la estimación de onsets es modelado como una tarea de clasificación en la que cada ventana de análisis es categorizada como conteniendo un onset o no. En este contexto, la modificación y actualización del modelo de clasificación se consigue modificando directamente el conjunto de entrenamiento de acuerdo a las correcciones de usuario. Por tanto, la cuestión de investigación a responder a estudiar es qué información es relevante para la mejora del modelo. Las conclusiones extraídas tras la exhaustiva experimentación llevada a cabo apuntan que cuanto mayor es la cantidad de información que se suministra al modelo por realimentación, mayor es la robustez del modelo y la reducción de carga de trabajo sobre el usuario final.

7. Un estudio sobre la potencial mejora que supone el uso de información de onsets en sistemas de transcripción de música. Cabe resaltar que la utilización de la información de onsets para transcripción no constituye una novedad por sí misma. Por ello, la novedad que se presenta en esta disertación viene dada por la mencionada evaluación formal de la

potencial mejora que se puede obtener con este tipo de información en comparación con sistemas que directamente la ignoran. Además, una contribución adicional en este contexto es la evaluación y estudio de la relación entre la bondad del algoritmo de estimación de onsets y la calidad de la transcripción final.

8. La introducción de una novedosa estrategia de conformado de notas (NT) enfocada como un problema de clasificación. En general, los sistemas NT suelen estar formados por una serie de reglas heurísticas que procesan la estimación de frecuencias fundamentales (MPE) para conformar un evento de nota musical. En este trabajo se ha considerado un sistema basado en un clasificador para que, de alguna manera, esas reglas fueran obtenidas de manera automática. Los resultados obtenidos muestran que el esquema propuesto es capaz de igualar a otras estrategias de NT aunque sin llegar a superarlas, al menos significativamente. Sin embargo, esta propuesta constituye un ejemplo de paradigma alternativo para casos de NT, lo cual ya constituye una contribución por sí misma.

Perspectivas para futura investigación

El trabajo presentado en esta disertación constituye una pequeña aportación a los campos de la transcripción automática de música (AMT) y el Reconocimiento de Formas (PR) las cuales, al menos de manera parcial, dan respuesta a las hipótesis planteadas al inicio. Sin embargo, y como en cualquier trabajo de investigación, el trabajo presentado abre una serie de caminos y perspectivas a considerar en futuras investigaciones.

Un primer punto a considerar es la extensión de los paradigmas interactivos propuestos a otras tareas de MIR. Como se ha demostrado, estos esquemas son capaces de reducir significativamente el esfuerzo del usuario en la anotación de eventos de onset. Es por tanto que parece prometedor el extender estas metodologías a otras tareas como la *estimación de acordes* o la *extracción del pulso musical* ya que, en estas tareas, la etiqueta dada a una ventana de análisis tiene gran influencia sobre las etiquetas de las ventanas posteriores.

El segundo trabajo futuro que se considera es continuar la exploración de la multimodalidad para la mejora de los sistemas AMT. En este contexto la utilización de información relacionada con la armoía de la música, como información de acordes o de tonalidad, debería suponer una mejora en los resultados ya que, de alguna manera, acotaría el espacio posible de soluciones. Mientras que algunos trabajos ya han explorado esta idea (por ejemplo, el trabajo de Benetos et al. (2014) en el cual se utiliza información sobre la tonalidad para restringir los valores de frecuencia fundamental posibles o el trabajo de Laaksonen (2014) en el que se utiliza información sobre acordes para la transcripción de melodías), los ejemplos en la literatura son todavía escasos, en gran medida por la dificultad intrínseca que conlleva la mezcla de ambas descripciones de la señal.

Estrechamente relacionado con la multimodalidad encontramos el concepto de reciente aparición de Modelo de Lenguaje Musical (del inglés Music Language Model, MLM). La idea de este paradigma es replicar el esquema de los sistemas de reconocimiento de voz en los que, a partir de una estimación inicial de la información presente en la señal por un modelo acústico, un modelo de lenguaje entrenado con información simbólica corrige errores que se hayan dado en la primera etapa. Aunque los resultados obtenidos hasta ahora por diferentes investigadores no son totalmente concluyentes acerca de la utilidad de los MLM en el contexto de la transcripción (un ejemplo de esto puede ser encontrado en el trabajo de Sigtia et al. (2016)), la intuición apunta a que este tipo de descripción debería ser de gran relevancia para la correcta consecución de la tarea. Es por ello que la principal cuestión de investigación parece residir en cómo aunar ambas fuentes de información.

Otra potencial fuente a explorar está relacionada con el procesado a bajo nivel de la señal por medio de esquemas de Aprendizaje Profundo (del inglés Deep Learning). La principal ventaja de este tipo de técnicas reside en que son capaces de obtener por sí mismas una representación en formato vector de características que se adecúan a la tarea en concreto (*feature learning*). En este sentido, dado el importante avance que este tipo de esquemas ha supuesto en otros campos como el procesado de imagen o incluso en otras disciplinas dentro del MIR, destaca como una vía importante a explorar.

Desde un punto de vista práctico destaca la necesidad de crear nuevos conjuntos de datos. En general, gran parte de la investigación en el campo de AMT está claramente orientada a timbres estilo piano debido a la relativa sencillez de crear conjuntos de datos de este tipo. Aunque existen colecciones de datos con otros timbres, tienden a ser las menos y, por tanto, los resultados que se obtienen no se pueden considerar del todo concluyentes. En ese sentido es necesaria la creación de conjuntos de datos con un tamaño relativamente grande que permitan evaluar los sistemas existentes de una manera concluyente.

Finalmente, un último punto a explorar es la posibilidad de adaptar los métodos de selección de prototipos (del inglés Prototype Selection, PS) para el caso particular en que las distribuciones de las clases de los datos no están balanceadas sin necesidad de un algoritmo previo que modifique los datos en sí para equilibrarlos artificialmente. Desde nuestro punto de vista esto constituye un punto de gran importancia a investigar ya que, mientras que los algoritmos de PS asumen que las clases en el conjunto de datos están balanceadas, en datos reales obtenidos fuera del laboratorio esto no suele ser así. Es por ello que consideramos esta investigación necesaria.

References

Abdallah, S., & Plumbley, M. (2003). Unsupervised onset detection: a probabilistic approach using ICA and a hidden Markov classifier. In *Proceedings of the Cambridge Music Processing Colloquium.* Cambridge, United Kingdom.

Alonso, M., Richard, G., & David, B. (2007). Tempo Estimation for Audio Recordings. *Journal of New Music Research*, *36*(1), 17–25.

Alvarado, P. A., & Stowell, D. (2016). Gaussian Processes for Music Audio Modelling and Content Analysis. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP).*

Angiulli, F. (2007). Fast Nearest Neighbor Condensation for Large Data Sets Classification. *IEEE Transactions on Knowledge and Data Engineering*, *19*(11), 1450–1464.

Argenti, F., Nesi, P., & Pantaleo, G. (2011). Automatic music transcription: from monophonic to polyphonic. In *Musical Robots and Interactive Multimodal Systems* (pp. 27–46). Springer.

Arı, İ., Şimşekli, U., Cemgil, A. T., & Akarun, L. (2012). Large Scale Polyphonic Music Transcription Using Randomized Matrix Decompositions. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)* (pp. 2020–2024).

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., ... Vilar, J.-M. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics*, *35*(1), 3–28.

Bay, M., Ehmann, A. F., & Downie, J. S. (2009, October). Evaluation of Multiple-F0 Estimation and Tracking Systems. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 315–320). Kobe, Japan.

Bellet, A., Bernabeu, J. F., Habrard, A., & Sebban, M. (2016). Learning discriminative tree edit similarities for linear classification – Application to melody recognition. *Neurocomputing*, *214*, 155–161.

Bello, J. P. (2003). *Towards the Automated Analysis of Simple Polyphonic Music: A Knowledge-based Approach* (PhD Thesis). Queen Mary University of London.

Bello, J. P., Daudet, L., Abdallah, S. A., Duxbury, C., Davies, M. E., & 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. P., Daudet, L., & Sandler, M. B. (2006). Automatic piano transcription using frequency and time-domain information. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(6), 2242–2251.

Bello, J. P., Duxbury, C., Davies, M., & Sandler, M. (2004). 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. P., & Sandler, M. (2003). Phase-based note onset detection for music signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Vol. 5, pp. 441–444).

Benetos, E. (2012). *Automatic transcription of polyphonic music exploiting temporal evolution* (PhD Thesis). Queen Mary University of London.

Benetos, E., Cherla, S., & Weyde, T. (2013, September). An efficient shift-invariant model for polyphonic music transcription. In *6th International Workshop on Machine Learning and Music (MML), In conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD).* Prague, Czech Republic.

Benetos, E., & Dixon, S. (2011). Polyphonic music transcription using note onset and offset detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 37–40).

Benetos, E., & Dixon, S. (2012). A shift-invariant latent variable model for automatic music transcription. *Computer Music Journal*, *36*(4), 81–94.

Benetos, E., & Dixon, S. (2013). Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model. *The Journal of the Acoustical Society of America*, *133*(3), 1727–1741.

Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., & Klapuri, A. (2012, October). Automatic Music Transcription: Breaking the Glass Ceiling. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR).* Porto, Portugal.

Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., & Klapuri, A. (2013). Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, *41*(3), 407–434.

Benetos, E., Jansson, A., & Weyde, T. (2014). Improving automatic music transcription through key detection. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio.*

Benetos, E., & Stylianou, Y. (2010). Auditory Spectrum-Based Pitched Instrument Onset Detection. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(8), 1968–1977.

Benetos, E., & Weyde, T. (2015). An efficient temporally-constrained probabilistic model for multiple-instrument music transcription. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 701–707). Málaga, Spain.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, *18*(9), 509–517.

Berg-Kirkpatrick, T., Andreas, J., & Klein, D. (2014). Unsupervised transcription of piano music. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 1538–1546).

Bernabeu, J. F., Calera-Rubio, J., Iñesta, J. M., & Rizo, D. (2011). Melodic Identification Using Probabilistic Tree Automata. *Journal of New Music Research*, *40*(2), 93-103.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* New York, New York, USA: Springer-Verlag.

Böck, S., Krebs, F., & Schedl, M. (2012). Evaluating the Online Capabilities of Onset Detection Methods. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 49–54). Porto, Portugal.

Böck, S., & Schedl, M. (2012). Polyphonic piano note transcription with recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 121–124).

Böck, S., Schlüter, J., & Widmer, G. (2013, September). Enhanced peak picking for onset detection with recurrent neural networks. In *6th International Workshop on Machine Learning and Music (MML), In conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD).* Prague, Czech Republic.

Böck, S., & Widmer, G. (2013a, November). Local Group Delay based Vibrato and Tremolo Suppression for Onset Detection. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 589–594). Curitiba, Brazil.

Böck, S., & Widmer, G. (2013b, September). Maximum Filter Vibrato Suppression for Onset Detection. In *Proceedings of the 16th International Conference on Digital Audio Effects (DAFx-13)* (pp. 55–61). Maynooth, Ireland.

Boulanger-Lewandowski, N., Bengio, Y., & Vincent, P. (2012). Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *Proceedings of the International Conference on Machine Learning (ICML).*

Box, J. (2013). *Evaluación de sistemas de detección de Onsets en señales musicales* (Diploma Thesis). University of Alicante, Spain.

Brighton, H., & Mellish, C. (2002). Advances in instance selection for instance-based learning algorithms. *Data Mining and Knowledge Discovery*, *6*(2), 153–172.

Brossier, P. M. (2006). *Automatic annotation of musical audio for interactive applications* (PhD Thesis). Queen Mary University of London.

Bunke, H., & Riesen, K. (2012). Towards the unification of structural and statistical pattern recognition. *Pattern Recognition Letters*, *33*(7), 811–825.

Calvo-Zaragoza, J. (2016). *Pattern Recognition for Music Notation* (PhD Thesis). Universidad de Alicante.

Calvo-Zaragoza, J., & Oncina, J. (2014). Recognition of Pen-Based Music Notation: the HOMUS dataset. In *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR)* (pp. 3038–3043). Stockholm, Sweden.

Calvo-Zaragoza, J., & Oncina, J. (2017). An efficient approach for Interactive Sequential Pattern Recognition. *Pattern Recognition*, *64*, 295–304.

Calvo-Zaragoza, J., Valero-Mas, J. J., & Rico-Juan, J. R. (2015a). Improving kNN multi-label classification in Prototype Selection scenarios using class proposals. *Pattern Recognition*, *48*(5), 1608–1622.

Calvo-Zaragoza, J., Valero-Mas, J. J., & Rico-Juan, J. R. (2015b, June). Prototype Generation on Structural Data using Dissimilarity Space Representation: A Case of Study. In R. Paredes, J. S. Cardoso, & X. M. Pardo (Eds.), *Proceedings of the 7th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)* (pp. 72–82). Santiago de Compostela, Spain: Springer.

Calvo-Zaragoza, J., Valero-Mas, J. J., & Rico-Juan, J. R. (2016a). Prototype generation on structural data using dissimilarity space representation. *Neural Computing and Applications*, 1–10.

Calvo-Zaragoza, J., Valero-Mas, J. J., & Rico-Juan, J. R. (2016b). Selecting promising classes from generated data for an efficient multi-class nearest neighbor classification. *Soft Computing*, 1–7.

Cambouropoulos, E. (2010). The Musical Surface: Challenging Basic Assumptions. *Musicae Scientiae*, *14*(2), 131–147.

Cañadas-Quesada, F. J., Ruiz-Reyes, N., Vera-Candeas, P., Carabias-Orti, J. J., & Maldonado, S. (2010). A Multiple-F0 Estimation Approach Based on Gaussian Spectral Modelling for Polyphonic Music Transcription. *Journal of New Music Research*, *39*(1), 93-107.

Cano, J. R., Herrera, F., & Lozano, M. (2006). On the Combination of Evolutionary Algorithms and Stratified Strategies for Training Set Selection in Data Mining. *Applied Soft Computing*, *6*(3), 323–332.

Cazau, D., Revillon, G., Krywyk, J., & Adam, O. (2015). An investigation of prior knowledge in Automatic Music Transcription systems. *The Journal of the Acoustical Society of America*, *138*(4), 2561–2573.

Cazau, D., Wang, Y., Chemillier, M., & Adam, O. (2016). An automatic music transcription system dedicated to the repertoires of the marovany zither. *Journal of New Music Research*, *45*(4), 343–360.

Cemgil, A. T. (2004). *Bayesian Music Transcription* (PhD Thesis). Radboud University Nijmege.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Cheng, T., Dixon, S., & Mauch, M. (2015). Improving piano note tracking by HMM smoothing. In *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)* (pp. 2009–2013).

Cheng, T., Mauch, M., Benetos, E., & Dixon, S. (2016, August). An Attack/Decay Model for Piano Transcription. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 584–590). New York, USA.

Chuan, C.-H., & Chew, E. (2008). *Audio Onset Detection Using Machine Learning Techniques: The Effect and Applicability of Key and Tempo Information* (Tech. Rep.). California, USA: University of Southern California Computer Science Department.

Cogliati, A., Duan, Z., & Wohlberg, B. (2015). Piano music transcription with fast convolutional sparse coding. In *Proceedings of the 25th IEEE International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1–6).

Collins, N. (2005). Using a Pitch Detector for Onset Detection. In *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 100–106). London, UK.

Conklin, D. (2013). Multiple Viewpoint Systems for Music Classification. *Journal of New Music Research*, *42*(1), 19-26.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21-27.

Dasarathy, B. V., Sánchez, J. S., & Townsend, S. (2000). Nearest Neighbour Editing and Condensing Tools-Synergy Exploitation. *Pattern Analysis and Applications*, 19–30.

de Andrade Scatolini, C., Richard, G., & Fuentes, B. (2015). Multipitch estimation using a PLCA-based model: Impact of partial user annotation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 186–190).

de Cheveigné, A. (2006). Multiple F0 Estimation. In *Computational auditory scene analysis: Principles, algorithms, and applications.* Wiley-IEEE Press.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, *39*(1), 1–38.

Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, *7*, 1–30.

Derrac, J., Cornelis, C., García, S., & Herrera, F. (2012). Enhancing evolutionary instance selection algorithms by means of fuzzy rough set based feature selection. *Information Sciences*, *186*(1), 73–92.

Dessein, A., Cont, A., & Lemaitre, G. (2010). Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 489–494).

Dittmar, C., & Abeßer, J. (2008, March). Automatic music transcription with user interaction. In *Proceedings of the 34th Jahrestagung für Akustik (DAGA).* Dresden, Germany.

Dixon, S. (2001). An interactive beat tracking and visualisation system. In *Proceedings of the International Computer Music Conference (ICMC)* (pp. 215–218). San Francisco, EEUU.

Dixon, S. (2006). Onset detection revisited. In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx)* (pp. 133–137). Montreal, Canada.

Dorran, D., & Lawlor, R. (2004). Time-scale modification of music using a synchronized subband/time-domain approach. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 225–228). Montreal, Canada.

Downie, J. S. (2003). Music information retrieval. *Annual Review of Information Science and Technology*, *37*(1), 295–340.

Duan, Z., Pardo, B., & Zhang, C. (2010). Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(8), 2121–2133.

Duan, Z., & Temperley, D. (2014, October). Note-level Music Transcription by Maximum Likelihood Sampling. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 181–186). Taipei, Taiwan.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification.* John Wiley & Sons.

Duxbury, C., Bello, J. P., Davies, M., & Sandler, M. (2003). Complex Domain Onset Detection for Musical Signals. In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx)* (pp. 90–93). London, UK.

Duxbury, C., Sandler, M., & Davis, M. (2002). A Hybrid Approach to Musical Note Onset Detection. In *Proceedings of the Digital Audio Effects Workshop (DAFx)* (pp. 33–38).

Ellis, D. P. W. (2007). Beat tracking by dynamic programming. *Journal of New Music Research*, *36*(1), 51–60.

Emiya, V., Badeau, R., & David, B. (2008). Automatic transcription of piano music based on HMM tracking of jointly-estimated pitches. In *Proceedings of the 16th European Signal Processing Conference (EUSIPCO)* (pp. 1–5).

Emiya, V., Badeau, R., & David, B. (2010). Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(6), 1643–1654.

Emiya, V., David, B., & Badeau, R. (2007). A parametric method for pitch estimation of piano tones. In *Proceedings of the IEEE International*

Conference on Acoustics, Speech, and Signal Processing (ICASSP) (Vol. 1, pp. 249–252).

Eshelman, L. J. (1990, July). The CHC Adaptive Search Algorithm: How to Have Safe Search When Engaging in Nontraditional Genetic Recombination. In *Proceedings of the first workshop on foundations of genetic algorithms* (pp. 265–283). Indiana, USA.

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

Fernández, A., García, S., & Herrera, F. (2011). Addressing the classification with imbalanced data: open problems and new challenges on class distribution. In *International Conference on Hybrid Artificial Intelligence Systems* (pp. 1–10).

Fix, E., & Hodges, J. L. (1951). Discriminatory analysis, nonparametric discrimination: Consistency properties. *US Air Force School of Aviation Medicine*, *Technical Report 4*(3), 477.

Freeman, H. (1961). On the Encoding of Arbitrary Geometric Configurations. *IRE Transactions on Electronic Computers*, *EC-10*(2), 260–268.

Friedman, J. H., Bentley, J. L., & Finkel, R. A. (1977). An Algorithm for Finding Best Matches in Logarithmic Expected Time. *ACM Transactions on Mathematical Software*, *3*(3), 209–226.

Fritsch, J. (2012). *High Quality Musical Audio Source Separation* (Master Thesis). Université Pierre et Marie Curie (UPMC) / Institut de Recherche et Coordination Acoustique/Musique (IRCAM) / Télécom ParisTech, France.

Frühwirth, M., & Rauber, A. (2002). Self-Organizing Maps for Content-Based Music Clustering. In *Proceedings of the 12th Italian Workshop on Neural Nets (WIRN)* (pp. 228–233). Vietri sul Mare.

Fuentes, B., Badeau, R., & Richard, G. (2011). Adaptive harmonic time-frequency decomposition of audio using shift-invariant PLCA. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 401–404).

Fuentes, B., Badeau, R., & Richard, G. (2012). Blind harmonic adaptive decomposition applied to supervised source separation. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)* (pp. 2654–2658).

Gallagher, M. (2009). *The Music Tech Dictionary: A Glossary of Audio-*

related Terms and Technologies. Course Technology.

García, S., Derrac, J., Cano, J. R., & Herrera, F. (2012). Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study. *IEEE Transactions on Pattern Analysis Machine Intelligence*, *34*(3), 417–435.

García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining.* Springer.

García, V., Sánchez, J., & Mollineda, R. (2007). An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. In *Iberoamerican Congress on Pattern Recognition* (pp. 397–406).

García, V., Sánchez, J. S., & Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, *25*(1), 13–21.

García-Pedrajas, N., & de Haro-García, A. (2014). Boosting instance selection algorithms. *Knowledge-Based Systems*, *67*, 342–360.

Gionis, A., Indyk, P., & Motwani, R. (1999). Similarity search in high dimensions via hashing. In *Proceedings of the very large data bases conference (vldb)* (pp. 518–529).

Giraldo, S., Ramírez, R., & Rollin, W. (2016, September). Onset detection using Machine Learning Ensemble methods. In *Proceedings of the 9th International Workshop on Machine Learning and Music (MML)* (pp. 21–25). Riva del Garda, Italy.

Glover, J., Lazzarini, V., & Timoney, J. (2011). Real-time detection of musical onsets with linear prediction and sinusoidal modeling. *EURASIP Journal on Advances in Signal Processing*, *2011*(1), 1–13.

Goto, M. (2001). An audio-based real-time beat tracking system for music with or without drum-sounds. *Journal of New Music Research*, *30*(2), 159–171.

Goto, M., Hashiguchi, H., Nishimura, T., & Oka, R. (2002). RWC Music Database: Popular, Classical and Jazz Music Databases. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)* (pp. 287–288). Paris, France.

Grindlay, G., & Ellis, D. (2011). Transcribing Multi-Instrument Polyphonic Music With Hierarchical Eigeninstruments. *Journal of Selected Topics in Signal Processing*, *5*(6), 1159–1169.

Grosche, P., Schuller, B., Müller, M., & Rigoll, G. (2012). Automatic transcription of recorded music. *Acta Acustica united with Acustica*, *98*(2), 199–215.

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing* (pp. 878–887).

Hart, P. (1968). The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, *14*(3), 515–516.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Herbert, T. (2009). *Music in Words: A Guide to Researching and Writing about Music: A Guide to Researching and Writing about Music*. Oxford University Press, USA.

Holzapfel, A., Stylianou, Y., Gedik, A. C., & Bozkurt, B. (2010). Three dimensions of pitched instrument onset detection. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(6), 1517–1527.

Hull, J. (1994). A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *16*(5), 550–554.

Iñesta, J. M., & Pérez-Sancho, C. (2013). Interactive multimodal music transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 211–215).

Kameoka, H., Nakano, M., Ochiai, K., Imoto, Y., Kashino, K., & Sagayama, S. (2012). Constrained and regularized variants of non-negative matrix factorization incorporating music-specific constraints. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 5365–5368).

Kameoka, H., Ochiai, K., Nakano, M., Tsuchiya, M., & Sagayama, S. (2012). Context-free 2D Tree Structure Model of Musical Notes for Bayesian Modeling of Polyphonic Spectrograms. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 307–312).

Kapanci, E., & Pfeffer, A. (2006). A Hierarchical Approach to Onset Detection. In *Proceedings of the International Computer Music Conference (ICMC)* (pp. 438–441). New Orleans, USA.

Kassler, M. (1966). Toward musical information retrieval. *Perspectives of New Music*, *4*(2), 59–67.

Kauppinen, I. (2002, July). Methods for detecting impulsive noise in speech and audio signals. In *Proceedings of the 14th International Conference on Digital Signal Processing (DSP)* (Vol. 2, pp. 967–970).

Kelz, R., Dorfer, M., Korzeniowski, F., Böck, S., Arzt, A., & Widmer, G.

(2016, August). On the Potential of Simple Framewise Approaches to Piano Transcription. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 475–481). New York, USA.

Kirchhoff, H. (2013). *A User-assisted Approach to Multiple Instrument Music Transcription* (PhD Thesis). Queen Mary University of London.

Kirchhoff, H., Dixon, S., & Klapuri, A. (2012). Shift-variant non-negative matrix deconvolution for music transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 125–128).

Kirchhoff, H., Dixon, S., & Klapuri, A. (2013). Missing template estimation for user-assisted music transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 26–30).

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

Klapuri, A. (2003). Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Speech and Audio Processing*, *11*(6), 804–816.

Klapuri, A. (2004a). Automatic Music Transcription as We Know it Today. *Journal of New Music Research*, *33*(3), 269–282.

Klapuri, A. (2004b). *Signal Processing Methods for the Automatic Transcription of Music* (PhD Thesis). Tampere University of Technology.

Klapuri, A., & Davy, M. (2007). *Signal processing methods for music transcription.* Springer Science & Business Media.

Koretz, A., & Tabrikian, J. (2011). Maximum a posteriori probability multiple-pitch tracking using the harmonic model. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(7), 2210–2221.

Kraft, S., & Zölzer, U. (2015). Polyphonic pitch detection by matching spectral and autocorrelation peaks. In *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)* (pp. 1301–1305).

Krumhansl, C. L. (2001). *Cognitive foundations of musical pitch.* Oxford University Press.

Kubat, M., & Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In *Proceedings of the fourteenth international conference on machine learning (icml)* (pp. 179–186).

Laaksonen, A. (2014). Automatic Melody Transcription based on Chord Transcription. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 119–124).

Lacoste, A., & Eck, D. (2007). A Supervised Classification Algorithm for Note Onset Detection. *EURASIP Journal on Advances in Signal Processing*.

Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. In *Proceedings of the conference on artificial intelligence in medicine in europe* (pp. 63–66).

Lee, C.-T., Yang, Y.-H., & Chen, H. H. (2012). Multipitch estimation of piano music by exemplar-based sparse representation. *IEEE Transactions on Multimedia*, *14*(3), 608–618.

Lerch, A., & Klich, I. (2005, April). *On the Evaluation of Automatic Onset Tracking Systems* (Tech. Rep.). Berlin, Germany: Technical University of Berlin.

Leveau, P., Daudet, L., & Richard, G. (2004). Methodology and Tools for the evaluation of automatic onset detection algorithms in music. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR)* (pp. 72–75). Barcelona, Spain.

Lichman, M. (2013). *UCI Machine Learning Repository.* Retrieved from http://archive.ics.uci.edu/ml

Lidy, T., Rauber, A., Pertusa, A., & Iñesta, J. M. (2007). Improving genre classification by combination of audio and symbolic descriptors using a transcription system. In *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 61–66). Vienna, Austria.

List, G. (1974). The Reliability of Transcription. *Ethnomusicology*, *18*(3), 353–377.

Liu, T., Moore, A. W., Yang, K., & Gray, A. G. (2004). An investigation of practical approximate nearest neighbor algorithms. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 825–832).

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113–141.

Marchi, E., Ferroni, G., Eyben, F., Gabrielli, L., Squartini, S., & Schuller, B. (2014). Multi-resolution Linear Prediction Based Features for Audio Onset Detection with Bidirectional LSTM Neural Networks. In

Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 2164–2168).

Marolt, M. (2004). A connectionist approach to automatic transcription of polyphonic piano music. *IEEE Transactions on Multimedia*, *6*(3), 439–449.

Marolt, M., & Divjak, S. (2002). On detecting repeated notes in piano music. In *Proceedings of the 3rd International Society for Music Information Retrieval Conference (ISMIR)* (pp. 273–274).

Marolt, M., Kavcic, A., & Privosnik, M. (2002). Neural Networks for Note Onset Detection in Piano Music. In *Proceedings of the International Computer Music Conference (ICMC).* Gothenburg, Sweden.

Masri, P. (1996). *Computer Modeling of Sound for Transformation and Synthesis of Musical Signals* (PhD Thesis). University of Bristol.

Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., . . . Dixon, S. (2015, May). Computer-aided Melody Note Transcription Using the Tony Software: Accuracy and Efficiency. In *First International Conference on Technologies for Music Notation and Representation (TENOR).* Paris, France.

Mitchell, T. M. (1997). *Machine Learning* (1st ed.). New York, NY, USA: McGraw-Hill, Inc.

Mohri, M., Pereira, F., & Riley, M. (2002). Weighted Finite-State Transducers in Speech Recognition. *Computer Speech & Language*, *16*(1), 69–88.

Moorer, J. A. (1977). On the Transcription of Musical Sound by Computer. *Computer Music Journal*, *1*(4), 32–38.

Muja, M., & Lowe, D. G. (2014). Scalable nearest neighbor algorithms for high dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(11), 2227–2240.

Müller, M., Konz, V., Bogler, W., & Arifi-Müller, V. (2011). *Saarland music data (SMD)* (Tech. Rep.). Miami, USA: Max-Planck Institute für Informatik.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective* (1st ed.). Cambridge, Massachusetts, USA: MIT Press.

Nam, J., Ngiam, J., Lee, H., & Slaney, M. (2011, October). A classification-based polyphonic piano transcription approach using learned feature representations. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 175–180). Miami, USA.

Nanni, L., & Lumini, A. (2011). Prototype reduction techniques: A comparison among different approaches. *Expert Systems with Applications*, *38*(9), 11820–11828.

Natarajan, N., Dhillon, I., Ravikumar, P., & Tewari, A. (2013). Learning with noisy labels. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 1196–1204).

Nettl, B. (2015). *The Study of Ethnomusicology : Thirty-Three Discussions.* Springfield: University of Illinois Press.

O'Hanlon, K., Nagano, H., & Plumbley, M. D. (2012). Structured sparsity for automatic music transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 441–444).

Ojima, Y., Nakamura, E., Itoyama, K., & Yoshii, K. (2016, August). A Hierarchical Bayesian Model of Chords, Pitches, and Spectrograms for Multipitch Analysis. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 309–315). New York, USA.

Oncina, J., & Vidal, E. (2011). Interactive structured output prediction: application to chromosome classification. In *Proceedings of the 5th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)* (pp. 256–264). Las Palmas de Gran Canaria, Spain.

Orio, N. (2006). Music Retrieval: A Tutorial and Review. *Foundations and Trends in Information Retrieval*, *1*(1), 1–96.

Ozerov, A., Vincent, E., & Bimbot, F. (2012). A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(4), 1118–1133.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peeling, P. H., & Godsill, S. J. (2011). Multiple pitch estimation using non-homogeneous Poisson processes. *IEEE Journal of Selected Topics in Signal Processing*, *5*(6), 1133–1143.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, *130*(5), 2902–2916.

Pekalska, E., Duin, R. P., & Paclík, P. (2006). Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, *39*(2), 189–208.

Pérez-García, T., Iñesta, J. M., Ponce de León, P. J., & Pertusa, A. (2011). A multimodal music transcription prototype: first steps in an interactive prototype development. In *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI)* (pp. 315–318).

Pérez-Sancho, C. (2009). *Stochastic Language Models for Music Information Retrieval* (PhD Thesis). Universidad de Alicante.

Pertusa, A. (2010). *Computationally efficient methods for polyphonic music transcription* (PhD Thesis). Universidad de Alicante.

Pertusa, A., & Iñesta, J. M. (2008). Multiple fundamental frequency estimation using Gaussian smoothness. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 105–108).

Pertusa, A., & Iñesta, J. M. (2012). Efficient methods for joint estimation of multiple fundamental frequencies in music signals. *EURASIP Journal on Advances in Signal Processing*, *2012*(1), 1–13.

Pertusa, A., Klapuri, A., & Iñesta, J. M. (2005, November). Recognition of Note Onsets in Digital Music Using Semitone Bands. In *Progress in pattern recognition, image analysis and applications: 10th iberoamerican congress on pattern recognition (ciarp)* (pp. 869–879).

Poliner, G., & Ellis, D. (2007). A discriminative model for polyphonic piano transcription. *EURASIP Journal on Applied Signal Processing*, *2007*(1), 154–154.

Poliner, G. E. (2008). *Classification-based Music Transcription* (PhD Thesis). Columbia University.

Ponce de León, P. J. (2011). *A statistical pattern recognition approach to symbolic music classification* (PhD Thesis). Universidad de Alicante.

Ponce de León, P. J., & Iñesta, J. M. (2002). Musical style identification using self-organising maps. In *Proceedings of the Second International Conference on Web Delivering of Music (WEDELMUSIC)* (pp. 82–89). Darmstadt, Germany.

Ponce de León, P. J., & Iñesta, J. M. (2007). A pattern recognition approach for music style identification using shallow statistical descriptors. *IEEE Transactions on Systems Man and Cybernetics C*, *37*(2), 248–257.

Quinlan, J. R. (2014). *C4.5: Programs for machine learning*. Elsevier.

Raczyński, S. A., Ono, N., & Sagayama, S. (2009). Note detection with dynamic bayesian networks as a postanalysis step for NMF-based multiple pitch estimation techniques. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 49–52).

Raczyński, S. A., Vincent, E., Bimbot, F., & Sagayama, S. (2010). Multiple pitch transcription using DBN-based musicological models. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 363–368).

Raczyński, S. A., Vincent, E., & Sagayama, S. (2013). Dynamic Bayesian networks for symbolic polyphonic pitch modeling. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(9), 1830–1840.

Randel, D. M. (1944). *The Harvard Dictionary of Music.* Belknap Press of Harvard University Press.

Raphael, C. (2005). A Graphical Model for Recognizing Sung Melodies. In *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 658–663).

Rico-Juan, J. R., & Iñesta, J. M. (2012). New rank methods for reducing the size of the training set using the nearest neighbor rule. *Pattern Recognition Letters*, *33*(5), 654–660.

Rizo, D. (2010). *Symbolic music comparison with tree data structures* (PhD Thesis). Universidad de Alicante.

Rosão, C., Ribeiro, R., & Martins de Matos, D. (2012). Influence of peak picking methods on onset detection. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 517–522). Porto, Portugal.

Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Upper Saddle River, New Jersey, USA: Pearson Education, Inc.

Ryynänen, M. P., & Klapuri, A. (2005). Polyphonic music transcription using note event modeling. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 319–322).

Sakoe, H., & Chiba, S. (1990). Readings in Speech Recognition. In A. Waibel & K.-F. Lee (Eds.), *Readings in Speech Recognition* (pp. 159–165). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Salamon, J., & Gómez, E. (2014). MIR.EDU: An open-source library for teaching Sound and Music Description. In *Late Breaking/Demo extended abstract, 15th International Society for Music Information Retrieval (ISMIR).* Taipei, Taiwan.

Schedl, M., Gómez, E., & Urbano, J. (2014). Music Information Retrieval: Recent Developments and Applications. *Foundations and Trends in Information Retrieval*, *8*, 127–261.

Schlüter, J., & Böck, S. (2013). Musical Onset Detection with Convolutional

Neural Networks. In *6th International Workshop on Machine Learning and Music (MML), In conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD).* Prague, Czech Republic.

Schlüter, J., & Böck, S. (2014). Improved musical onset detection with convolutional neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 6979–6983).

Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., ... Widmer, G. (2013). *Roadmap for Music Information ReSearch.* The MIReS Consortium.

Sigtia, S., Benetos, E., Boulanger-Lewandowski, N., Weyde, T., d'Avila Garcez, A. S., & Dixon, S. (2015). A hybrid recurrent neural network for music transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 2061–2065).

Sigtia, S., Benetos, E., Cherla, S., Weyde, T., d'Avila Garcez, A. S., & Dixon, S. (2014). An RNN-based music language model for improving automatic music transcription. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 53–58).

Sigtia, S., Benetos, E., & Dixon, S. (2016). An End-to-End Neural Network for Polyphonic Piano Music Transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(5), 927–939.

Smaragdis, P., & Brown, J. C. (2003). Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 177–180).

Smaragdis, P., & Mysore, G. J. (2009). Separation by "humming": user-guided sound extraction from monophonic mixtures. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 69–72).

Smaragdis, P., Raj, B., & Shashanka, M. (2006). A Probabilistic Latent Variable Model for Acoustic Modeling. In *Advances in Neural Information Processing Systems (NIPS).*

Stowell, D., & Plumbey, M. D. (2007, August). Adaptive whitening for improved real-time audio onset detection. In *Proceedings of the International Computer Music Conference (ICMC)* (pp. 312–319). Copenhagen, Denmark.

Su, L., & Yang, Y.-H. (2015). Combining Spectral and Temporal Representations for Multipitch Estimation of Polyphonic Music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(10), 1600–1612.

Tomek, I. (1976). Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-6*(11), 769–772.

Toselli, A. H., Romero, V., Pastor, M., & Vidal, E. (2010). Multimodal interactive transcription of text images. *Pattern Recognition*, *43*(5), 1814–1825.

Toselli, A. H., Vidal, E., & Casacuberta, F. (2011). *Multimodal interactive pattern recognition and applications.* Springer Science & Business Media.

Tsai, C.-F., Eberle, W., & Chu, C.-Y. (2013). Genetic algorithms in feature and instance selection. *Knowledge-Based Systems*, *39*, 240–247.

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10*(5), 293–302.

Valero-Mas, J. J., Benetos, E., & Iñesta, J. M. (2016, September). Classification-based Note Tracking for Automatic Music Transcription. In *Proceedings of the 9th International Workshop on Machine Learning and Music (MML)* (pp. 61–65). Riva del Garda, Italy.

Valero-Mas, J. J., Benetos, E., & Iñesta, J. M. (2017, June). Assessing the Relevance of Onset Information for Note Tracking in Piano Music Transcription. In *Proceedings of the Audio Engineering Society (AES) International Conference on Semantic Audio.*

Valero-Mas, J. J., Calvo-Zaragoza, J., & Rico-Juan, J. R. (2016). On the suitability of Prototype Selection methods for kNN classification with distributed data. *Neurocomputing*, *203*, 150–160.

Valero-Mas, J. J., Calvo-Zaragoza, J., Rico-Juan, J. R., & Iñesta, J. M. (2016). An experimental study on rank methods for prototype selection. *Soft Computing*, 1–13.

Valero-Mas, J. J., Calvo-Zaragoza, J., Rico-Juan, J. R., & Iñesta, J. M. (2017, June). A study of prototype selection algorithms for nearest neighbour in class-imbalanced problems. In *Proceedings of the 8th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA).* Faro, Portugal.

Valero-Mas, J. J., & Iñesta, J. M. (2015, October). Interactive onset detection in audio recordings. In *Late Breaking/Demo extended abstract, 16th International Society for Music Information Retrieval Conference*

(ISMIR). Málaga, Spain.

Valero-Mas, J. J., & Iñesta, J. M. (2017, July). Experimental assessment of descriptive statistics and adaptive methodologies for threshold establishment in onset selection functions. In *Proceedings of the 14th Sound and Music Computing Conference (SMC).* Espoo, Finland.

Valero-Mas, J. J., Iñesta, J. M., & Pérez-Sancho, C. (2014, November). Onset detection with the user in the learning loop. In *Proceedings of the 7th International Workshop on Machine Learning and Music (MML).* Barcelona, Spain.

Valero-Mas, J. J., Salamon, J., & Gómez, E. (2015, July). Analyzing the influence of pitch quantization and note segmentation on singing voice alignment in the context of audio-based Query-by-Humming. In *Proceedings of the 12th Sound and Music Computing Conference (SMC)* (pp. 371–378). Maynooth, Ireland.

Vidal, E. (1986). An algorithm for finding nearest neighbours in (approximately) constant average time. *Pattern Recognition Letters*, *4*(3), 145–157.

Vidal, E., Rodríguez, L., Casacuberta, F., & García-Varea, I. (2008). Interactive Pattern Recognition. In *Proceedings of the 4th International Conference on Machine Learning for Multimodal Interaction (MLMI)* (pp. 60–71). Brno, Czech Republic.

Vincent, E., Bertin, N., & Badeau, R. (2010). Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(3), 528–537.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*(2), 260–269.

Wagner, R. A., & Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the ACM*, *21*(1), 168–173.

Weninger, F., Kirst, C., Schuller, B., & Bungartz, H.-J. (2013). A discriminative approach to polyphonic piano note transcription using supervised non-negative matrix factorization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 6–10).

West, K., & Cox, S. (2005). Finding An Optimal Segmentation for Audio Genre Classification. In *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 680–685). London, UK.

Wilkinson, R. A., Geist, J., Janet, S., Grother, P. J., Burges, C. J., Creecy, R., . . . Wilson, C. L. (1992). *The First Census Optical Character Recognition System Conference* (Tech. Rep.). US: Department of Commerce.

Wilson, D. L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man and Cybernetics*, *SMC-2*(3), 408–421.

Wilson, D. R., & Martinez, T. R. (1997). Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, *6*, 1–34.

Wilson, D. R., & Martinez, T. R. (2000). Reduction techniques for instance-based learning algorithms. *Machine learning*, *38*(3), 257–286.

Yang, Y.-H., & Chen, H. H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology*, *3*(3), 40.

Yeh, C. (2008). *Multiple fundamental frequency estimation of polyphonic recordings* (PhD Thesis). Université Paris VI - Pierre et Marie Curie.

Yoshii, K., & Goto, M. (2012). A nonparametric Bayesian multipitch analyzer based on infinite latent harmonic allocation. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(3), 717–730.

Zappa, F., & Vai, S. (1982). *The Frank Zappa Guitar Book*. Hal Leonard Publishing.

Zhou, R., Mattavelli, M., & Zoia, G. (2008). Music Onset Detection Based on Resonator Time Frequency Image. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(8), 1685–1695.

Zhou, R., & Reiss, J. D. (2007). *Music Onset Detection Combining Energy-Based and Pitch-Based Approaches* (Tech. Rep.). Vienna, Austria: Centre for Digital Music, Queen Mary University of London.