FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Multi-Channel Approaches For Musical Audio Content Analysis

João Carlos Couto Antunes Fonseca

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Supervisor at INESC TEC: Matthew E. P. Davies, Ph.D.

Supervisor at FEUP: Gilberto Bernardes, Ph.D.

July 31, 2020

Resumo

Detecção Automática de Início de Notas Musicais e Separação Musical são dois dos tópicos mais amplamente estudados no domínio da Recuperação de Informação Musical (MIR). Vários métodos foram propostos em pesquisas anteriores, no entanto estes são maioritariamente adaptados para a música ocidental, com menor foco noutras tradições musicais do mundo.

O objectivo deste projecto de investigação é realizar uma avaliação crítica das capacidades das estratégias de detecção automática de início de notas musicais e de separação musical para a observação de micro-ritmos numa prática musical não-ocidental, o "Maracatu de Baque Solto" do Brasil. Esta tradição é originária da região do Pernambuco no nordeste do Brasil, e é caracterizada pela sua natureza altamente percussiva, onde os percussionistas tocam em estreita proximidade entre si e, tipicamente, tocam tão alto e tão rápido quanto possível.

Esta dissertação contrasta três diferentes formas de análise de conteúdo micro-rítmico neste estilo de música afro-latino americano, nomeadamente através de: i) gravações misturadas em estéreo ou mono; ii) sinais separados obtidos através de técnicas de separação musical do estado da arte; e iii) sinais multipista perfeitamente separados, obtidos via microfones de contacto. Esta análise será realizada num conjunto de sinais individuais de cada instrumento do Maracatu que formam o conjunto de dados de trabalho e que foram totalmente anotados.

As abordagens elaboradas mostraram que é possível observar perfis micro-rítmicos através das visualizações calculadas com as estimativas de início de notas a partir de microfones de contacto. No entanto, as estruturas micro-rítmicas desaparecem nas visualizações calculadas com as estimativas de início de nota obtidas através de sinais misturados ou sinais separados automaticamente.

Abstract

Automatic Onset Detection and Music Source Separation are two of the most widely studied tasks in the domain of Music Information Retrieval (MIR). Various methods have been proposed in previous research but these are largely adapted for western music, with less focus on other music traditions of the world.

The goal of this research project is to undertake a critical evaluation of the capabilities of automatic onset estimation strategies and musical source separation for microtiming visualisation in a very specific non-western musical practice, Brazilian "Maracatu de Baque Solto". This music originates from the Pernambuco region in north eastern Brazil, and is characterised by its highly percussive nature, where the percussionists perform in close proximity to one another and typically play as loud and as fast as possible.

The work in this dissertation will contrast three different means for undertaking the analysis of micro-rhythmic content in this Afro-Latin American music, namely through the use of: i) stereo or mono mixed recordings; ii) separated sources obtained via state of the art musical audio source separation techniques; and iii) perfectly separated multi-track stems obtained via contact microphones. This analysis will be conducted on a Maracatu dataset of individual instrument signals which has been collected and annotated.

The devised approaches showed that it is possible to observe microtiming profiles when working with onset estimations obtained from contact microphones. However, the micro-rhythmic structures vanish when working with the onset information obtained from mixtures or automatically separated signals.

Agradecimentos

Em primeiro lugar, quero agradecer aos meus pais por todo o amor incondicional que sempre me dedicaram. Apesar das adversidades, a minha formação foi sempre colocada em primeiro lugar e esta dissertação representa também o culminar dos seus esforços neste longo caminho. Obrigado Pai. Obrigado Mãe.

Agradeço especialmente à minha namorada Filipa, pela ternura, amabilidade e coragem, por me tornar uma pessoa melhor, por nos termos encontrado e por ter aceite caminhar ao meu lado.

Agradeço à minha família por serem a base onde tudo está alicerçado, em especial à minha avó Esmeralda e ao meu padrinho Rui.

A todos os amigos que fui fazendo ao longo da vida e que formam uma família com diferentes apelidos, a vós: Miguel, Borges, Jorge, Rita, Adães, Francisco, Teresa, Pedro, Pinto, Ricardo, Mariana, e ainda ao núcleo duro da Engenharia Rádio: Hugo, Gonçalo, Gomes, Aranha, Wolfs.

Finalmente, ao Matthew, por ter tornado real a oportunidade de realizar este trabalho, pela amizade e conhecimento que sempre partilhou e pelos quais agradeço. Agradeço ainda ao Gilberto, pelo acompanhamento e pelos seus esforços na minha rápida integração no grupo de Sound and Music Computing.

João Fonseca

"The archer sees the mark upon the path of the infinite,
and He bends you with His might
that His arrows may go swift and far.
Let your bending in the archer's hand
be for gladness;
For even as He loves the arrow that flies,
so He loves also the bow that is stable."

Kahlil Gibran

Contents

List of Figures

List of Tables

Abbreviations

HELP-MD	The Healing and Emotional Power of Music and Dance
MIR	Music Information Retrieval
MSS	Music Source Separation
BPM	Beats Per Minute
TF	Time-Frequency
STFT	Short-Time Fourier Transform
GWF	Generalized Wiener Filter
ICA	Independent Component Analysis
PCA	Principal Component Analysis
IID	Inter-channel Intensity Differences
IPD	Inter-channel Phase Differences
LGM	Local Gaussian Model
KAM	Kernel Additive Modeling
NMF	Non-negative Matrix Factorization
KL	Kullback-Leibler
NMFD	Non-Negative Matrix Factor Deconvolution
DNN	Deep Neural Network
FNN	Feed-Forward Neural Network
RNN	Recurrent Neural Network
MWF	Multi-channel Wiener Filter
BLSTM	Bidirectional Long Short-Term Memory
DL	Deep Learning
AQO	Audio Quality Oriented
SO	Significance Oriented
ODF	Onset Detection Function
CARAT	Computer-Aided Rhythm Analysis Toolbox

Chapter 1

Introduction

Music has always been a key factor in social and cultural practices throughout history. Whether we refer to timeless human emotions like celebration or fear, music has assumed various forms through time and spread over many cultures since the early civilisations. Understanding our musical patrimony means understanding our pathway as a species, and where we're headed [4].

How we create, consume, or connect with music today has changed drastically in the last decades. The digital era opened many doors and gave rise to many new research areas that focus on understanding, at a deeper level, through computational means, the intricacies of our musical heritage. Along the way, music research has provided technical solutions to catalogue, produce, and proliferate the deep-rooted connections that humans have with music.

Music information retrieval (MIR) is one of those research areas. Defined as a multidisciplinary research area that gathers knowledge from audio signal processing, musicology, and machine learning, its aims include extracting meaningful information from music sources, in a process known as feature extraction, and indexing music collections using the extracted features enabling music classification or computing similarity between two musical pieces [5].

All of the mentioned tasks are the foundations for the development of different types of applications that retrieve musical information. For instance, acoustic fingerprinting allows track identification by analysing only a fragment of the audio [6] or chord identification that recognises chords of any given music file [7].

1.1 Context

The bond between music and sacred traditions has long been a fascination of ethnomusicological field research. To the same extent, the use of music to promote health and well-being within a particular cultural context widens the allure. Communities from around the globe have developed rituals with symbolic, religious, and emotional meanings that culminate in expressive strategies for playing together that differ in subtle ways to those in Western culture.

During Carnival season, the community of Zona da Mata Norte region, in Northeast Brazil, gathers itself around the "Maracatu de Baque Solto", a mystical performance-ritual that allows people to prevent spiritual and physical attacks that could result in illness and even death.

Inserted in this theme is the FCT-funded research project HELP-MD, "The Healing and Emotional Power of Music and Dance" [8, 9], to which this dissertation is aligned, that explores the hypothesis that if music is widely associated with healing practices in many societies from around the world, this could be due to its potential to elicit and control emotions, whether this happens through symbolic associations or aesthetic meanings attributed to musical forms. To this end, MIR techniques, such as automatic onset detection and microtiming analysis, provide mechanisms to perform an exploratory analysis of the qualities of the music in the percussionists' performances.

1.2 Motivation

The task of deriving insights from a music tradition is preceded of computational strategies of segmenting a music piece into its structural parts. Segmentation strategies allow observation of the organisation of music pieces, whether looking into the metrical structure or the interrelating temporal scales like beat and onset information, that can characterise a particular style.

The work presented in this dissertation undertakes an evaluation of signal representations for musical audio content analysis. In particular, it will evaluate the capabilities of onset estimation strategies for the analysis of micro-rhythmic content in Afro-Latin American music, namely through the use of: i) perfectly separated multi-track stems obtained via contact microphones; ii) stereo or mono mixed recordings; and iii) separated sources obtained via state of the art musical audio source separation techniques.

1.3 Goals

The emphasis of this thesis is on the evaluation of onset estimation approaches for the study of musical micro-rhythm in Maracatu. Our research goals are summarised as follows:

- Compile and annotate a dataset of mixed and multi-channel recordings of the Brazilian "Maracatu de Baque Solto" tradition;
- Conceive methods for the visualisation of rhythmical micro-variations and pattern recognition;
- Evaluate the performance of automatic onset estimation approaches in three different signal scenarios;
- Explore the effectiveness of music source separation when analysing micro-rhythmic content;
- Compare the rhythmic analysis obtained from the three different signal scenarios regarding microtiming identification.

1.4 Structure

This dissertation is structured in five chapters. Excluding this introductory chapter, the remainder are structured as follows:

- *Chapter 2: Background and State of the Art* - This chapter provides an overview of concepts addressed in this thesis. It introduces the concepts of microtiming, automatic onset estimation, and music source separation followed by a bibliographical review of each research area.

- *Chapter 3: Maracatu de Baque Solto* - This chapter shortly presents the "Maracatu de Baque Solto" tradition and the instruments used by the performers. It also describes the data acquisition process a summary of the collected dataset to be used in this work.

- *Chapter 4: Proposed Approach* - This chapter unveils the formulation of the problem tackled in this work along with a flowchart describing the designed strategy for the proposed solution.

- *Chapter 5: Automatic Onset Estimation* - This chapter details the annotation process of the collected dataset, presents the scenarios in which we perform the automatic onset estimations and measure the performance of each approach.

- *Chapter 6: Microtiming Visualisation* - This chapter demonstrates the microtiming visualisation that is possible when we use manually annotated beat and onset information and contrast it with the visualisation computed with the estimations obtained from the three signal scenarios.

- *Chapter 7: Conclusion* - This chapter concludes this dissertation by presenting the main findings from this work and alluding to promising directions for future research.

1.5 Project Acknowledgment and Contributions

The work in this dissertation is conducted within the context of the project "The Healing and Emotional Power of Music and Dance" (HELP-MD), PTDC/ART-PER/29641/2017 which is supported by funds through the FCT - Foundation for Science and Technology, I.P.

Results from this dissertation contributed to the following paper:

- Matthew E. P. Davies, Magdalena Fuentes, João Fonseca, Luís Aly, Marco Jerónimo and Filippo Bonini Baraldi. "Moving in Time: Computational Analysis of Microtiming in Maracatu de Baque Solto" To appear in 21st International Society for Music Information Retrieval Conference, ISMIR, 2020 [10].

1.6 Sound Examples

A set of demonstrative audio and video excerpts of the Maracatu de Baque Solto instruments can be found here.[1]

[1] https://tinyurl.com/ybrc8ux8

Chapter 2

Background and State of the Art

2.1 Introduction

In this chapter, we introduce the relevant concepts related to this work. We start by introducing the central music concepts to better understand the terminology used throughout this document. Subsequently, we present an overview of the three main areas of research encompassed by this work and summarise the main techniques, relevant previous work, and how they apply in the context of this dissertation.

2.2 Definition of Musical Concepts

In this section we briefly introduce the relevant musical terminology commonly used in the music information retrieval research field for a deeper understanding of the concepts explained further in the document.

Metrical Structure is a written representation of organised pulsations that are periodically grouped and form patterns of points in time, hierarchically organised into levels as shown in Figure 2.1 (here divided in beat and downbeat levels).

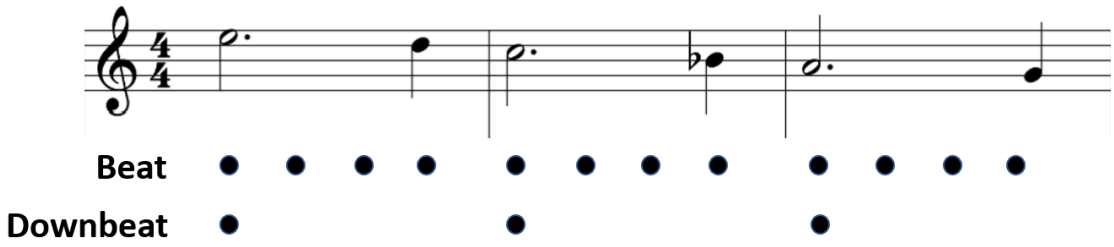

Figure 2.1: Illustration of metrical structure. Excerpt of "Blue in Green" by Miles Davies. *Adapted from:* [1].

Beat is the most salient pulsation of a metrical level, which matches the tap or clap of a person when listening to a music piece. The beats can be grouped into bars. The first beat of each bar is called the downbeat.

Bar is a segment of time corresponding to a specific number of beats in which each beat is represented by a particular note value and the boundaries of the bar are indicated by vertical bar lines.

Timbre is the unique auditory sensation of any given instrument that enables a listener to distinguish two nonidentical sounds, identically presented with the same loudness and pitch.

Tempo is related to the frequency of the main pulsation of a music piece and it can be variable over time. It is usually measured in beats per minute (BPM).

Onset is the time instant of the detectable start of a melodic note, a harmonic change or percussion stroke. Onset times provide information about the actual timing of the articulated events, independently of the metrical structure.

Microtiming is the small-scale deviation of the articulated onsets that are not synchronised with the metrical structure. Different from other forms of tempo variations, like *rubato* or *accelerando*, and more like conscious events shifts at a constant tempo throughout a music piece. Microtiming can be "unsystematic", i.e. resulting from natural variation in performance (so-called "motor noise" in human movement) or "systematic", i.e. intentional deviations from strict metronomic timing.

2.3 Microtiming Analysis

2.3.1 Introduction

Rhythmic patterns can often be used to identify a particular musical style. They can be defined as a series of structured and hierarchical organised pulsations inserted into a metrical structure that regulates them as points in time. In Western music theory, rhythmic phenomena are translated evenly on a score.

In some music genres, notes are sometimes played with a slight deviation from the instants that the score transmits. These rhythmic and micro-rhythmic structures contribute to a specific character of the music that humans describe as "swing" or "groove" [11]. Throughout many Afro-rooted music traditions, small but systematic timing deviations from straight patterns can be found, meaning that, the occurrence is a stylistic requirement and a conscious decision that Western music theory struggles to define.

It is a distinct attribute of music performance and should not be confused with tempo variations, like *rubato* or *accelerando*, but rather, understood as event shifts at a constant tempo [12], known in the scientific community as *microtiming*.

This *micro-rhythm* (the result of the microtiming) is dependent on the position in the metrical structure of the musical piece so that the expected position can be compared to the observed microtiming deviation. Microtiming is recognised in many musical styles such as Jazz, characterised by a specific rhythmic pattern deviation known as *swing* or *swing feel* where consecutive eighth-notes are played unevenly, "performed as long-short patterns" [13, 14], or in the Uruguayan Candombe drumming that shows another type of time granularity [14].

On numerous occasions, microtiming deviations have been identified by researchers in many Brazilian *samba* variations [15, 16]. The type of microtiming in this Brazilian rhythm, on par with Uruguayan Candombe and various *samba* related genres, is usually specified by the onset anticipations around the third and fourth semiquavers (or "16th notes") in a beat [15, 17], considering quarter-note-long patterns.

2.3.2 Previous Work

Microtiming studies have been carried out in the past addressing *samba* related genres such as Samba de Roda, Samba Carioca, Samba de Enredo, Partido-Alto [16, 15]. To study the expressiveness of these timing patterns related to quarter-note segments it is essential to first recognise the expected position of the event within the metrical structure. The deviating percussive strokes can be obtained from the onset times, tempo, or beat structure [18].

Gouyon [16] studied microtiming in Samba de Roda under the premise that deviations occurred around quarter-note segments. In the first stage of the investigation, a semi-automatic beat tracking algorithm was used to segment the musical data orienting the tracking to quarter-notes instead of bars, a different level of the metrical hierarchy. This was followed by a second stage of the analysis, namely, feature extraction using an onset detection function (ODF) to compute the complex spectral difference to segment meaningful patterns, in this case, quarter-note patterns. This action allowed the detection of slight differences in the Inter-Beat Intervals (IBI) where the variations occurred. In the final stage, the patterns extracted from the feature vectors were grouped using *k-means*, a clustering algorithm, that allowed the characterisation of the recording. It was concluded and observed that "both the third and fourth 16th note beats are slightly ahead of their corresponding quantized positions" [16].

A similar approach was taken in [15], where Naveda et al. explored the microtiming and its effect on several musical properties such as intensity, meter, and timbre on a large dataset (106 excerpts) that covered "Samba Carioca", "Samba de Enredo", "Partido-Alto" and "Samba de Roda". In this work the audio segmentation is a two-part process: first, segmentation of the frequency domain spectrum to average out the loudness auditory curves to three spectrum regions (low, mid and high frequency); The second part consists of segmentation of three metric levels in the temporal domain, features of lengths 1, 2 and 4 beats. To compute the microtiming features, each of the segments is analysed concerning the mathematical semiquaver subdivision where an

algorithm, for each spectral region, searched for the highest peak situated around the first beat and then computed the average peak position of the three spectral regions to bridge imprecise segmentation. Similar to [16], to find common patterns between these events, Naveda et. al applied a *k-means* clustering algorithm and concluded that the measured deviations "could be a strategy to induce tension, ambiguity, and flexibility in the rhythmic texture" [15].

Both approaches match the conclusion that the majority of patterns present a local maximum at the 16th note level and that the third and fourth 16th notes in the analysed patterns tend to be executed ahead of their correspondent positions.

On an alternative route of analysis, Dittmar et. al [19] proposed a different way of representing microtiming deviations, specifically in jazz. This tool was dubbed the *swingogram*, a time-swing ratio representation that measures the level of *swing* of jazz solo recordings and that aims to provide tools that alleviate the annotation process in a semi-automatic manner. The authors inspect the rhythmic patterns composed of two 8th notes and do so by determining the signals' similarity using autocorrelation retrieved from an ODF as a feature, obtaining the relative position of events for further examination. Laroche et. al in [20] tackle the problem of jointly estimate the tempo, swing ratio, and beat positions in audio recordings assuming constant tempo. The method uses a preliminary transient detection stage where note onsets or percussion hits are detected, followed by a maximum likelihood estimation of the tempo, swing, and downbeat. These are obtained from the transient times or the inter-transient elapsed times as long as the transient detection can identify salient features in the audio track.

Fuentes et al. in [21, 18] developed a combined model for automatic tracking of beats and microtiming profiles within rhythmic patterns with four articulated sixteenth notes. The proposal exploits a class of statistical modeling methods called Conditional Random Fields, often used for pattern recognition and prediction, and uses beat and onset activations derived from deep learning models as observations to study beat–length rhythmic patterns of timekeeper instruments. In this system a microtiming descriptor is proposed that shows how the articulated sixteenth notes deviate within the rhythmic pattern from their isochronous expected positions. This descriptor is composed of three microtiming–ratios, whose inter-onset intervals are defined concerning the beginning of the beat instead of the previous onset. The model was applied to Afro-Latin American music (Brazilian *samba* and Uruguayan *canbombe*) and used a ground-truth derived from annotated onsets.

2.3.3 Evaluation in the Context of this Work

In the context of this work, we will evaluate the presence of microtiming qualitatively. Our goal here is not to explain the reasons behind this phenomenon, which form part of the larger aims of the "HELP-MD" FCT research project, nor to characterise it in a quantitative fashion, but rather to identify some patterns present on different tracks and on different instruments.

In order to order to visualise the microtiming, we'll use the CARAT library[1], an open-source

[1] https://carat.readthedocs.io

library for computer-aided rhythm analysis from audio recordings. This toolkit was developed in Python and aims to provide ways for the analysis of individual audio files or collections of music data, enabling visualisation and listening to the user [22].

This visualisation is possible by taking as input the onsets of the timekeeper instrument, that is responsible for marking the beats, and the onsets of another instrument, usually with a more expressive role in the performance. Later in this work, we'll try to assess if this qualitative evaluation made with ground-truth data is possible with automatically generated data in different signal separation scenarios.

2.4 Automatic Onset Detection

2.4.1 Introduction

Music is an event-based phenomenon described by musical concepts such as rhythm, melody, or harmony. Onset detection is the process of locating events in a music signal, whether we refer to a percussive stroke or a played note in a piano.

An onset marks the beginning of a single event and is a fundamental part of segmenting a music signal. Unlike other studies that focus on beat and tempo estimation, the onset detection process can be considered independent of periodicity [23].

The established form of detecting onset locations is to identify the "transient" regions in the signal where it is usual to observe a sudden burst of energy [2], as shown in Fig. 2.2.

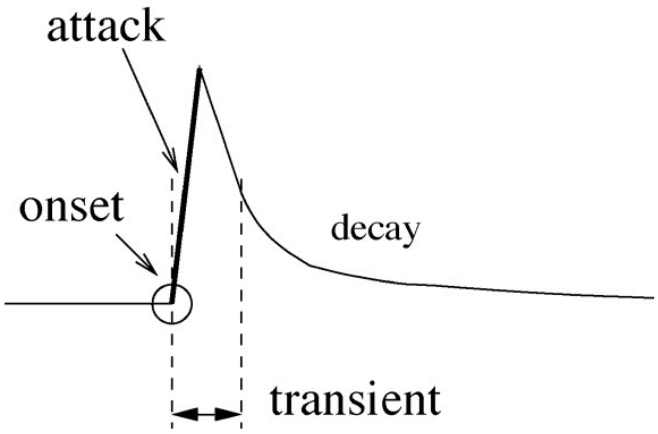

Figure 2.2: Illustration of the ideal case of a single note onset. Adapted from: [2].

Fig. 2.2 shows a simple case of an isolated note, where it is clear the difference between "onset", "attack" and "transient". The onset of a note is the exact instant in which the musical event starts and coincides with the beginning of the transient. The attack is the time interval during which we observe an increase in the amplitude envelope. The transient is the period during which the attack occurs and the abrupt decrease right after it [2].

Automatic onset detection applications are usually divided into online detectors, used for real-time practices in a live performance for example, and offline detectors, that are the majority of detectors, applied in digital audio workstations, or within a processing pipeline of other MIR tasks.

Fig. 2.3 illustrates the basic procedure employed in the majority the of onset detection algorithms that are divided into three parts: i) signal pre-processing, that aims to emphasise relevant parts of the signal; ii) reduction, where the onset detection function is computed, a function whose peaks indicate probable note onsets and iii) peak detection, where most offline methods use averaging over (past and future) time to compute dynamic thresholds for the final extraction of onset times [2, 24, 23, 25].

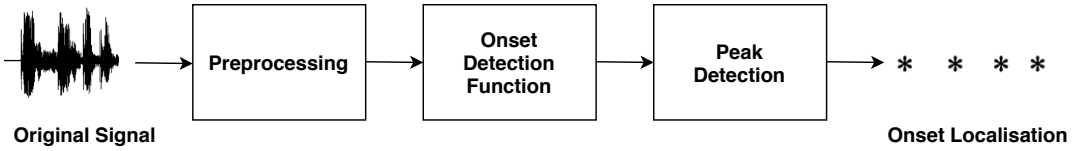

Figure 2.3: Common onset detection workflow.

Moreover, the basis of the onset detection approaches presented in the next section use either the phase (phase deviation), energy (spectral flux), or a combination of both, of the signals, in their methods. More recent approaches rely on machine learning techniques. Generally, these have higher computational costs and depend on large datasets for training and testing.

2.4.2 Previous Work

Earlier onset detection methods relied on the bursts of energy that percussive sounds represent at the beginning of each event. The detection function follows the amplitude envelope of the signal that reflects an increase of the signal's amplitude at each onset occurrence. Schloss in [26] used the energy contour of the waveform to find the attacks of percussive sounds, with an energy envelope follower as in Eq. 2.1:

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

where $w(m)$ is a smoothing window to evaluate the average energy over the window of width N. However, and despite efficient results of these methods when strong percussive transients are predominant in a noiseless background, it fails for non-percussive signals and when transient energy overlaps in complex mixtures [27].

To reflect changes in the spectral structure of the signal, several onset detection functions have found it useful to represent the signal over both time and frequency. Time-frequency (TF) representations can be obtained using several frequency bands or using short-time Fourier transforms [2].

In [28], Goto and Muraoka published a beat tracking system that makes use of filter banks to analyse transients across frequencies. They presented a multiple-agent architecture to detect rhythmic patterns that allows multiple examinations of beat positions in parallel. The system slices the spectrogram into spectrum strips and recognises onsets by detecting sudden changes in energy.

Masri [29] proposed, in his Ph.D. Thesis, the High-Frequency Content (HFC) function that emphasises energy bursts which occurs in the higher part of the spectrum. The mechanism produces sharp peaks during attack transients, linearly weighting each bin's contribution in proportion to its frequency as in Eq. 2.2:

$$HFC = \sum_{k=1}^{n} \left(k \left| X_k(n) \right|^2 \right) \tag{2.2}$$

where $X_k(n)$ is the kth bin of the STFT taken at time n. Detection results are notably robust when faced with percussive onsets. However, the function is less successful at identifying non-percussive onsets that do not present such wide-band bursts like bowed strings or wind instruments.

Other methods measure the changes in the harmonic content of the signal, the spectral difference or spectral flux, that calculate the detection function based on the difference between the spectral magnitude of two successive STFT frames as formulated in Eq. 2.3 [2, 30].

$$SD(n) = \sum_{k=-N/2}^{N/2-1} \left\{ H\left(|X_k(n)| - |X_k(n-1)| \right) \right\}^2 \tag{2.3}$$

where $H(x) = (x + |x|)/2$, i.e., zero for negative arguments [2].

This approach quantifies the amount of change found from one frame to another, rather than frame-by-frame measurements. On one hand, Masri [29] uses the L_1-norm of the difference between magnitude spectra, on the other, Duxbury et. al [31] use the L_2-norm on the rectified difference, that affects only those frequencies where there is an increase in energy and is intended to emphasise onsets rather than offsets.

Bello and Sandler [32] tested a different approach that measures the temporal instability of phase. Tonal and percussive onsets are identified by deviations from an expected phase based on the underlying assumption of constant frequency. The phase deviation for each bin is quantified as Eq. 2.4:

$$\phi'_k(n) = princarg\left(\frac{d^2}{dn^2} \left(\phi_k(n) \right) \right) \tag{2.4}$$

where *princarg* maps the phase to the $[-\pi, \pi]$ range. The onset detection function is generated as in Eq. 2.5:

$$PD_\phi(n) = \sum_{k=0}^{N} \left| \phi'_k(n) \right| \tag{2.5}$$

During the steady-state part of the signal, it is expected that phase deviations tend to zero, thus the distribution is strongly peaked around this value. With this in mind, the phase delay (angular speed in the unit circle) is assumed to be constant, and without acceleration. During attack transients, values increase, widening and flattening the distribution measured by the inter-quartile range and the kurtosis of the distribution. A drawback of this function is that important phase changes may also occur at places not related to a musical change: noisy components of the signal will usually present an unstable phase. Although this may not affect tonal events with strong harmonic components, large variations may occur as the signal becomes more percussive and noisy [2, 32].

In addition, Bello et al. in [27] introduced a study on the combined use of energy and phase information for the detection of onsets in musical signals. The proposed approach works with Fourier coefficients in the complex domain, where both phase and amplitude information work together, offering a generally more robust onset detection scheme. The stationarity of the spectral bin is quantified by calculating the Euclidean distance between the observed $X_k(n)$ and that predicted by the previous frames, $X'_k(n)$. These distances are summed across the frequency-domain to generate an onset detection function that is sharp at the position of onsets and smooth everywhere else [27, 2]. This approach proved itself effective for a large range of audio signals while remaining computationally cheap and because of this, suited for online purposes.

Recent approaches have focused on the use of machine learning techniques for automatic onset detection. In general, these methods tend to be more robust when applied to different types of music with the vast majority working with the magnitude spectra of signals [33, 34, 23]. Since most of these methods are purely data-driven, they become highly dependent on large datasets for training and are in general computationally more demanding, which makes them unsuited for online processing [25].

Eyben et al. [23] present a universal onset detector that is claimed to work in all kinds of music, including complex music mixes. This method is based on auditory spectral features and relative spectral differences processed by a bidirectional Long Short-Term Memory recurrent neural network, which acts as a data reduction function. The authors relied on a large database of onset data that covers various genres and onset types to train the network. Broadly, this system transforms audio data to the frequency domain via two parallel STFTs with different window sizes (W=1024 and W=2048 samples), obtaining the magnitude spectra and their first-order differences to feed as inputs to the bidirectional Long Short-Term Memory (BLSTM) network, which produces an onset activation function at its output. The onsets are represented by the local maxima of the onset detection function, defined as the points where activation values are greater than the threshold, which can have a constrained range of variation between 0.1 to 0.3.

Böck et al. [24] extended the system proposed in [23] and modified it in order to enable the system to work in real-time online scenarios without delay. Because bidirectional neural networks violate causality (they average over past and future time to compute dynamic thresholds), they are not suitable for online purposes and thus was replaced by a unidirectional one. Also, the Long Short-Term Memory (LSTM) units used in the hidden layer were replaced by standard units

with a softmax activation function to allow the performance on shorter temporal context, in the case of online detection, limited to a few frames. The modifications reduced the computational complexity of the system and made it suitable for real-time processing. It achieves performance close to offline onset detection algorithms while introducing zero delay between the audio signal and the reporting of an onset.

2.4.3 Evaluation in the Context of this Work

For evaluation, we use the generic F-measure (F1) score for onset estimation with varying tolerance windows. The estimations will be compared against ground-truth data.

In [2, 35], an onset is considered as correct if it is detected within a 100 ms window around ($\pm$ 50 ms) the annotated ground truth onset position. In [36], the *mir_eval* metrics use a tolerance window of $\pm$70 ms. In this work, we align with a smaller tolerance window of $\pm$25 ms as in [37, 23, 24].

The F-measure, F in Eq. 2.8, is calculated using two quantities, precision, p in Eq. 2.6, and recall, r in Eq. 2.7. Precision indicates the proportion of the onsets that are correct, i.e. those that fall within the tolerance window around the annotation. Recall quantifies the proportion of the total number of correct onsets that were identified.

$$p = \frac{c}{c + f^{+}} \tag{2.6}$$

$$r = \frac{c}{c + f^{-}} \tag{2.7}$$

$$F = \frac{2pr}{p + r} = \frac{2c}{2c + f^{+} + f^{-}} \tag{2.8}$$

where c is the number of correct detections, f^{+} is the number of false positives and f^{-} is the number of false negatives.

2.5 Music Source Separation

2.5.1 Introduction

Source separation aims to separate a set of source signals from a set of mixed signals with little to no information about the sources or the mixing process [38]. Under this broad category are audio source separation, which operates with audio signals, where it groups speech separation (cocktail party problem [39]), sound separation (separation of natural sound sources from our daily life, e.g., people walking, birds singing or cars passing) and music source separation.

Music source separation is the task of decomposing music mixtures into its constitutive components, as illustrated in Figure 2.4, by estimating each source instrument or group of instruments

that share a similar timbre, e.g. a group of trombones may be perceived as a single source in the mixture, depending on the applied algorithm [40]. The problem of source separation would not arise if all multi-channel studio recordings were available. In practice, the vast majority of published music is distributed in stereo format, i.e., with just two channels. Furthermore, in many older recordings, many instruments were recorded into one channel due to hardware limitations. Given this fact, MSS unfolds into several challenges that have been tackled over the years [40], from lead and accompaniment separation [41], singing voice separation [42] to music remixing and upmixing [43], to name just a few.

Due to the particularities of each one of these problems, the separation approach must be designed taking into account the nature of the signals, the mixing system of the input, the recording environment and the amount of prior information available about the sources (scores, hours of instruments recordings, etc).

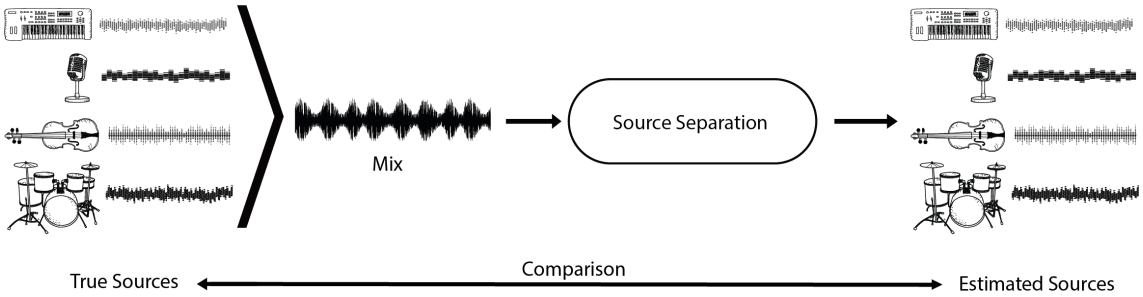

Figure 2.4: Music source separation illustration.

2.5.1.1 Signal Model and Representation

Multi-channel signals consist of several waveforms, captured by more than one microphone. The mixing process presents various nuances that directly or indirectly affect the way sources are combined such as room reverberation or microphone leakage [44].

Considering a linear mixing model (only amplitude scaling before mixing), the representation of the mixture signal in the time domain *x(t)* can be expressed as a weighted sum of the sources $s(t)$ [45] as formulated in Equation 2.9:

$$x_i(t) = \sum_{j=1}^{J} m_{i,j} \cdot s_j(t) \tag{2.9}$$

where $i = 1, ..., I$, given that I represents the total number of channels, J is the total number of sources and m represents the scalar amplitude (gain) associated with each channel. It is important to note that Equation 2.9 does not take into consideration any residual component originating from noise.

The vast majority of recordings are performed in reverberant locations and in such environments the mixture definition in Equation 2.9 is better formulated with a convolution between the

sources $s(t)$ and the mixing matrix $m_{i,j}$ holding the impulse responses corresponding to each pair of source and microphone as shown in the Equation 2.10 [45]:

$$x_i(t) = \sum_{j=1}^{J} (m_{i,j} * s_{i,j}(t)) \tag{2.10}$$

where $*$ designates convolution.

The constraint of this convolution mixing model is that each source must be located at a precise point in space corresponding to an impulse response, which makes it impossible for spatially spread sources. Vincent et al. presented in [46] an alternative, described in Equation 2.11, of the convolution mixing model that determines the contribution of each source in each channel:

$$x_i(t) = \sum_{j=1}^{J} (\sum_{j=1}^{J} (m_{i,j} * s_{i,j}(t))) \tag{2.11}$$

In most MSS systems, the first step is to transfer the input time-domain signal to a real-valued time-frequency representation [47]. The reason is that, in this type of representation, sources tend to be less overlapped when compared to the time-domain waveform [48]. The TF representation that most research in MSS has focused on is the short-time Fourier transform (STFT) [49]. Like other representations, it also outputs the spectrogram that is the magnitude (or power) of the STFT, a time-varying representation of the spectral content of a real audio signal. When a mixture is multi-channel, each channel has its own TF representation, leading to a three-dimensional array: frequency, time, and channel.

2.5.1.2 Source Filtering

In many models, the separated sources are obtained by filtering the mixture, usually in a final post-processing stage. Typically, this is done by using a TF masking approach, where each source estimate is obtained by multiplying the TF representation of the mixture with a time-varying filter (TF mask) that changes every few milliseconds. Loosely speaking, this TF mask can be understood as an equalizer with periodic settings changes that select which frequencies are attenuated and which ones are not. The most common form of designing such masks is the generalized Wiener filter (GWF) [50].

2.5.2 Previous Work

One of the first used techniques in MSS was the Independent Component Analysis (ICA) that exploits the spatial position of the sources. ICA is a computational signal processing method that aims to separate a set of multi-channel mixed signals into a set of additive source signals or independent components [51].

This approach relies on the statistical independence of the sources and their *non-Gaussian* distribution to estimate a demixing matrix for the mixture signal [52]. ICA requires the mixture to be determined, i.e., containing the same number of channels as music sources. However, usually,

this is not the case for music signals, where the vast majority contain more music sources than channels, i.e., they are *under determined*. Also, ICA algorithms performance decreases when reverberation increases, since demixing filters become harder to estimate [53].

Non-negative Matrix Factorisation (NMF) is a spectrogram factorisation technique that estimates the input spectrogram as a linear product of two matrices as in Equation 2.12, first proposed by Lee and Seung [54]:

$$M \approx W \cdot H \tag{2.12}$$

where M is the magnitude spectrogram of the mixture signal. W is a matrix of basis vectors, which is a dictionary of spectral templates modelling spectral characteristics of the sources, and a matrix of weights, time gains, or activation H, both restricted to be non-negative. W can be associated and learned for instrument pitches and H can yield the estimation of their temporal activation. Here, NMF improves source separation by using musical context-specific knowledge.

The dimension of the input matrix M is n x m, and the dimensions of approximation are n x r for matrix W and r x m for matrix H, where r displays the rank of the factorisation. An example NMF decomposition is shown in Figure 2.5:

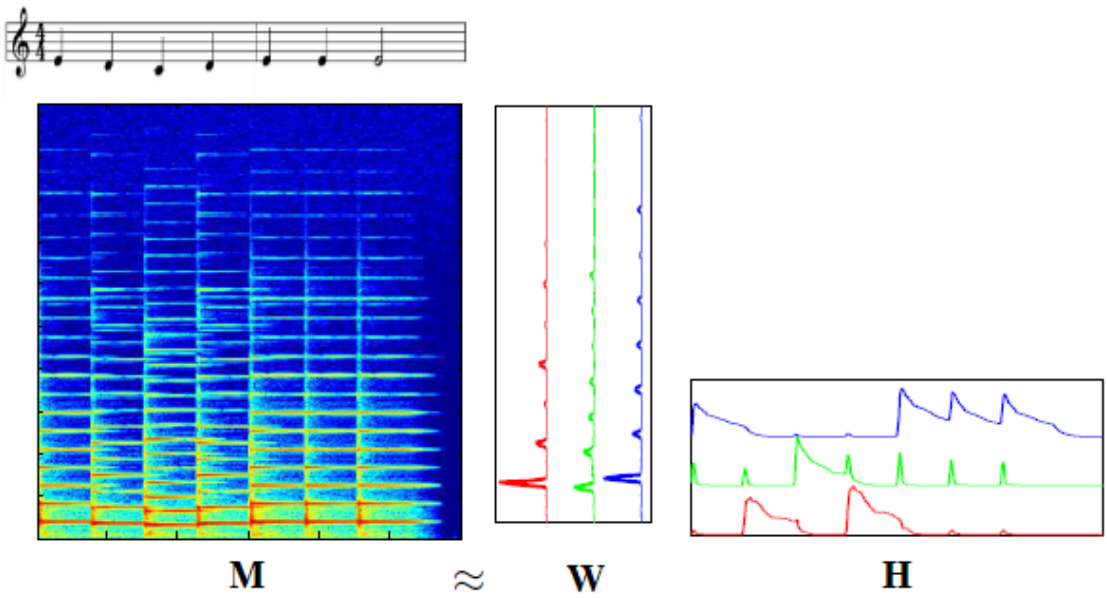

Figure 2.5: NMF example with musical score (on top) of children's song "Mary Had a Little Lamb" where M is the recording spectrogram, W the matrix of basis vectors and H, the activation matrix [3].

Here we can notice that W captures the harmonic content and H saves the time onsets and gains of each individual note. Through Figure 2.5 we can decipher that either the short-time segments of the mixture of M can be seen as columns and these approximate as a weighted sum of basis vectors or that the mixture matrix M is approximated as a sum of matrix layers [55].

In order to quantify the quality of the approximation, Lee and Seung [56] defined two possible cost functions. One, displayed in Equation 2.13, relies on the calculation of the divergence D (reconstruction error) between M and M_{approx} (= $W \cdot H$).

$$D(M \parallel M_{approx}) = \sum_{ij} \left(M_{ij} \, log \, \frac{M_{ij}}{M_{approx}} - M_{ij} + M_{approx_{ij}} \right) \tag{2.13}$$

where the divergence D reduces to the Kullback-Leibler divergence (KL) when $\sum_{ij}(M_{ij}) = \sum_{ij}\left(M_{approx_{ij}}\right) = 1$ [56].

Associated with this cost function is an update rule theorem, know as *multiplicative update rule*, for H (Equation 2.14) and W (Equation 2.15) [56]:

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{\sum_i W_{i\alpha} V_{i\mu} \, / \, (W\,H)_{i\mu}}{\sum_k W_{\kappa\alpha}} \tag{2.14}$$

$$W_{i\alpha} \leftarrow W_{i\alpha} \frac{\sum_\mu H_{\alpha\mu} V_{i\mu} \, / \, (W\,H)_{i\mu}}{\sum_\upsilon W_{\alpha\upsilon}} \tag{2.15}$$

where the multiplicative gain is formed to assure non-negativity in updating parameters $H_{\alpha\mu}$ and $W_{i\alpha}$. Helén et al. presented a work on NMF separation of drums from polyphonic music [57] using NMF to minimise the KL divergence, with a random initialisation of W and H.

Dittmar et al. investigated a method for real-time transcription and separation of drum recordings based on NMF decomposition [58]. It is based on an NMF decomposition initialised with prior spectral basis templates for the expected drums and therefore assumes that the isolated drum sounds are available for training. In this approach, the authors introduce a modification imposing semi-adaptive behaviour on W during the NMF iterations that reverts a prior basis vector W_p to the NMF decomposition of every individual frame [58].

NMF can be reformulated under a probabilistic model with the non-negativity constraints as in [59], which introduces Probabilistic Latent Component Analysis (PLCA) as formulated in Equation 2.16:

$$P(x) = \sum_z P(z) \, \Pi_{j=1}^{N} P(x_j \,|\, z) \tag{2.16}$$

where $P(x)$ is an N-dimensional distribution of the random variable $x = x1, x2, ..., x_N$. The z is a latent variable, and the $P(x_j|z)$ are one dimensional distributions. This model proposes a solution to discover the probability distribution by estimating both $P(x_j|z)$ and $P(z)$ from an observed $P(x)$ similarly to the NMF structure. PLCA aims to maximise independence, forcing the results to be positive. It is an Expectation-Maximisation algorithm that models a mixture of independent distributions, each considered an independent component latent with the data [59]. This approach, when compared with standard NMF has a lower rank of factorisation that enables dealing with more complex mixtures like multi-track recordings in a simplified way.

Smaragdis published another important work where he presents an extension to the Non-Negative Matrix Factorization algorithm which is capable of identifying components with tem-

poral structure [60]. Here, Smaragdis proposes an extension of Equation 2.12 to Equation 2.17 naming it Non-Negative Matrix Factor Deconvolution (NMFD):

$$M \approx \sum_{t=0}^{T-1} W_t \cdot H^{t\rightarrow} \tag{2.17}$$

where the matrix M is the input that is going to be decomposed, W and H are the basis vectors and weights matrices. The three matrices maintain their non-negativity [60]. The $(\cdot)^{i\rightarrow}$ operator shifts the columns of its argument by i spots to the right. This approach enables the extraction of more expressive basis functions.

Recent strategies in MSS focus on the use of deep neural networks (DNN). DNN methods take advantage of supervised learning grounded in large datasets that include mixture and the isolated signals from various recordings. They operate by training the input parameters of nonlinear functions to minimise the reconstruction error of the output. The inputs are the magnitude spectrograms of the audio mixtures whereas the outputs are either the magnitude spectrograms of each of the sources or their separating masks [61, 62, 42, 63].

The major aspect of the training of a DNN is the type and quantity of available data. Training such methods require enormous amounts of data that try to cover a high number of possibilities heard in real recordings so that the trained model can be as general as possible and able to infer the underlying spectral characteristics of the musical sources. In reality, the individual recordings for each instrument are typically unavailable for the general public, meaning that smaller datasets with freely available multi-track stems are often used for training of these algorithms [64, 65].

$Open-Unmix$ [66] is based on the *bidirectional LSTM* model from Uhlich et al. [61] and *Spleeter* [67] is built with $U-net$ pre-trained models that follow the convolutional neural network (CNN) architecture from Jansson et al. [42]. These are two of the state-of-the-art algorithms that deliver the best separation so far. The employed DL techniques are both data-driven where, instead of making any assumptions about the music mixture, the models can be learned from a large and representative database of examples.

2.5.3 Evaluation in the Context of this Work

Regarding the quality of the separated sounds, MSS tasks fall into two categories [68]:

- Audio Quality Oriented (AQO): AQO applications aim to fully unmix any given mixture at the highest possible quality so that the extracted sources can be listened to straight after separation or after some post-processing audio effects [68]. It can be applied in tools for unmixing, remixing and upmixing, hearing aids, or post-production.

- Significance Oriented (SO): SO applications aim for the extracted sources and/or mixing parameters to obtain sufficient quality to facilitate the semantic analysis of complex signals. In this case, the obtained information will serve at more abstract levels, to find a representation of the observations related to human perception [68]. It can be applied in tools for MIR, polyphonic transcription, and object-based audio coding to name but a few.

In the context of this work, the focus will be on the significance oriented separation since obtaining the highest possible separation quality is not our goal here. We aim to obtain sufficient separation quality to enable the automatic onset detection models to run on the separated signals to assess if the observation of microtiming patterns, through the outputted onsets, is possible.

The separated sources will be evaluated indirectly. The standard performance measures like Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), Signal to Artifact Ratio (SAR), and others will be left out since the nature of the original signals and the separated ones are fundamentally distinct. The original signals are obtained via contact microphones that capture mechanical vibrations induced in the instruments. Although we can perceive them as audio waveforms, these signals have not propagated through the air and thus have no reverberation. The separated signals derive from stereo mixtures that were obtained with traditional electro-acoustic microphones. For this reason, they cannot be compared in an objective manner as would be required for the calculation of standard source separation metrics.

Chapter 3

Maracatu de Baque Solto

3.1 Introduction

"Maracatu de Baque Solto", also known as "Maracatu Rural", is a performance of music, dance, and poetry that takes place during the Carnival season in Zona da Mata Norte region, in the interior of the state of Pernambuco (Northeast Brazil), that can gather up to 200 people.

According to the symbolic, religious and emotional meanings of *Maracatu*, the ritual helps to prevent spiritual and physical attacks that could result in illness and even death. *Maracatu* musicians and dancers struggle to achieve a high level of group cohesion locally known as *consonância* (consonance), a valued human and aesthetic feature which is opposed to *desmantelo* (fracture, breaking up). It is not yet clear how this consonance is achieved through music, although these concepts point to subtle manners of producing rhythmically interactions during the performances.

A Maracatu performance is a neighbourhood party, a poetic competition, a cult of ancestors, a ritual of protection of the community of the inhabitants of Zona da Mata Norte region. Maracatu percussive music is highly repetitive, and is played as loud and as fast as possible, ranging from 160 to 180 BPM [10].

The most ancient *maracatus* were rural workers, cane workers, sugarcane cutters, between the end of the 19th and early 20th century. It has been part of the Brazilian National Historical and Artistic Heritage Institute [1] in the Register Book of Forms of Expression since December 2014.

In December 2019, the Maracatu de Baque Solto group "Leão de Ouro de Condado" was invited to Lisbon in the context of a FCT-funded project titled "The Healing and Emotional Power of Music and Dance" [8, 9]. This was the first time that these musicians flew out of Brazil to share their culture overseas. The recordings were made on two different occasions: first, a parade through the city of Lisbon followed by live performance, and second, in the motion capture laboratory located at the Faculty of Human Kinetics of the University of Lisbon. An example of the Maracatu de Baque Solto parade in Lisbon is shown in Figure 3.1.

[1] http://portal.iphan.gov.br/pagina/detalhes/505/

Figure 3.1: Maracatu de Baque Solto “Leão de Ouro de Condado” during a parade in the center of Lisbon, Portugal. Photo credit: Filippo Bonini Baraldi.

3.2 Instruments

In Maracatu, the five percussion instruments of the group are named as follows [10]:

- *Tarol* - similar to a small snare drum, but thinner. Illustrated in Fig. 3.2.
- *Porca* - a friction drum, played with a damp cloth holding the stick, also know as *Cuíca*. Illustrated in Fig. 3.3 and Fig. 3.4.
- *Mineiro* - a metal tube filled with beads or other small objects, which is shaken to create a rattle type sound. Illustrated in Fig. 3.5.
- *Bombo* - a bass drum like instrument played with two sticks, one for each side. We refer to Bombo High as the upper skin, and Bombo Low as the lower. Illustrated in Fig. 3.6.
- *Gonguê* - an iron instrument comprised of two bells of different pitches. We refer to Gonguê High as the higher, and Gonguê Low as the lower pitched bell. Illustrated in Fig. 3.7.

Figure 3.2: Tarol

Figure 3.3: Porca

Figure 3.4: Back of the Porca

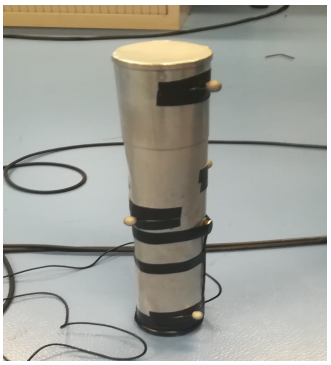

Figure 3.5: Mineiro

Figure 3.6: Bombo

Figure 3.7: Gonguê

3.3 Data Acquisition

In the context of this project, the aim is to analyse each percussionist's performance in isolation. The recording process was designed in a way that leakage effects would be eliminated (or at least heavily minimised) in order to ease the analysis and annotation processes.

Although other attempts of multiple eletro-acoustic microphone setups have successfully captured isolated audio for similar microtiming analysis [69], the physical arrangement of the Maracatu percussionists in a tight circle, together with their very loud playing style, makes this approach impractical and highly prone to leakage. The proposed alternative was to use contact microphones that converted the vibrational energy of the drum surfaces into electrical energy.

For this study, aligned with the work in [10], we used the Schertler Basik Set universal contact microphone. This particular set includes a phantom-power adaptor box, with a 1.5 V AA battery, which delivers a line-level signal along an unbalanced 1/4" jack connector, and a frequency response ranging from 60 Hz to 15 kHz. The sensitivity on the instrument is -34 dB (time-averaged sound level). We found these microphones could provide high-quality audio recordings with minimal spillage and distortion.

For the microphone placement, we sought to balance the optimum location for sound capture on each instrument in terms of its physical properties, while minimizing any impact to each musi-

cian's playing style. Given the small size of the Schertler pickup (less than 1 cm in diameter), this aspect was relatively straightforward.

Concerning the number of pickups used per instrument, we placed two on the Gonguê – one per bell, and two on the Bombo – one on each skin of the drum. For the remaining instruments, the Tarol, Porca, and Mineiro we used a single pickup. For the first two instruments, we placed one microphone on each surface for capturing pitch nuances. On the other hand, instruments such as Porca, Tarol, and Mineiro afford one pitch-only. For these instruments, we placed one contact microphone in a sweet-spot, ensuring optimal frequency response, time-average sound levels and musician's play ability.

The contact microphones were individually connected to a Motu UltraLite-mk3 Hybrid (USB/ Firewire) with nominal gain at the input and no further processing. The recording session consisted of discrete, synchronised tracks, one for each microphone, recorded in a Pro Tools 2019 mixing session. For the input settings, we configured the system to record at 44kHz sampling rate with 16-bit depth.

The stereo mixtures were obtained by exporting the audio from the video recordings of the live performances of the Maracatu group. The video recordings were made with the Sony Handycam HDR-CX160 camera with 5.1 surround audio capture which were mixed-down to stereo using the default settings of the open source video encoding and decoding utility FFMpeg.

3.4 Summary

The dataset consists of the stereo mixture signal and the isolated signals of each considered instrument of 34 pieces of a "Maracatu de Baque Solto" performance, that totals approximately 22 minutes. Across 7 channels, this led to a total of 238 acquired contact microphone signals having minimum and maximum lengths of 24.9 s and 123.3 s (2.055 min), respectively.

Chapter 4

Proposed Approach

In this chapter, we formulate the problem that this dissertation addresses and we detail the proposed solution and its constituent steps, illustrated in a flowchart.

4.1 Problem Formulation

The problem consists of undertaking a critical evaluation of automatic signal segmentation for microtiming analysis purposes, in the case of this work, microtiming deviations present in "Maracatu de Baque Solto", a highly percussive music style.

In this work, we will contrast automatic onset estimation capabilities in three different signal scenarios: i) the use of perfectly separated multi-track stems obtained via contact microphones. ii) stereo or mono mixed recordings; and iii) separated sources obtained via state of the art musical audio source separation techniques.

The small deviations present in the microtiming phenomenon require a rigid definition of what a considered correct onset estimation is and thus, the tolerance window around the estimation must not be too wide that the microtiming vanishes from the rhythmic visualisations. On this basis, we align with Böck et al. in [70], stating that the detection of percussive onsets is not an entirely solved problem, in particular, when dealing with unusual instrument waveforms with long attack times, such as the Porca, or unclear onset locations, such as the Mineiro, as we'll see in the next chapter.

4.2 Proposed Solution

A flowchart of the proposed solution is shown in Fig. 4.1.

The first step is dedicated to recording a dataset with the type of music that we'll work with, detailed in Chapter 3, with the isolated instrument signals and the corresponding stereo mixture.

The second step consists of manually annotating the collected dataset for the isolated instrument signals from which we'll derive our ground-truth data, indispensable for evaluating the quality of the different automatic onset estimations scenarios.

The third step comprises the section in which we obtain the separated sources of each considered instrument from the stereo mixtures using the PLCA implementation described in [59]. This step will help to understand how far away are the state of the art music source separation algorithms if one wants to start a micro-rhythmic visualisation from a mixture signal. While in Chapter 2 we discuss the prominence of deep neural networks for musical source separation, in this work we focus on the separation of different percussion instruments from one another, and largely in the absence of other musical instruments. To this end, we do not believe that systems such as Spleeter and Open-Unmix would be appropriate. Notwithstanding the fact that, the types of musical instrument sounds here are unlikely to have formed part of the training data when these networks were trained.

The fourth step groups the three different signal scenarios that we are working with and, in it, we perform several automatic onset estimation approaches with adaptations for each type of signal. Throughout this step we evaluate, in a qualitative way, the preliminary outcome of each approach to motivate the following adjusted approach.

The fifth step evaluates, in a quantitative way, the performance of the automatic onset estimation approaches applied to the three different signal scenarios. We look into the mean and distribution F-measure values in function of tolerance window and to the ratio of true positives, false negatives and false positives to conclude about the over or under detection of each approach.

The sixth step contains the microtiming visualisations obtained with the manual onset annotations and the estimations of the onsets obtained from the automatic detectors.

The seventh step compares the microtiming visualisations and, finally, in the eighth step we draw the major conclusions from the designed solution.

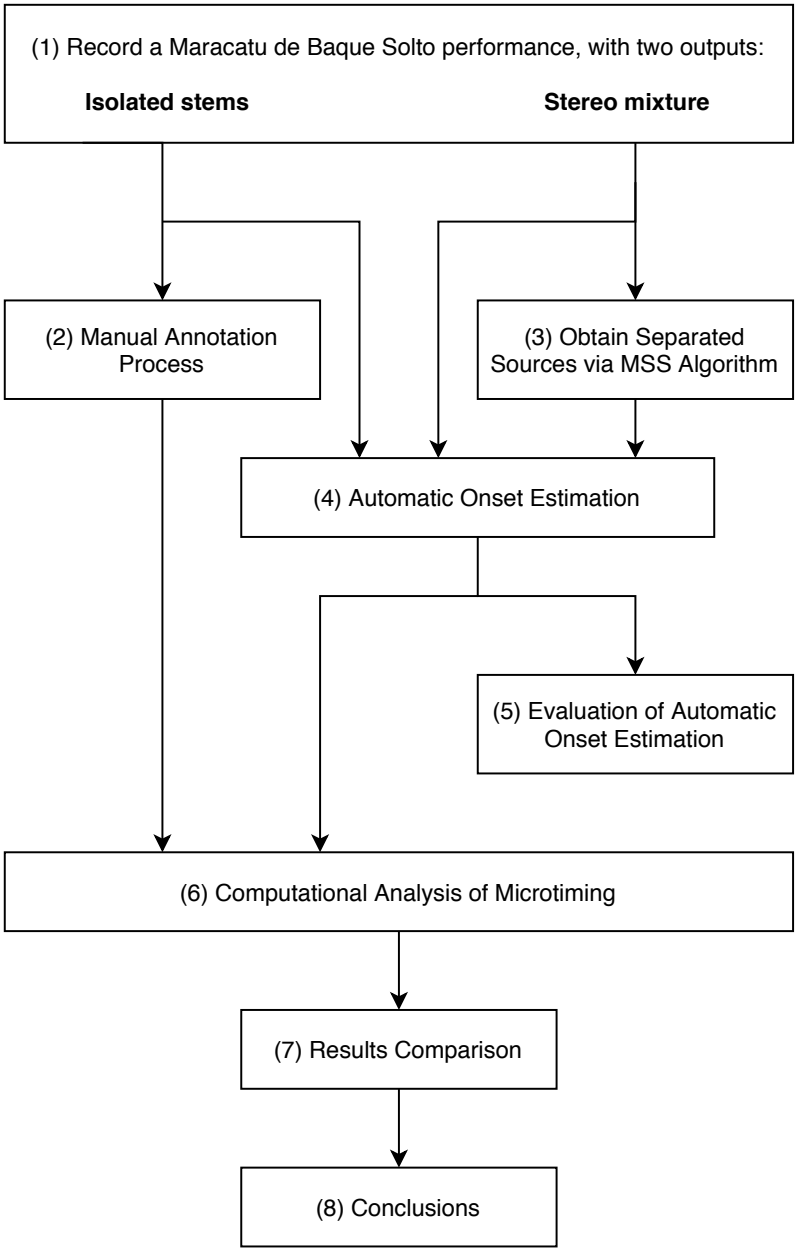

Figure 4.1: Flowchart of the proposed solution.

Chapter 5

Automatic Onset Estimation

5.1 Introduction

In this chapter, we disclose the adopted strategies for automatic onset detection applied to Brazilian Maracatu de Baque Solto and evaluate their capabilities in different signal scenarios. We aim to discover whether each approach delivers good enough results that enable microtiming visualisation.

Throughout this chapter we elaborate on the contribution of annotating the newly acquired dataset, we evaluate the retraining of a deep neural network onset detection algorithm specifically for this genre and evaluate its performance in three different scenarios: when applied to isolated signals, mixture signals and separated signals.

5.2 Manual Annotation Process

The process of manually annotating a musical dataset is a thorough and time-consuming task. It consists of manually label onset locations for the entire collection of sound files using a standard sound editor, like Sonic Visualiser [71], in the case of this work. First, onsets were identified using the visualisation capabilities of Sonic Visualiser by plotting spectrograms with different STFT lengths to precisely capture the timing of the percussive event. Secondly, a perceptive evaluation was done by playing back a slowed version of each track with the overlaid annotations.

Throughout the process, we found that the recordings of the Mineiro were very challenging to annotate in a precise and consistent way due to the unusual waveform shapes, and for this reason, we chose not to include them for our analysis. Ultimately, we selected four instruments: two instruments with time-keeping roles (Porca and Gonguê Low) and two rhythmically expressive instruments in which to observe microtiming (Tarol and Bombo High) [72]. An overview of the number of onsets labelled for the considered instruments in all 34 tracks is shown in table 5.1.

To illustrate the nature of the signals that we're working on, as well as providing some indication of the speed at which the percussive events occur in Maracatu music, we point out that 39

Table 5.1: Number of onsets annotated per considered instrument for entire dataset.

Instrument	**# Onsets**
TAROL	20511
PORCA	5072
BOMBO HIGH	14781
GONGUÊ LOW	5192
TOTAL	45556

onsets are labelled in just 1.6 s in the lowest plot of Fig. 5.1, that illustrates an excerpt of track #27 of our dataset.

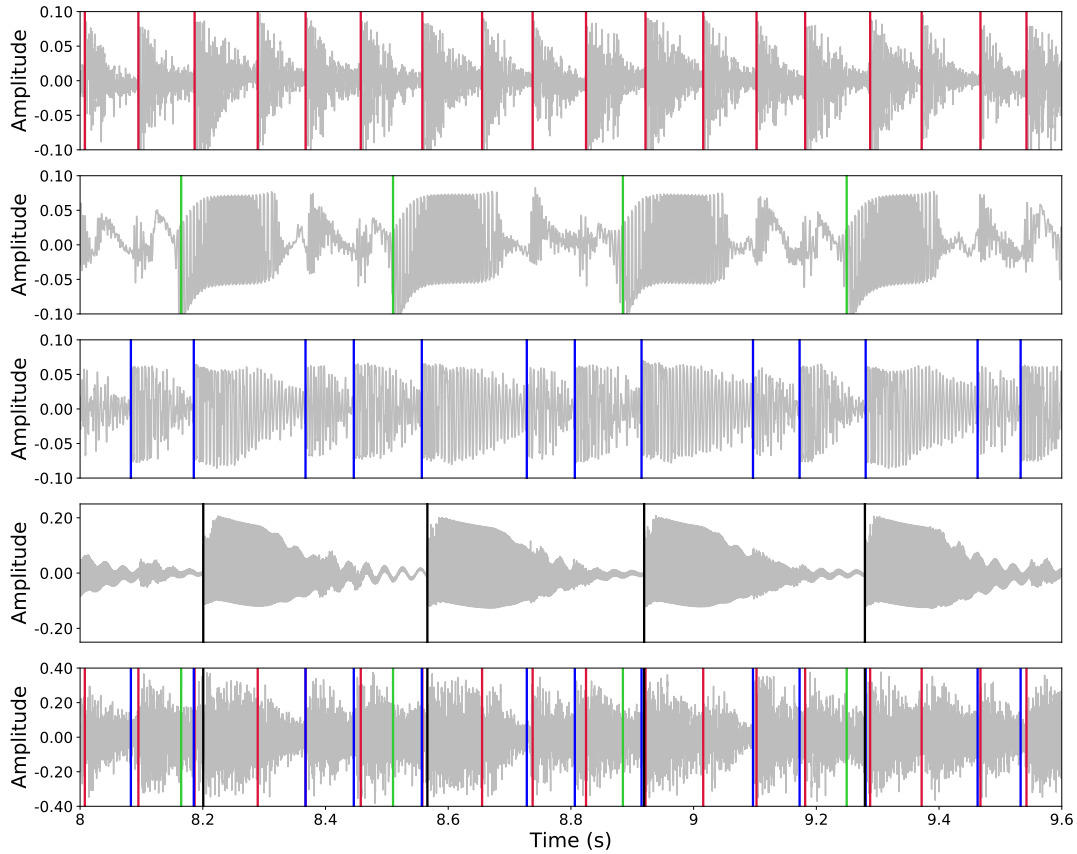

Figure 5.1: Illustration of instrument signals with manually annotated onsets in approximately one bar of track #27. Top to bottom: Tarol with annotations in red solid lines; Porca with annotations in green solid lines; Bombo High with annotations in blue solid lines; Gonguê Low with annotations in black solid lines; Audio mixture with overlaid onsets of the four instruments.

5.3 Automatic Onset Estimation Scenarios

In this section, we describe how onsets are automatically estimated through offline detection methods, on three different signal scenarios: using isolated signals from each individual instrument obtained via contact microphones; using stereo and mono mixed audio signals; and using separated signals of each instrument obtained via a music source separation solution. A quantitative evaluation of the presented approaches is shown in Section 5.4 where we contrast the qualitative evaluation made progressively throughout this section with data measurements.

5.3.1 On Contact Microphone Signals

Given our proposed signal acquisition method using contact microphones, we start our estimation process with the isolated signals. Even though the isolation between channels is very satisfactory, some leakage can still occur, as shown in Fig. 5.1, specially when we refer to the two bells of the Gonguê which are physically connected, and for which only the low bell is annotated.

5.3.1.1 Madmom Library

The first step of our process lies in obtaining the onsets of each individual instrument signal by applying a state-of-the-art onset detector method that uses deep neural networks [23]. This method is implemented in an audio signal processing library, called Madmom [73]. A more in-depth overview of this algorithm is presented in Section 2.4.

A multi-track view of an excerpt of the signals is shown in Fig. 5.2, with the resulting Madmom generated onsets overlaid.

In practice, what we found was that the unfamiliar waveform shapes of the Gonguê and Porca events created numerous problems for the onset detection system both on the temporal location of the events and on the amount of extra detections (insertions). Since this is a data-driven algorithm, we can conclude that Porca and Gonguê signals are unfamiliar to the training dataset of this algorithm. On the contrary, we can observe that its training dataset contained similar signals to the ones of Tarol and Bombo High, hence the much more promising onset estimation results.

At this point, we determined that the Madmom approach was not sufficient for us to achieve the goal of observing microtiming patterns, using fully automatic outputs. Nevertheless, we take closer look at the Madmom detections that indeed correspond to a percussive event and inspect the timing offset between these and the manual annotations (absolute error).

In Fig. 5.5, we show a zoomed-in version of signal excerpts to illustrate these unwanted deviations which directly affect the later microtiming analysis. Here, we should focus on sub-figures (a), (d), (g), and (j) that correspond to Tarol, Porca, Bombo High, and Gonguê Low, respectively.

A considerable offset is noticeable, in the four instruments, when comparing the coloured lines, that represent the estimated onsets, with the black dashed lines, that represent the position of the manual annotations. For microtiming analysis purposes, we believe that these sutble deviations

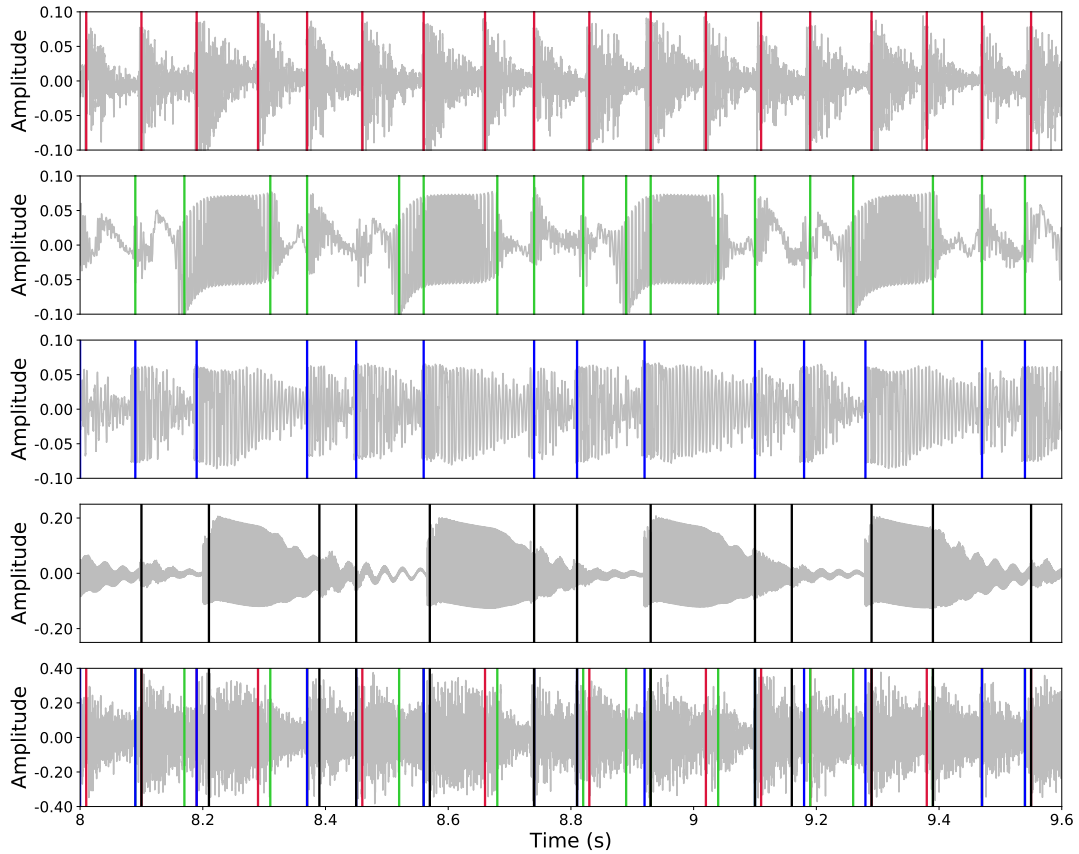

Figure 5.2: Illustration of instrument signals with estimated onsets obtained via Madmom audio signal processing library in approximately one bar of track #27. Top to bottom: Tarol with annotations in red solid lines; Porca with annotations in green solid lines; Bombo High with annotations in blue solid lines; Gonguê Low with annotations in black solid lines; Audio mixture with overlaid onsets of the four instruments.

cannot be ignored. Therefore, to improve our results both on accuracy and precision, we make use of an in-development tool designed in the context of the HELP-MD project.

5.3.1.2 Retraining of DNN Algorithm Approach

In the context of the ongoing HELP-MD project, an existing deep neural network approach was retrained on a per-instrument basis in order to model each specific instrument based on a subset of manually annotated onsets. More specifically, the temporal convolutational network (TCN) approach in [74] which was first presented for musical audio beat tracking was adapted for the task of onset detection. While it would have been possible to retrain a version of the Madmom architecture, the TCN offers the advantage that is particularly fast to train (even without GPUs) and uses proportionally far fewer weights than the recurrent neural network approach in Madmom. The retraining methodology, which we only present briefly here (since it is not yet published and is being developed by other team members of the HELP-MD project), involves first training the

network from scratch on an existing onset detection dataset [25]. Next, for each of the instruments of the Maracatu, a separate instrument-adapted network is obtained by freezing all but the shallowest layers of the network (i.e. those closest to the musical surface) and the final output layer, and retraining on a short manually annotated section. In this sense, these retrained networks are not a direct contribution of this dissertation, but were used in "black-box" manner to allow the comparison against an onset detection approach trained to operate under more general and diverse conditions. Nevertheless, we deemed it important to provide a high-level description of the approach for the purposes of understanding the remainder of the dissertation.

The main advantage of the approach is that, through the learning process, the network familiarises the onset detector function with the waveform shapes of the considered instruments and thus it should obtain more satisfactory results.

Shown in Fig. 5.3 is an excerpt of the signals of each instrument with overlaid onset estimations of the retrained network in which we can recognise a substantial improvement both in temporal location and in extra detections (zero in the presented excerpt). At a glance, it would be difficult to distinguish it from Fig. 5.1, with the manual annotations.

Regardless of this qualitative evaluation, we zoom-in again, now on this version of the signals and estimations, displayed in Fig. 5.5, sub-figures (b), (e), (h) and (k) that correspond to Tarol, Porca, Bombo High and Gonguê Low respectively, to visually inspect the absolute errors that might be present.

We observe an approximation of the coloured lines to the black dashed lines which means that onset estimation performance improved. Nevertheless, there's still some visible error associated with this approach. In the case of this excerpt, the absolute error is in the order of a 5 ms deviation from manual annotations. To minimise this gap even further, a time-correction algorithm was applied to adjust the temporal location of the estimated onsets of the retrained DNN algorithm.

5.3.1.3 Time-Correction of the Retrained DNN Models Estimations

To lessen the small errors associated with retrained DNN estimations, a time-correction algorithm was developed and applied to the output of the retrained DNN models to adjust the temporal location of each estimated onset.

The time-correction process takes as input: the signal, the corresponding annotation file, sampling rate, and a window length, *winlen*, and outputs the time-corrected onset estimations.

The algorithm is structured as follows: for each annotation location we create a large region around it with two times the size of the input *winlen*, i.e., the location of the annotation $\pm$ *winlen*. This defines the space in which the algorithm will search for a precise onset position of the drum stroke. Next, two smaller regions with the $1/4$ of the size of *winlen* will iterate through each sample within the larger region. For each sample, p, within the larger region, it will calculate the change in amplitude and spectral flux by computing the difference of each of these quantities between the smaller two regions. The two smaller regions are placed sequentially (i.e., one right after the other) having the first sample of the first small region be p. Finally, the new onset position

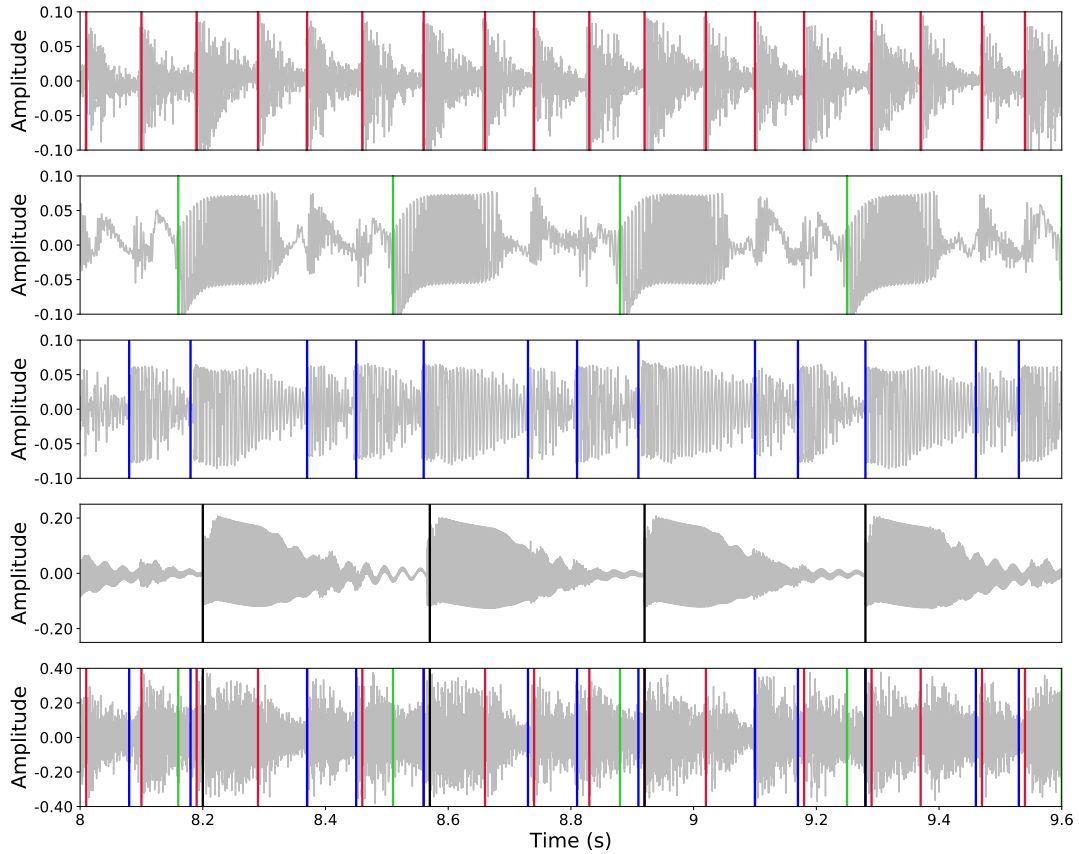

Figure 5.3: Illustration of instrument signals with estimated onsets obtained via retrained DNN algorithm in approximately one bar of track #27. Top to bottom: Tarol with annotations in red solid lines; Porca with annotations in green solid lines; Bombo High with annotations in blue solid lines; Gonguê Low with annotations in black solid lines; Audio mixture with overlaid onsets of the four instruments.

is defined at the sample where it is identified the maximum of the sum of the change in amplitude and spectral flux. The amplitude, A , of each region is computed using Eq. 5.1:

$$A(n) = \sum_{i=reg_n} (|x(n)|) \tag{5.1}$$

Where reg_n is the space defined by reg_1 and reg_2 and $|x(n)|$ is the absolute value of signal x. The amplitude change is obtained by computing the amplitude difference of reg_2 and reg_1 throughout each sample within the large region.

The spectral flux measures changes in the spectrum as a "distance" between successive short-term Fourier spectra [2], in this case, we use the Euclidean distance, as Eq. 5.2 describes:

$$SF(n) = \sqrt{\sum_{i=reg_n} \left((|X_2(i) - X_1(i)|)^2 \right)} \tag{5.2}$$

Having X_n as the magnitude spectra of each region reg_n, reg_1 and reg_2, calculated by means of the Fast-Fourier Transform (FFT).

The final step is for post-processing where from the array of collected differences in amplitude and from the array of collected differences in flux, the algorithm selects as the new onset location, the sample at which the sum of amplitude change and spectral flux is maximum, that corresponds to peak.

In essence, this algorithm is illustrated in Fig. 5.4 and it can be summarised as follows:

- Definition the large regions of interest in which to adjust the temporal location of the given estimations: location of each given annotation $\pm$ *winlen*;
- For each sample within each large region, slide two smaller regions with $1/4$ of the size of *winlen* each and calculate the amplitude difference and spectral flux between the two smaller regions;
- The adjusted onset location is found at the sample in which the sum of the amplitude difference and spectral flux is maximum plus the length of one smaller region samples.

Both the Madmom library and the DNN approach calculate at a 10ms resolution and, with signals sampled at 44.1kHz, leads to 441 sample resolution. Consequently, for the input of the algorithm we chose the closest power of two window length, that is 512 samples.

Since this mechanism adjusts very small temporal offsets, the multi-track view of the time-correction will be very similar to Fig. 5.3. However, a zoomed-in version of the signals can provide a more informative illustration and is shown in sub-figures (c), (f), (i) and (l) of Fig. 5.5, that correspond to Tarol, Porca, Bombo High and Gonguê Low, respectively.

A close analysis of Fig. 5.5 indicates that the time-correction approach narrows the absolute error between estimations and manual annotations to practically zero. While we recognise that it is difficult to precisely annotate even sharp percussive events at the audio sample level, we observed this approach was able to provide a highly consistent annotation of the onsets across multiple instances the same instrument and following a sensible methodology of looking for the point of maximal energy change and spectral flux.

5.3.2 On Mixed Audio Signals

Until this point, we showed how onsets can be estimated and corrected if we have clean isolated signals for each instrument. The best-case scenario for any evaluation of this kind.

The process of manually annotating a mixed signal is impractical to, sometimes, virtually impossible. In this section, our goal is to test the previously demonstrated approaches and apply them to mixed audio signals. Since it might not always be possible to capture signals individually, this represents a more challenging scenario. We start with stereo audio signals and observe the performance of the algorithm and also experiment with artificial mixtures where we have control of which instruments are mixed.

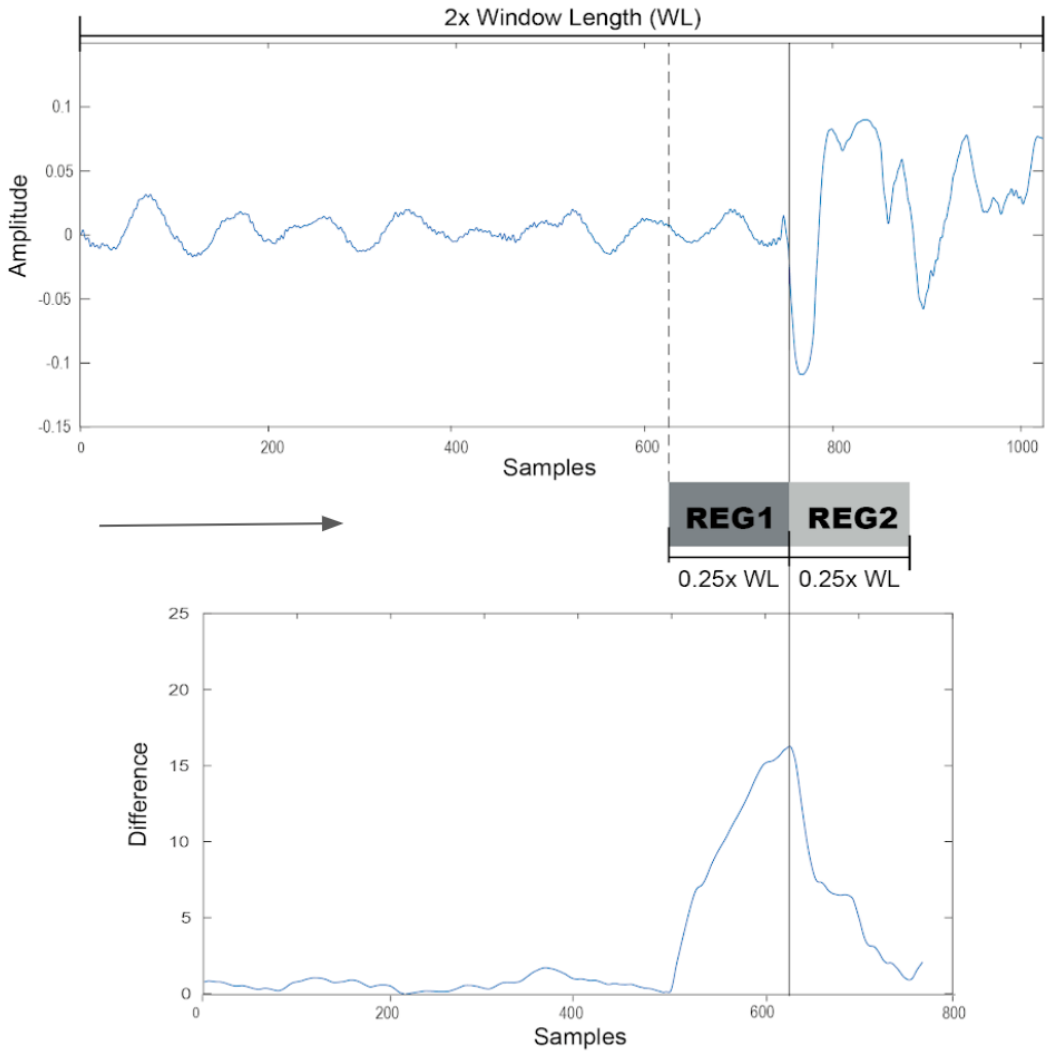

Figure 5.4: Example of the peak resulting from the maximum of the sum of amplitude change and spectral flux. Top figure illustrates the large region in the instrument signal. Bottom figure illustrates the peak of the difference between the two smaller regions, REG1 and REG2.

5.3.2.1 Stereo Mixture Signals

The goal in this section is to understand how the previous approaches perform when other sources are present in the signal and we seek to estimate the onsets of each instrument individually.

Listening back the audio recordings of our dataset we can hear, on top of the five instruments presented in Section 3.2, a whistle ("apito"), a human voice when the poet improvises short verses and a wind instrument, the trumpet. All together, they make the estimation process much more challenging.

The first distinction found from estimating in isolated signals is that the Madmom detection algorithm does distinguish the onsets from different instruments. If applied to mixed signals, the resulting output is the detection of every percussive event present in the mixture with no distinction whatsoever between instruments and so, it is not useful for our intent.

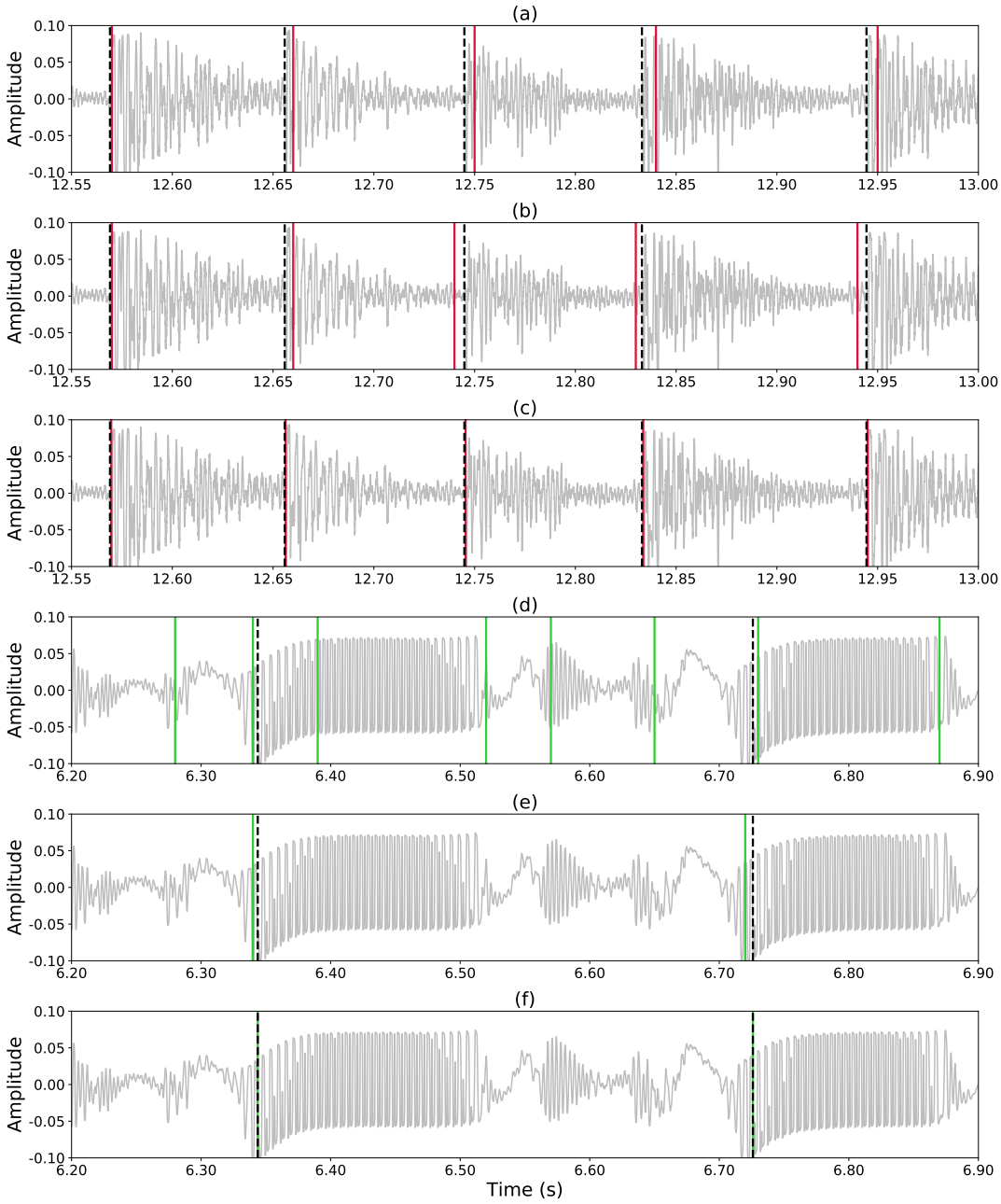

Figure 5.5: Part 1: Illustration of zoomed-in instrument signals with estimated onsets of track #27 coloured represented. Manual annotations in black dashed-line. Top to bottom: (a) Tarol with Madmom annotations in red solid lines; (b) Tarol with retrained DNN annotations in red solid lines; (c) Tarol with time-corrected annotations in red solid lines; (d) Porca with Madmom annotations in green solid lines; (e) Porca with retrained DNN annotations in green solid lines; (f) Porca with time-corrected annotations in green solid lines.

However, we can apply each of the four instrument-specific models from the retrained DNN approach to the same mixture track. For visualisation purposes, in Fig. 5.6, we exhibit the resulting onset estimations obtained from the mixture signal plotted onto the isolated signals.

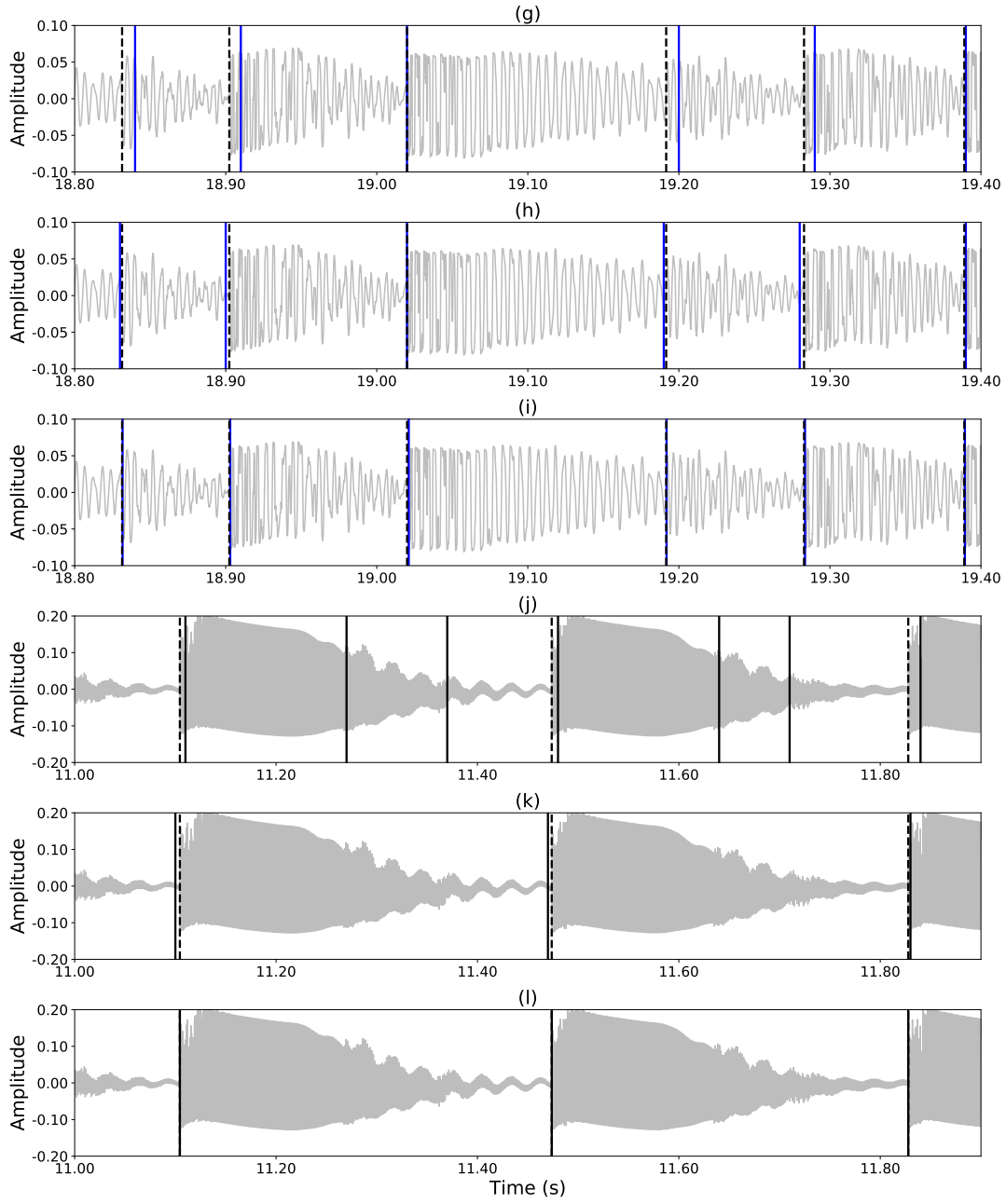

Figure 5.5: Part 2: Illustration of zoomed-in instrument signals with estimated onsets of track #27 coloured represented. Manual annotations in black dashed-line. Top to bottom: (g) Bombo High with Madmom annotations in red solid lines; (h) Bombo High with retrained DNN annotations in red solid lines; (i) Bombo High with time-corrected annotations in red solid lines; (j) Gonguê Low with Madmom annotations in green solid lines; (k) Gonguê Low with retrained DNN annotations in green solid lines; (l) Gonguê Low with time-corrected annotations in green solid lines.

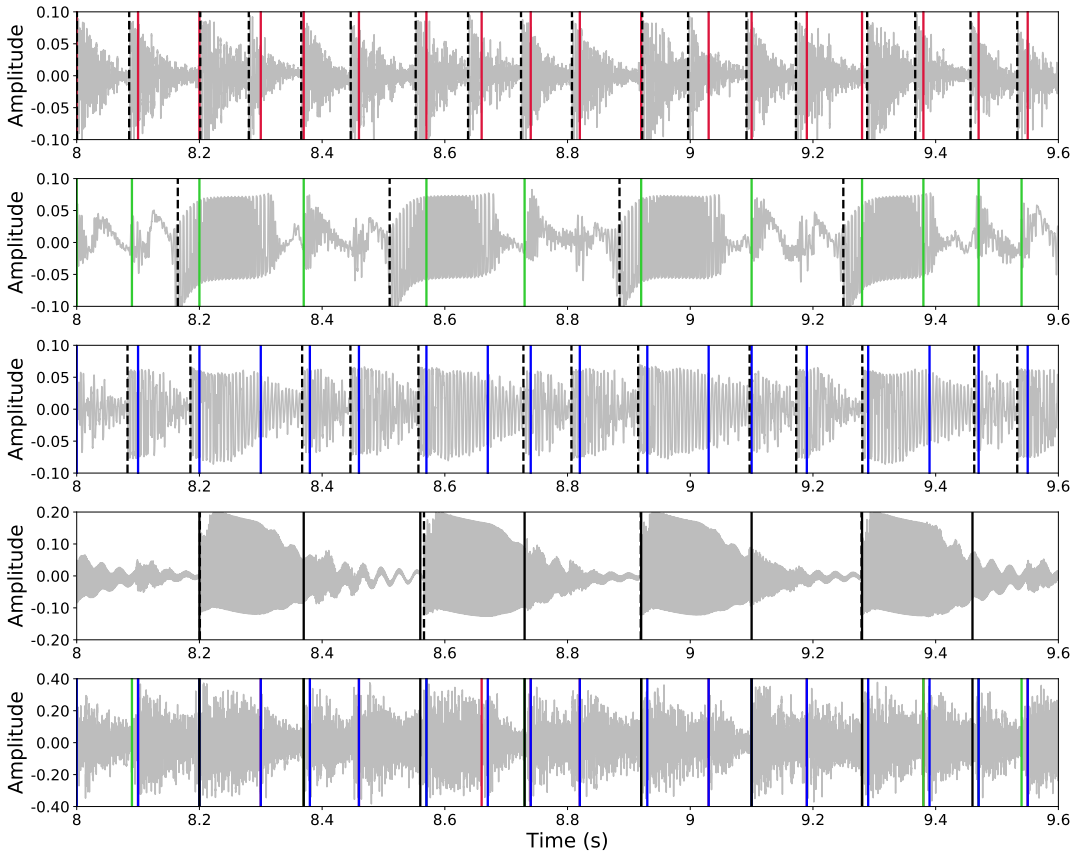

Figure 5.6: Illustration of isolated signals with overlaid estimated onsets from the stereo mixture signal obtained via retrained DNN models in approximately one bar of track #27. Top to bottom: Tarol with annotations in red solid lines; Porca with annotations in green solid lines; Bombo High with annotations in blue solid lines; Gonguê Low with annotations in black solid lines; Audio mixture with overlaid onsets of the four instruments. Manual annotations in black dashed-line.

Comparing it with previous figures, it's clear that difficulties were found when trying to estimate the onsets' location. In the Tarol and Bombo High tracks, the number of events detected is close to the manual annotations, however, the precision is much worse. In the Porca track there are some extra insertions plus the temporal location also has a significantly higher error associated.

Not surprisingly, the Gonguê Low track shows the best results in temporal location precision. The reason behind it, we argue, is that the sound produced by the Gonguê is higher-pitched in relation to the other instruments and hence is quite prominent in the stereo mixtures. It makes sense that, the extra detections correspond to the other bell of the Gonguê, the Gonguê High, since both have a very similar timbre since they are different parts of the same instrument. On this basis, we can predict that the performance of the DNN Gonguê model may be satisfactory for mixture signals if we ignore (for now) the extra insertions.

As the results in stereo mixture signals present themselves as imperfect for microtiming analysis, we take a step back and try to compute the retrained DNN models on a simpler mixture.

5.3.2.2 Mono Artificial Mixtures Signals

To assess if the onset estimation approach improves with a simpler mixture signal than the stereo one, we created artificial mixtures using the contact microphone signals from the dataset and mixed them, accordingly, on an open-source sound editor, Audacity, and exporting them as a mono mixed signal.

Our hypothesis is as-follows: if a mixture signal has less interference from other instrument's sound, then the retrained DNN instrument-specific models should provide better results. In Fig. 5.7 we show the multi-track view of the signals. By examining the bottom plot, corresponding to the mono mixture signal, we observe a cleaner signal when compared to the real, multi-instrument recording.

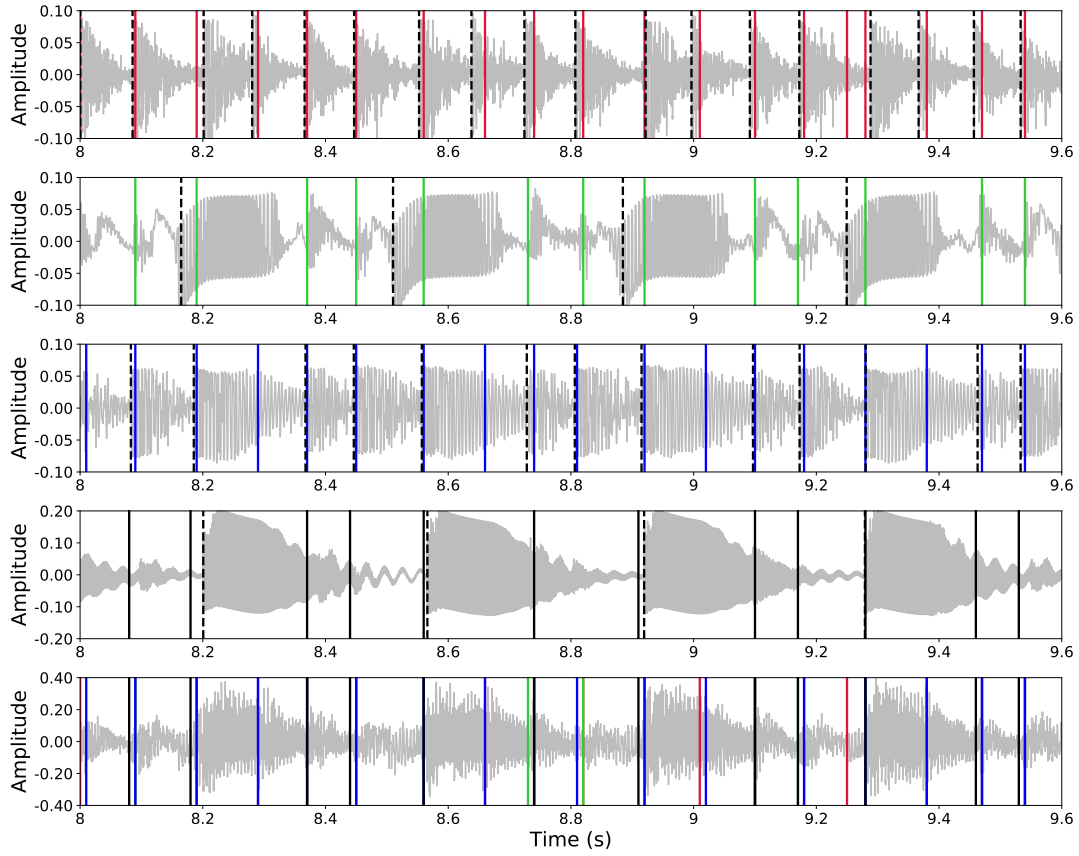

Figure 5.7: Illustration of isolated signals with overlaid estimated onsets from the artificial mixture signal obtained via retrained DNN models in approximately one bar of track #27. Top to bottom: Tarol with annotations in red solid lines; Porca with annotations in green solid lines; Bombo High with annotations in blue solid lines; Gonguê Low with annotations in black solid lines; Audio mixture with overlaid onsets of the four instruments. Manual annotations in black dashed-line.

Comparing with previous figures, our qualitative analysis quickly concludes that no substantial improvement is achieved when working with simpler and artificial mixtures. The absolute error in

the temporal location of the onset estimations appears transversal to all signals. For this reason, this approach is discarded.

Instead, we try a music source separation approach on the stereo signals to perform the same analysis and evaluate the outcome.

5.3.3 On Separated Signals

In the previous section, we observed how the DNN instrument-specific models estimate the onsets directly from mixture files. A preliminary evaluation allowed us to understand that some work is still needed to approximate the onset estimations from those obtained when working with isolated signals. For this reason, we now explore the possibility of applying a music source separation (MSS) algorithm.

Given a music mixture as input, MSS algorithms provide a separated estimation of each sound source as output, as detailed in Section 2.5. Thus, they offer a way of minimising interference from other instruments and a chance for the DNN models to obtain a higher percentage of corrected onset estimations.

5.3.3.1 Baseline Method

To obtain the separated sources from a stereo mixture at the highest possible quality, we scanned through the state-of-the-art MSS algorithms, as explained in Section 2.5. The deep neural network approaches are the ones that register the best separation results. However, the currently available models are trained for completely different types of sounds and require large amounts of data in the learning process. Therefore, we chose to work with the state-of-the-art NMF-based approach, specifically the PLCA algorithm from [59, 75].

The implementation applied in this work is presented in [59], in which the main function is described in 5.3 and we followed the procedures for semi-supervised source separation presented in [75].

$$[w,h,z] = plca(x,K,iter,sz,sw,sh,z,w,h,pl,lw,lh) \tag{5.3}$$

Inputs:

x - input distribution
K - number of basis functions
$iter$ - number of EM iterations
sz - sparsity of z
sw - sparsity of w
sh - sparsity of h
z - initial value of z
w - initial value of w
h - initial value of h

pl - plot flag

lw - columns of w to learn

lh - rows of h to learn

Outputs:

w - vertical bases

h - horizontal bases

z - component priors

The first step consisted of allocating track #1 of our dataset (4 isolated signals, one per considered instrument) to have the PLCA algorithm learn several spectral and temporal profiles, i.e., vertical and horizontal bases respectively. To that end, first, the contact microphone signals of the considered instruments of track #1 were converted to the TF domain through an STFT using a 4096-point Hanning window, with a 75% overlap. Subsequently, the PLCA function in 5.3 was computed four times, having each instrument signal as input distribution for each computation, 100 iterations, and the number of bases was set to 80.

The separation of the remaining 33 stereo mixtures of the dataset was done under the assumption that the activation matrices of each considered instrument of track #1, learned from the PLCA, are similar enough throughout the dataset to allow their use. This is accomplished by first converting the stereo mixture n to the TF domain using the same STFT parameters and averaging both channels into a single-channel mixture. The PLCA function was then computed, this time, inserting the fixed learned bases as an initial value of w. The number of bases was set to 240 bases to model the remaining of the signal and computed with 200 iterations each.

In Fig. 5.8 is shown an excerpt of the development of the separation process. The first subplot illustrates an excerpt of the isolated Tarol #27 spectrogram on which we can easily identify where the onsets occur by looking at the high-frequency vertical lines. The second subplot illustrates an excerpt of the stereo mixture #27 spectrogram that is more populated with frequencies from the other instruments making it less clear where the Tarol onsets are. The third subplot illustrates an excerpt of the separated Tarol from mixture #27 on which we can verify a cleaner spectrogram when compared with the mixture one, but a less informative spectrogram when compared to the first. The fourth subplot illustrates the mean activation function from the 240 bases, where each peak corresponds to the temporal activation of the modelled Tarol sound.

Based on informal listening to the separation results, we consider the overall quality to be satisfactory, except for the Porca, that had some interference from the other instruments. This difficulty can be explained when observing the isolated signal of the instrument and realising that Porca produces an harmonic shifting sound and, in consequence, the PLCA algorithm might need some extra spectral profiles to model each sound in the mixture. For this reason, we computed again the learning of the profiles with 120 basis vectors only for this specific instrument, which was

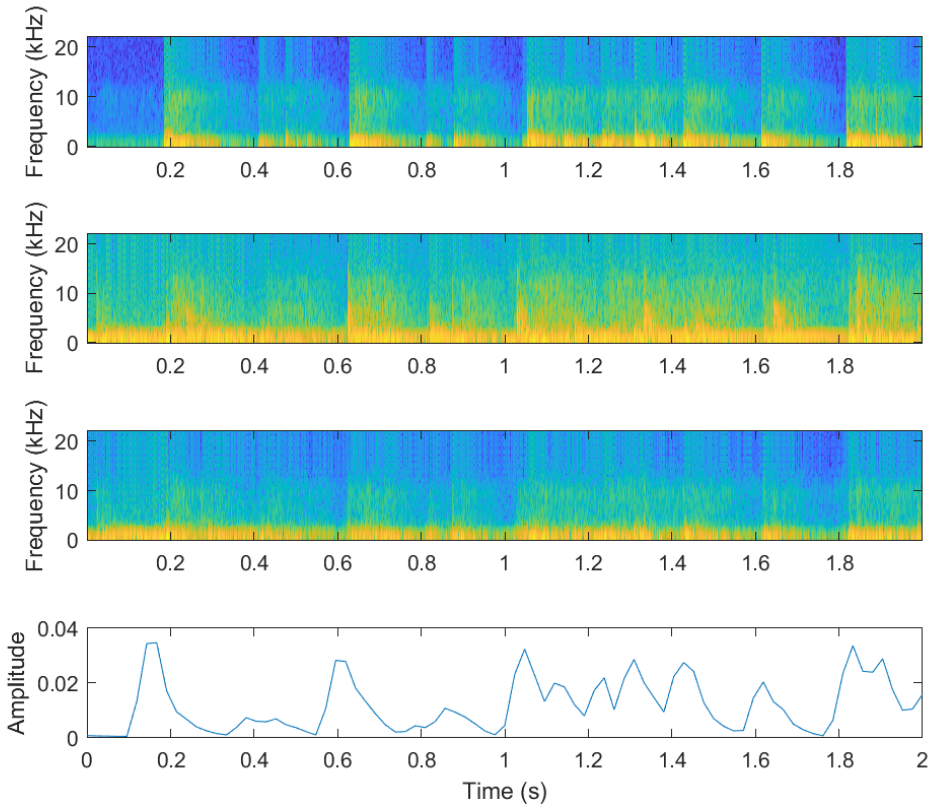

Figure 5.8: An example of spectrograms excerpts of track #27. Top to bottom: Spectrogram of Tarol isolated signal (from contact microphone). Spectrogram of mixture signal. Spectrogram of Tarol separated signal. Mean activation function of the 240 bases used in separation algorithm.

the chosen number of basis vectors, through trial and error, where perceptive separation quality stagnated, yet still worse when compared with the remaining instruments.

5.3.3.2 Applying Instrument-Specific DNN Models

Once the separated sources were obtained, the retrained DNN models from Section 5.3.1.2 could now be applied to automatically estimate the onsets on the separated sources. Fig. 5.10 shows the multi-track view of onset estimations on the separated sources. Due to the many signal processing operations in PLCA, we found, only for Gonguê Low separated signal, attenuation in amplitude to half of its original value (isolated signal), that severely affected the onset activation function (OAF) and, for this reason, we had to change the threshold value on the peak-picking stage, through trial and error, from 0.3 to 0.01, as depicted in Fig. 5.9 (note vertical axis scale in OAF in Gonguê Low separated signal). A set of demonstrative audio excerpts mentioned in the figures can be found here.[1]

[1] https://tinyurl.com/ybrc8ux8

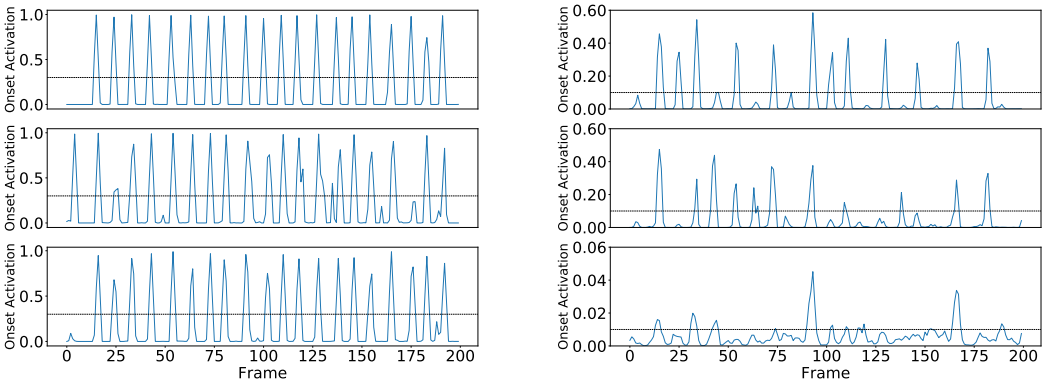

Figure 5.9: Onset activation function (output of DNN models) for a 2-second excerpt of track #27. Left column corresponds to OAFs of Tarol. Right column corresponds to OAFs of Gonguê Low. For both instruments: first row corresponds to OAF of isolated signal; second row corresponds to OAF of mixture signal; third row corresponds to OAF of separated signal.

Confirming our previous hypothesis, the worst performing onset estimation in Fig. 5.10 corresponds to the Porca separated signal. For the remaining instruments, the estimations have minor absolute errors. The Gonguê Low signal has some extra insertions that derive from poorer separation of the other bell likely arising from the Gonguê High, and thus interfering with the results. The time correction algorithm was applied in these separated signals but it didn't produce any improvement because the signals are not as clean the isolated ones, and any interference in the separated signals would limit the effectiveness of the time-correction approach, which is only designed to work under quite strict constraints.

5.4 Results and Discussion

In this section, the proposed onset estimations schemes were tested with the entire dataset signals, through a quantitative analysis.

5.4.1 On Contact Microphone Signals

The mean F1 scores calculated as a function of the width of the tolerance window for the isolated signals are depicted in Fig. 5.11 for the Madmom estimations, in Fig. 5.12 for the retrained DNN models estimations and in Fig. 5.13 for the time-corrected DNN estimations. In each case, the results show the performance evolution of each approach when the tolerance window length increases.

The first point to note is that there is some variation among the tested implementations, with a visible increase in performance from each method to the next.

Considering the Madmom estimation results in function of the tolerance window, we observe that it is possible to achieve a satisfactory mean F1 score from 15 ms on for Tarol and Bombo

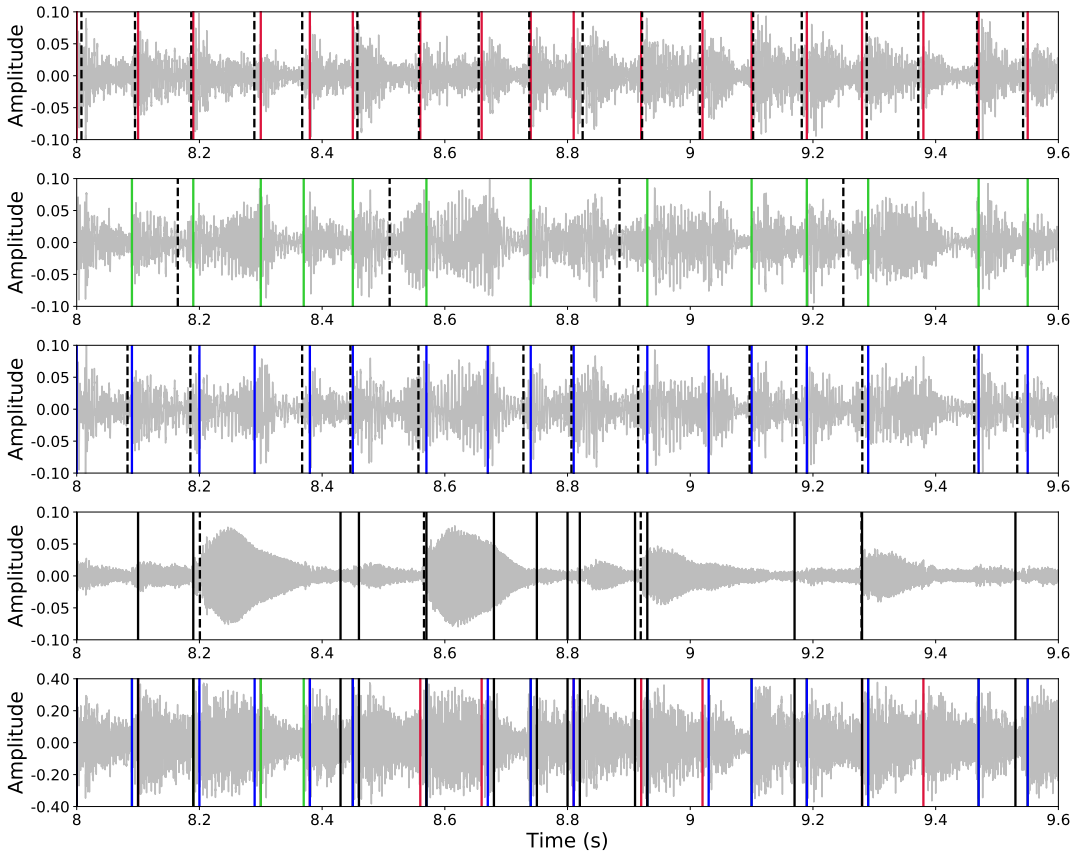

Figure 5.10: Illustration of separated sources signals with estimated onsets obtained via retrained DNN models in approximately one bar of track #27. Top to bottom: Tarol with annotations in red solid lines; Porca with annotations in green solid lines; Bombo High with annotations in blue solid lines; Gonguê Low with annotations in black solid lines; Audio mixture with overlaid onsets of the four instruments. Manual annotations in black dashed-line.

High. However, for Porca and Gonguê Low, the plotline stagnates around 15 ms at a mean F1 score below 0.6.

The impact of the retraining of the DNN instrument-specific models is quite clear from Fig. 5.12. The training process was done using data of the same genre from a homogeneous dataset and led to a substantial increase of performance, i.e., higher mean F1 scores for smaller tolerance windows. Together, the four considered instruments hit a mean F1 score above 0.9 at a tolerance window as low as 10 ms. The performance for the Gonguê Low signal almost doubled and for the Porca signal, more than doubled. Although, the Porca mean F1 performance is slightly below the other instruments perhaps due to the wider variety of sound types produced and which were not sufficiently modelled by the small amount of training dataset used to adapt the networks (in the case of each instrument this was just 5s).

The time-correction of the DNN estimations reveals itself as the best performing scheme to approximate the automatic estimations to manual annotations. With this method, the mean F1 scores reach their highest value for the four instruments for the smallest tolerance window lengths.

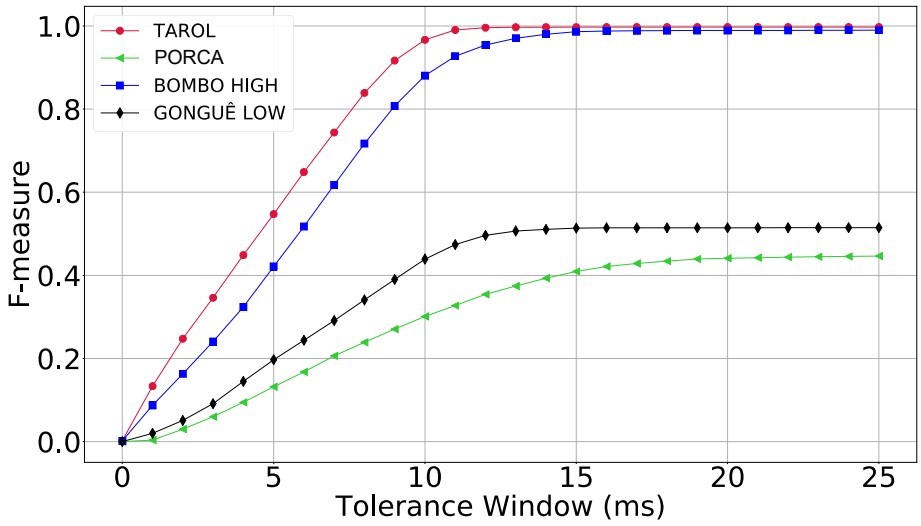

Figure 5.11: Mean F-measure score of Madmom estimations for the four considered instruments in function of the tolerance window. Tarol line in red. Porca line in green. Bombo High line in blue. Gonguê Low line in black.

From the tolerance window equal to 3 ms on, the mean F1 scores are all above 0.8. Compared to the previous method, at the same tolerance window, all mean F1 scores were below 0.6. This observation confirms our initial hypothesis declared in Section 5.3.1.3.

We now turn to another indicator of the performance of the employed methods, which show the evolution of the number of correctly detected onsets (true positives), incorrectly detect onsets (false positives), and incorrectly ignored onsets (false negatives) as a function of the tolerance window, for the entire dataset. In Fig. 5.14, Fig. 5.15 and Fig. 5.16 are shown the ratios between the number of true positives (TP), false positives (FP) and false negatives (FN) onset events for the Madmom estimations, retrained DNN models estimations and time-correction of the retrained DNN estimations, respectively. The sum of TP and FN events correspond to the number of ground-truth events as presented in Table 5.1 from Section 5.2.

In Fig. 5.14, the results show a tendency for extra detections (insertions), reaching up to two times the real number of events in the case of Tarol and Bombo High (top row) or even over four times the real number of events in the case of Porca and Gonguê Low (bottom row). These visualisations confirm the ones from Fig. 5.11, for example, in the cases of Tarol and Bombo High, that eliminate almost all FP detections with a tolerance window greater than or equal to 10 ms.

Similarly, in Fig. 5.15, the results show the improvement of the retraining of the DNN models. The number of false positives drastically decreases for the Porca and Gonguê Low estimations

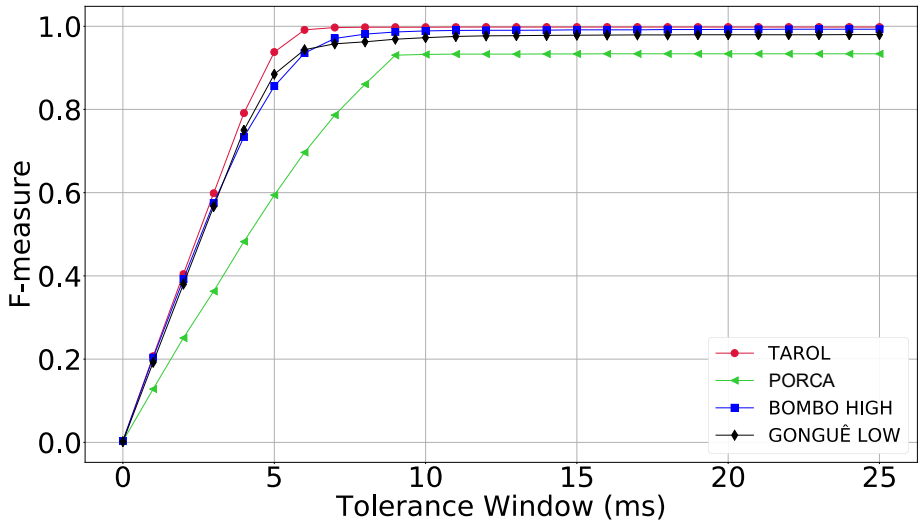

Figure 5.12: Mean F-measure score of DNN models estimations for the four considered instruments in function of the tolerance window. Tarol line in red. Porca line in green. Bombo High line in blue. Gonguê Low line in black.

where the ratio between TP, FN, and FP now have a similar behaviour for the four instruments. The residual number of FP that are identified in Porca estimations for tolerance windows larger than 10 ms are the reason for the slight F1 underperformance when estimating the onsets of this instrument.

Culminating with the results of the time-correction method, present in Fig. 5.16, we see the practical elimination of FP for tolerance windows larger than 2 ms, with the exception of the Porca time-corrected estimations where the residual FP events were maintained.

Overall, these results show that when using an instrument-specific onset detector together with a temporal localisation adjustment on clean contact microphone signals, it is possible to achieve a high level of performance on this dataset. In turn this suggests that the microtiming visualisation computed with these fully automatic estimations may be close to the visualisations computed with manual annotations.

5.4.2 On Mixed Audio Signals

Moving on to the results on mixed signals, we proposed, in Section 5.3.2.1, to extract the onset locations of each considered instrument directly from the stereo mixture in which are present up to 8 different sound sources. To evaluate the performance of our approach on these signals on

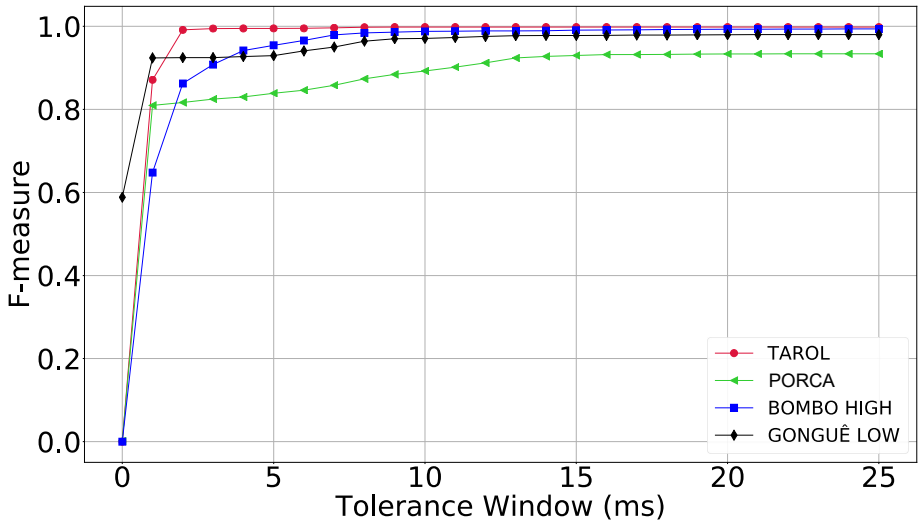

Figure 5.13: Mean F-measure score of time-correction of DNN estimations for the four considered instruments in function of the tolerance window. Tarol line in red. Porca line in green. Bombo High line in blue. Gonguê Low line in black.

the same parameters, the mean F1 scores with varying tolerance windows were computed and are depicted in Fig. 5.17.

Not surprisingly, the performance decreases substantially to the point of achieving mean F1 score above 0.5 for Bombo High and Gonguê Low only for tolerance windows larger than 17 ms. It is worth noticing that the Gonguê Low plotline in this figure is quite similar to the plotline of the same instrument in Fig. 5.11 of the Madmom estimations with isolated signals that allows us to draw an interesting parallel about the performance of the retrained DNN instrument-specific models. The estimations of the Tarol score high in F1, to a maximum of 0.8 at 24 ms. The worst performing estimation is for Porca, with a mean F1 score of only 0.3 at the maximum window length. Because this instrument produces a very low-pitched sound, which is less prominent in the mixture recordings, the results are unsuprisingly worse.

To confirm our observations, the number of TP, FN, and FP counted from the mixture signal estimations are shown in Fig. 5.18. At a glance, the results show a considerable increase in the number of FP for all tolerance window lengths for all instruments, with the exception for the Gonguê Low. Also, the numbers of TP seem to be small, especially for Porca and Gonguê Low estimations.

Fig. 5.18 shows that the numbers of TP for the Tarol and Bombo High are close to 90% for tolerance windows larger than 20 ms. In consequence, the same figure helps us understand that

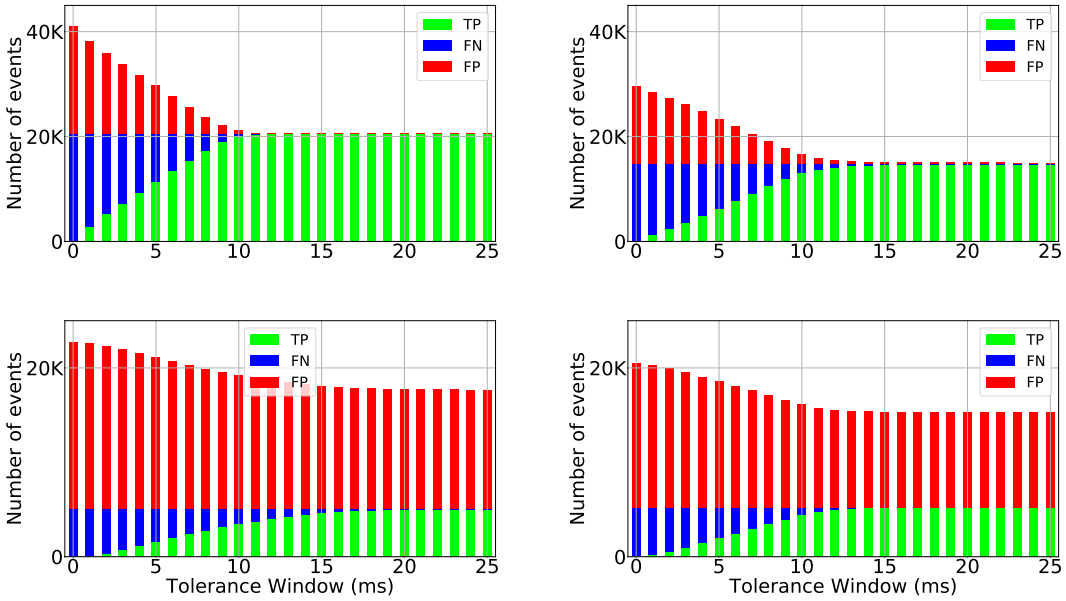

Figure 5.14: Number of TP (in green), FP (in red) and FN (in blue) for each instrument of the Madmom estimations in function of the tolerance window. Top left corresponds to Tarol. Top right corresponds to Bombo High. Bottom left corresponds to Porca. Bottom right corresponds to Gonguê Low.

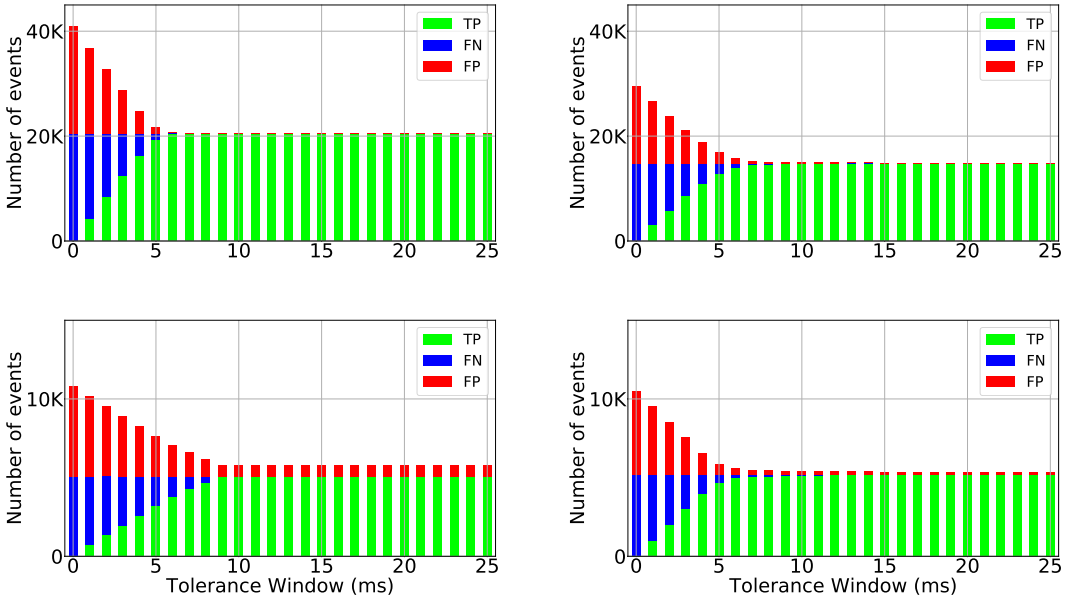

Figure 5.15: Number of TP (in green), FP (in red) and FN (in blue) for each instrument of the retrained DNN estimations in function of the tolerance window. Top left corresponds to Tarol. Top right corresponds to Bombo High. Bottom left corresponds to Porca. Bottom right corresponds to Gonguê Low.

the lower F1 scores occur because of the large number of extra detections (FP). This behaviour tends to decrease in Tarol for larger tolerance windows but not as much for the Bombo High that

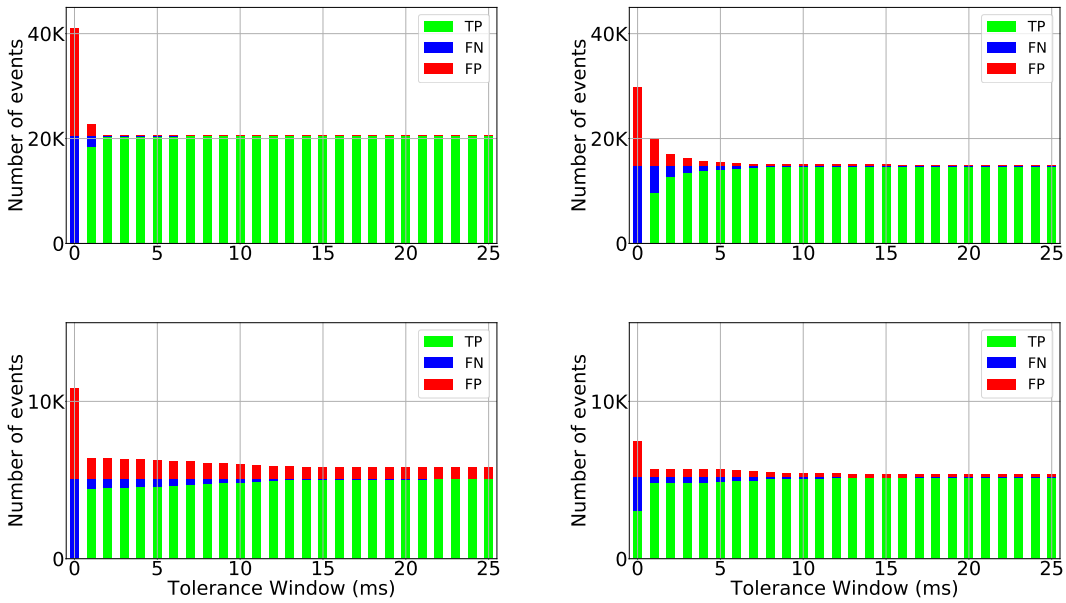

Figure 5.16: Number of TP (in green), FP (in red) and FN (in blue) for each instrument of the time-correction of the retrained DNN estimations in function of the tolerance window. Top left corresponds to Tarol. Top right corresponds to Bombo High. Bottom left corresponds to Porca. Bottom right corresponds to Gonguê Low.

registers a number of FP similar to the sum of FN and TP for maximum tolerance window length, presumably, due to the similar timbre of the sound produced by the lower skin of the Bombo, Bombo Low. On the contrary, the results for the Gonguê Low estimations show that the worse performing F1 score is due to the narrow number of TP detected rather than the excessive amount of FP detections. Porca estimations are again the worst performing estimations having the number of FP more than double when compared to the actual percussive events. Additionally, the number of TP are low and close to zero for tolerance windows shorter than 5 ms.

Overall, the outcome of estimating the onsets directly from the mixture signal proved to be a challenging task. We highlight the presence of other sources (up to 8) in the signal and similar timbre between some of the instruments as the main reasons for the obtained results and, for this reason, the main motivation to do a performance analysis with separated sources.

5.4.3 On Separated Signals

We now switch to the evaluation of the estimations computed with the separated sources where the mean F1 scores in function of the tolerance window are shown in Fig. 5.19.

The results show an important improvement on the Tarol and Bombo High scores across wider tolerance window lengths. At 25 ms, both instrument estimations score around 0.8, with Tarol scoring near 0.9, an increase of 0.1, and Bombo High scoring slightly below 0.8, representing almost a 0.2 increase. It is also worth mentioning that the highest F1 scores for the Tarol and Bombo High obtained on the mixture signals are achieved, with the separated sources, at 18 ms

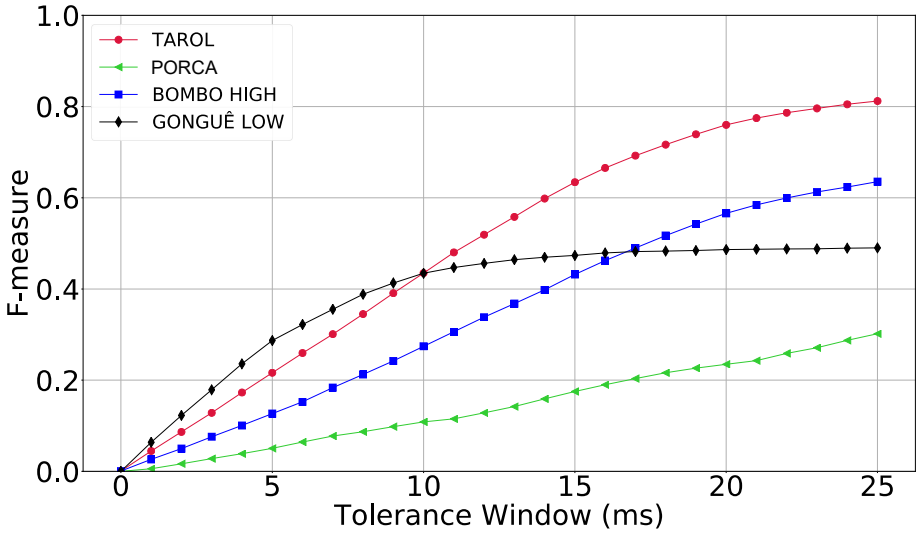

Figure 5.17: Mean F-measure score of the stereo mixture estimations obtained via retrained DNN models in function of the tolerance window. Tarol line in red. Porca line in green. Bombo High line in blue. Gonguê Low line in black.

tolerance window, that represents a 7 ms decrease in tolerance window length. The F1 score evolution of the Gonguê Low and Porca estimations maintained a similar behaviour to the previous approach. For the Gonguê Low, the maximum F1 score is around 0.5 at 25 ms. The slower increase of F-measure plotline for this instrument was the only identified difference. For the Porca, the plotline has also a similar behaviour, although the maximum score is lower than the previous approach. In this scheme, Porca scores a maximum F1 of 0.25 at 25 ms whereas the estimations on the mixture signal the maximum F1 score was 0.3 for the same window length. This small decrease in performance can be explained by the imperfect separation for this specific instrument as we refer to in Section 5.3.3.1.

The visualisations from Fig. 5.20 corroborate our observations. The increase of performance in Tarol and Bombo High (top row) is evident and comes from the reduction of the number of FP when compared to mixed signals. The Gonguê Low estimations maintain a similar behaviour although we notice an increased number of FP, with the stacked bars being over 10k for all tolerance windows where on mixed signals, they were below 10k. The mean F1 score line is similar to the one of mixed signals because, despite the higher number of FP, the number of TP also increased substantially, compensating the over-detections. The Porca estimations saw a slight decrease on the already small number of TP allied to the increase of FP events justify the not-satisfying performance of the estimations in this instrument.

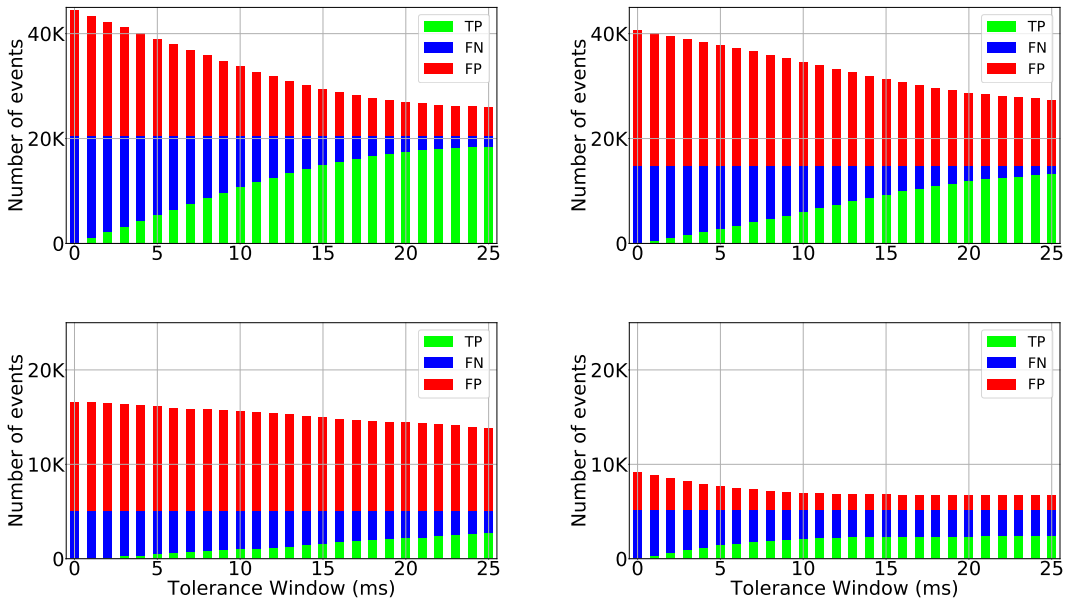

Figure 5.18: Number of TP (in green), FP (in red) and FN (in blue) for each instrument of the stereo mixture estimations obtained via retrained DNN models in function of the tolerance window. Top left corresponds to Tarol. Top right corresponds to Bombo High. Bottom left corresponds to Porca. Bottom right corresponds to Gonguê Low.

Overall, we observed an increase of performance on the both rhythmically expressive instruments which is a promising outcome when considering the subsequent visualisation of micro-timing patterns in these instruments. The downside of this approach relies on the fact that its performance is proportional to the separation quality, which is evident when compared with the results of the perfectly isolated signals.

5.4.4 Overview Of The Three Signal Scenarios

To give us a sense of the distribution of the F1 scores from which we calculated the mean in previous figures, presented in Fig. 5.21, Fig. 5.22 and Fig. 5.23 are box-plot distributions in function of the tolerance window for the best performing approach on isolated signals (time-corrected estimations), for the mixture signals and the separated sources, respectively.

The F1 distributions of the time-corrected estimations are very narrow, meaning that scores did not differ much from each other and tending to 1. Again, only for the Porca there was a bigger variance of the F1 scores. But, when compared to the next two figures, Fig. 5.22 for mixture signals and Fig. 5.23 for separated source signals, the challenges found are observable through the extended boxplots and the rather wide range of the upper and lower whiskers. As a consequence, there are almost no outliers present in both figures.

In conclusion, it became clear that the time-corrected estimations from the retrained DNN models achieved the best results by having very high F1 scores, close to 1, on isolated signals and

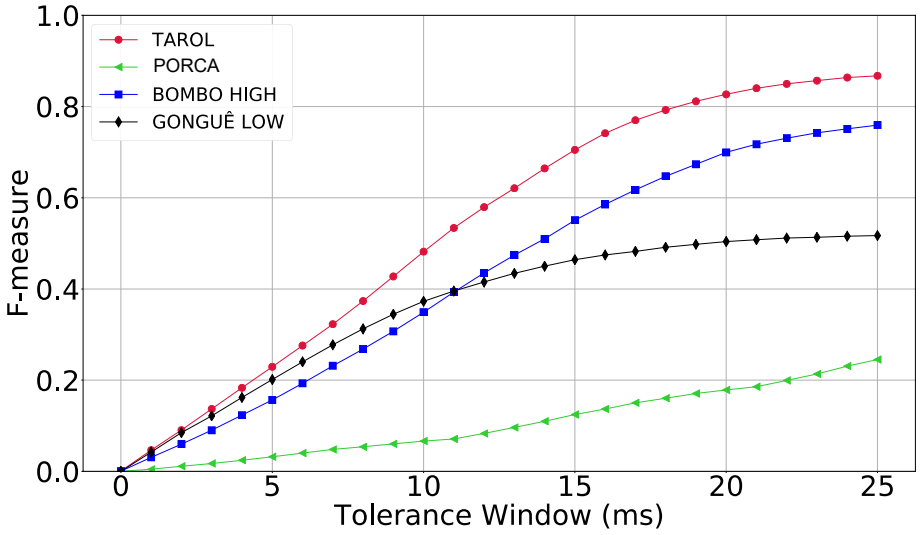

Figure 5.19: Mean F-measure score of the separated sources estimations obtained via retrained DNN models in function of the tolerance window. Tarol line in red. Porca line in green. Bombo High line in blue. Gonguê Low line in black.

thus, hinting at a good basis for their usage in microtiming visualisation. The attempt of estimating the onsets directly from the mixture signal with the retrained instrument specific DNN models revealed insufficient results and proved to be too difficult of a task for this type of music that motivated source separation from mixture signals. In this approach we observed improvements, specially for the Tarol and Bombo High that are two rhythmically expressive instruments and thus, suggesting the possibility of a better microtiming visualisation presented in the next chapter.

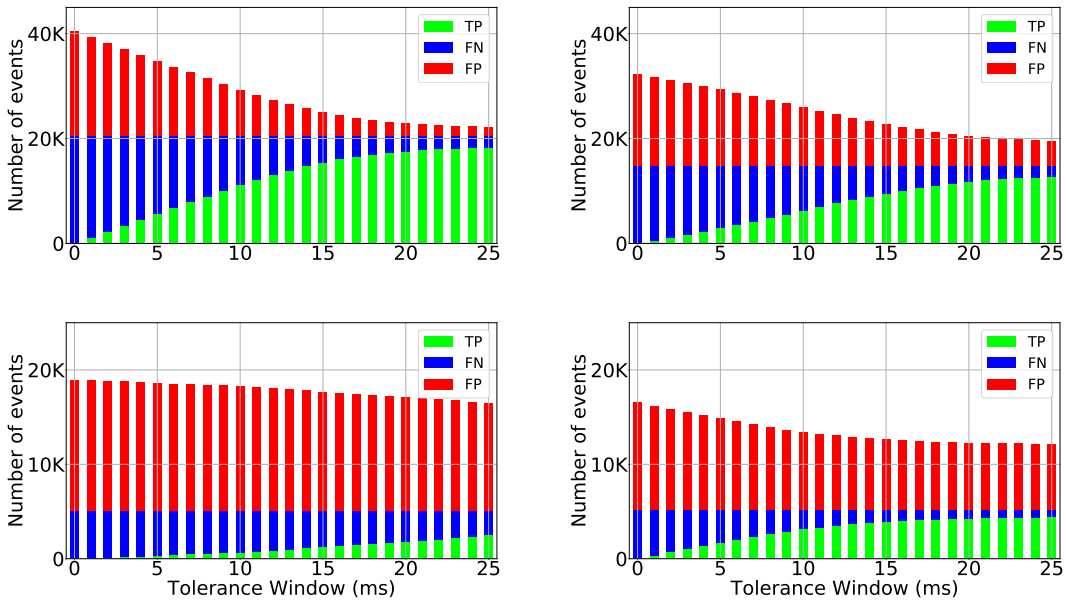

Figure 5.20: Number of TP (in green), FP (in red) and FN (in blue) for each instrument of the separated sources estimations obtained via retrained DNN models in function of the tolerance window. Top left corresponds to Tarol. Top right corresponds to Bombo High. Bottom left corresponds to Porca. Bottom right corresponds to Gonguê Low.

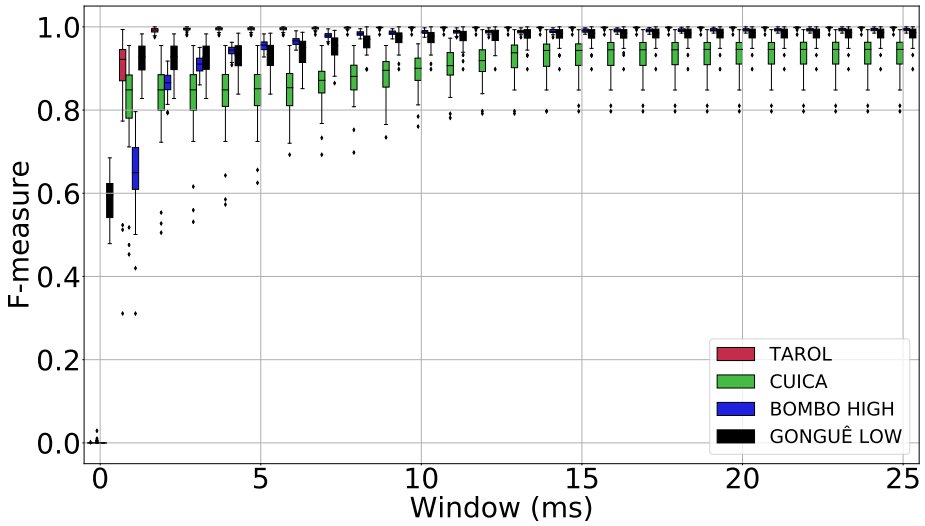

Figure 5.21: F-measure score distribution of time-corrected estimations on the isolated signals in function of the tolerance window. Tarol score distribution in red box. Porca score distribution in green box. Bombo High score distribution in blue box. Gonguê Low score distribution in black box.

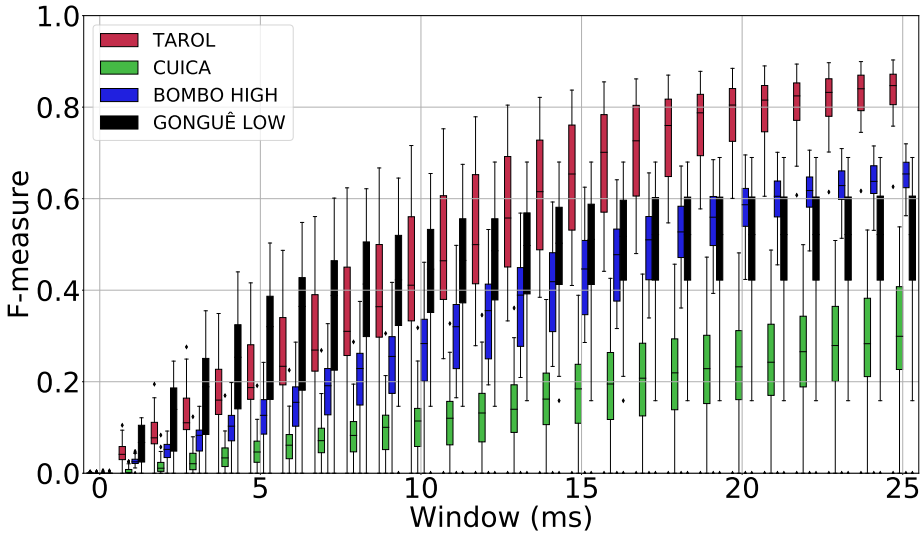

Figure 5.22: F-measure score distribution of retrained DNN models on the mixture signals in function of the tolerance window. Tarol score distribution in red box. Porca score distribution in green box. Bombo High score distribution in blue box. Gonguê Low score distribution in black box.

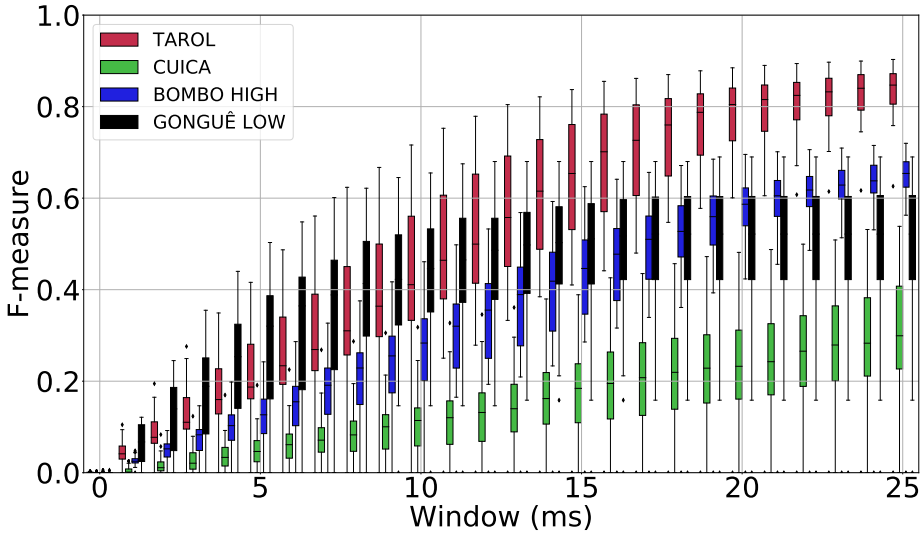

Figure 5.23: F-measure score distribution of retrained DNN models on the separated source signals in function of the tolerance window. Tarol score distribution in red box. Porca score distribution in green box. Bombo High score distribution in blue box. Gonguê Low score distribution in black box.

Chapter 6

Microtiming Visualisation

6.1 Introduction

In this chapter, we focus on observing the microtiming structures that might be present on the recordings of the Maracatu dataset. With the onset estimations provided in the previous chapter, it is now possible to compute a visualisation of the onset locations in a rhythmically meaningful way. The goal of this chapter is to discover the impact that the different onset estimation scenarios have on the microtiming visualisation of different instruments.

To visualise the microtiming profiles, we use the capabilities of the CARAT library [22] to plot the evolution of the onset positions over multiple beats and investigate the presence of systematic and unsystematic microtiming patterns over time. Note that this visualisation allows us to focus on identifying microtiming patterns because the selected Maracatu tracks have only one main rhythmic pattern through the entire piece [72].

The remainder of this chapter will focus on describing the frame that enables microtiming visualisation followed by the comparison of the microtiming profiles of the two rhythmically expressive instruments, Tarol and Bombo High, with the Gonguê Low as the reference timekeeper and finally, the discussion about whether or not it is possible to identify microtiming deviations on the different signal scenarios.

6.2 Microtiming Profiles

To undertake any evaluation of the microtiming structure present in recordings, we need to obtain temporal markers that indicate the note onset positions. Following previous works in microtiming analysis [15, 21, 14], we create a reference beat grid to be able to compare the locations of performed onsets with a quantised beat and/or sub-beat positions.

Due to the tempo variations possibly present in the recordings, it is not possible to analyse timing data in absolute duration. For this reason, each beat interval can be assigned a normalised duration of 100%, and thus a rhythmic pattern containing four equal sub-divisions would occur at normalised positions 0%, 25%, 50%, and 75%. The onsets are converted to their relative position

with regards to the beats and are assigned to a position in an isochronous metrical grid (equally distributed subdivisions within the beat) [14].

Although previous approaches studied within-instrument rhythmic patterns, this work, aligned with [10], focuses on identifying the between-instrument microtiming deviations, assuming the Gonguê Low onsets as the beat reference (since it always plays the beat) and Tarol and Bombo High as the rhythmically expressive onsets [72].

In Fig. 6.1 and Fig. 6.2 are illustrated the microtiming profiles for the Tarol and Bombo High, respectively, computed with manual annotations on isolated signals, best-performing automatic onset estimation on isolated signals (time-correction method), DNN models on mixture signals and DNN models on separated source signals. The red marks represent the location of the onsets within each beat, represented horizontally through a beat-length normalisation, where the .1 tick corresponds to the first sub-beat position (that coincides with the beat) and the remaining .2, .3 and .4 ticks correspond to the second, third and fourth sub-beat positions. The beat evolution across time is displayed vertically, with time increasing bottom-up. At the bottom of each sub-figure is a histogram of the onset locations providing a measure of the mean location of events within a group (extended by the light-blue dashed lines) and their amount of dispersion.

6.3 Discussion

Regarding Fig. 6.1, we can see microtiming patterns of track #27 consistent with a sub-division of the beat into four 1/16th notes for Tarol.

Looking at the microtiming profiles computed with manual annotations (top left), we can recognise the controlled fluctuations around the mean deviation value, moving back and forth the quantised positions that indicate a dynamic microtiming. Furthermore, from looking at the histograms, we see that the onsets of the Tarol have the following patterns: [1%, 26%, 49%, 73%]. These values indicate, on average, that the first sub-beat is 1% behind the beat, the second sub-beat follows the same behaviour, the third sub-beat is 1% ahead and the fourth sub-beat is 2% ahead of their quantised positions.

Turning to the microtiming profiles computed with the estimations of the time-correction of the DNN models estimations, the best performing approach on isolated signals, where we can observe a similar microtiming dynamics for the Tarol throughout the piece. Even if, the mean deviations present in the histograms don't perfectly match, they are pretty close with a 1% deviation on the first and fourth sub-beats, approximating these to the quantised positions.

From the estimations extracted of mixture signals (bottom left), we verify that the microtiming structures from the previous observations almost disappear leaving no trace of the microrhythmically rich arrangements. We observe a heavy concentration of the position of the onsets near the deviation line that could be the result of a quantisation effect in the estimation process. This is special prominent on the first sub-beat histogram where we see a peak that reflects that superposition of several onsets onto the same grid location. On a smaller scale, the same applies to the third sub-beat position.

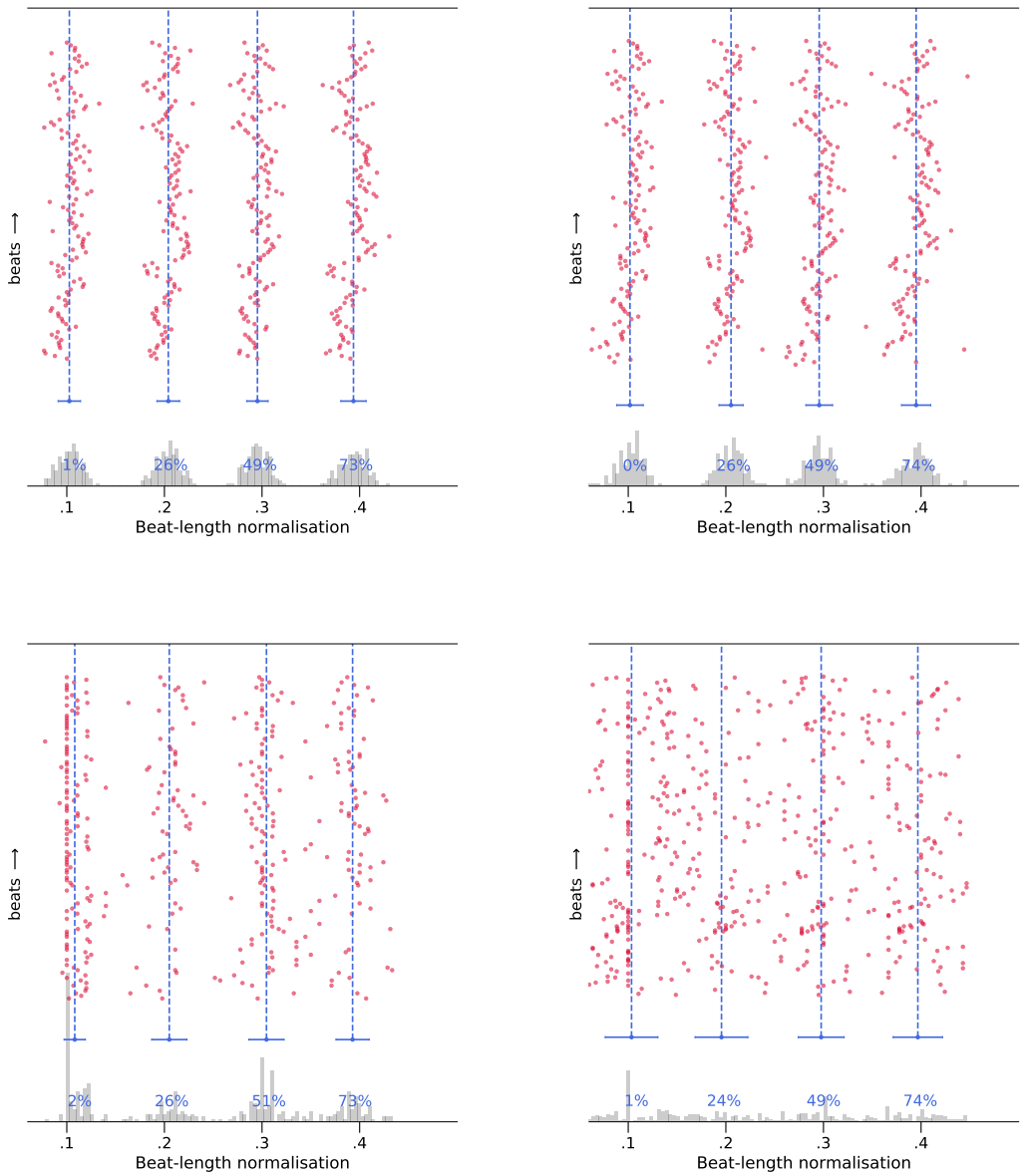

Figure 6.1: Microtiming profiles in track #27 for Tarol with Gonguê Low as the beat reference. Top left profile computed with manual annotations. Top right profile computed with estimations from isolated signals. Bottom left profile computed with estimations from mixture signals. Bottom right profile computed with estimations from separated signals.

In the opposite direction, the estimations obtained from the separated sources (bottom right) are more dispersed through the beat-length normalisation to the point of not being clear to which group belongs the onsets in between sub-beat positions. On the histograms, the onset distributions across the horizontal axes are visible showing no mean deviation value, except for a small one for the sub-beat.

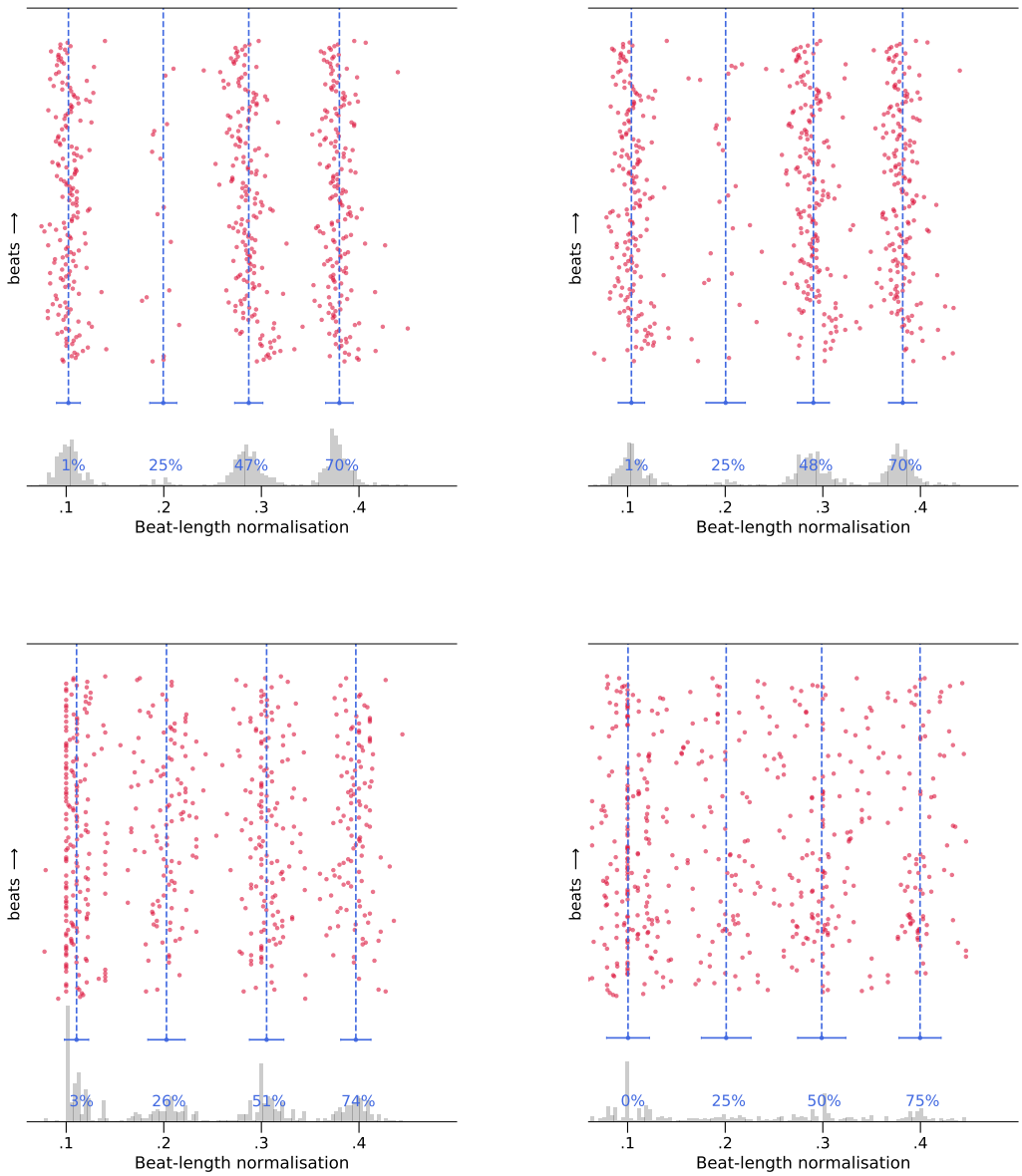

Figure 6.2: Microtiming profiles in track #23 for Bombo High with Gonguê Low as the beat reference. Top left profile computed with manual annotations. Top right profile computed with estimations from isolated signals. Bottom left profile computed with estimations from mixture signals. Bottom right profile computed with estimations from separated signals.

Overall, for this instrument, we can conclude that the visualisation of microtiming patterns can be observed and identified with the onsets estimated from the isolated signals in a very reliable way. However, for the other two signal scenarios, the same does not apply, confirming the hypothesis of under-performance pondered as we observed the F-measure boxplot score distributions from the previous chapter. Part of the problem with the visualisations on the mixture and separated signals also lies in the fact that Gonguê Low estimations, the timekeeper instrument, are

way below from satisfactory as we could read from the mean F1 scores, resulting in unclear beat references that affect the evaluation.

Relating to Fig. 6.2 we can see microtiming patterns of track #23 consistent with a sub-division of the beat into four 1/16th notes, with the second 1/16th note not played for Bombo High.

Looking at the microtiming profiles computed with manual annotations (top left), we can recognise a similar microtiming dynamic as in Tarol. The deviation patterns for Bombo High are: [1%, 47%, 70%]. Since this instrument plays only three notes per beat in Maracatu, the deviation percentage present on the second sub-beat should be ignored. The rare onsets that are marked around the second sub-beat in the visualisation with manual annotations (top left) are a combination of incorrectly annotated events and strokes with large deviations from their beat positions that seldom occur at the beginning of the tracks when the musicians are not yet in perfect synchrony. From the same sub-figure, we can see that in Bombo the mean anticipations in the third and fourth sub-beats are higher than in Tarol, specifically, a 2% and 5% anticipation, respectively.

This figure also confirms the hypothesis from the previous chapter, that for this instrument too it is possible to observe a very similar microtiming behaviour when computed with the time-corrected estimations obtained from the isolated signals (top right). Also, the histograms show a very similar mean deviation as in the visualisation with manual annotations.

On the mixture and separated signals it was expected a worse visualisation for the same reasons of the Tarol, adding to the fact that for Bombo High, the F-measures scored below the Tarol with the similar dispersion of results. Note that, in the bottom two sub-figures, not only the microtiming structures are not visible, but the rhythmic structure of the Bombo High onsets is also completely lost, showing onsets in the four sub-beats, which is inaccurate. For the separated signals in particular, the estimated onsets are so dispersed that there is no identifiable trace of deviations as we can see from the histograms fitting the all onsets in their quantised positions. Since the Gonguê Low onsets are also used in these visualisations as the beat reference, the same difficulty as the previous applies.

Small quantitative differences aside, the same behaviour examined in Fig. 6.1 and Fig. 6.2 of both instruments was observed throughout the Maracatu dataset.

From this analysis, we conclude that estimating onsets automatically from perfectly isolated contact microphone signals provides the highest probability of a successful microtiming visualisation and, consequently, microtiming analysis. These visualisations can serve as the basis for spreading the practice of capturing signals with contact microphones for microtiming analysis purposes and, to a greater extent, to even skip the exhaustive process of manually annotating a dataset because, as we could see, the utilised approach on isolated signals had very high performance. On the other hand, if it's only possible to access mixed signals, then the hopes of observing microtiming should remain low, even with the possibility of separating the sources through a music source separation techniques, because it is very likely that the microtiming structures cannot be easily identified from the current automatic onset estimation methods. An interesting area for future work would be to consider how possible it is to perform an accurate manual annotation of onset and beat positions from separated sources.

Chapter 7

Conclusion

7.1 Summary and Future Work

In this work we performed an evaluation on the performance of several automatic onset estimation approaches for microtiming visualisation purposes in the understudied Brazilian Maracatu de Baque Solto. In the process, we recorded and annotated a 34 track dataset of this music tradition in the context of the ongoing HELP-MD project. The signal acquisition process was made using contact microphones that were placed on the surface(s) of the instruments with the aim of obtaining clean signals with very little leakage from other sources. The annotations took four full weeks to be finished, in a process that led to labelling over 45,500 onsets over four instruments.

Furthermore, we showed that the state-of-the-art onset detector encountered difficulties when estimating the onsets of this style, verifying that automatic onset estimation is not a solved problem yet within the MIR research. We then applied an in-development tool adapted specifically to detect the onsets of the instruments used in Maracatu followed by a correction algorithm that outperformed the state-of-the-art onset detector algorithm.

Then, we took it one step further and tested this approach on mixture signals with the objective of estimating the onsets of each instrument, separately. Ultimately, we found this approach delivered unsatisfying results when compared with the results derived from the estimations on isolated signals.

To test it from another angle, we applied a state-of-the-art music source separation algorithm on the mixture signal to try to approximate to the conditions of the isolated signals. The results showed a small improvement for the two rhythmically expressive instruments but still far satisfactory scores.

Finally, we computed the microtiming visualisations using the manual annotations on isolated signals, the best performing estimations on isolated signals, the estimations on mixture signals and on separated signals. In these conditions, we could confirm that microtiming visualisation is currently only possible when working with isolated signals. The resulting visualisations from mixture and separated signals are still far from the goal of observing and analysing microtiming structures, leaving the door open for future research to adapt existing onset detecting methods that

precisely estimate the musical events on this and similar genres directly from the mixture signal, or even use the Maracatu dataset to evaluate a universal onset detector performance.

Also, there is the opportunity to enhance the quality of the music source separation algorithms to enable micro-rhythmic analysis and bridge the gap of when isolated signals are not possible to capture. In addition, we showed the long and non-trivial challenges we had to overcome in order to visualise the microtiming in Maracatu in a computational way.

Lastly, this work could serve as basis to document the possibilities of capturing instrument signals via contact microphones for microtiming analysis purposes that, when combined with existing automatic onset estimators, could eliminate or heavily reduce the exhaustive process of manual annotating datasets and preserve microtiming information.

In essence, the conclusions from this work could help future research to accelerate the repetitive task of annotations/estimating a dataset with the characteristics of the Maracatu so that the focus can be pointed at the analysis of microtiming in this and other traditions, to associate this quantitative information to more abstract concepts of playing collectively and to explore musically phenomena invisible through the Western lens of understanding music.

7.2 Perspectives

In the broader context of the HELP-MD project to which the work in this dissertation contributes, we believe the findings obtained illustrate that it remains extremely challenging to obtain sufficiently high quality automatic analysis to perform reliable ethnomusicological analysis. Substantial effort was required first to compile and annotate the dataset in order to be able to retrain and test automatic onset detection methods. This manual annotation was critical to provide the reference against which automatic methods could be compared. Concerning the specific results obtained, we believe that is already a success to demonstrate similar microtiming visualisations, even with the restricted domain of contact microphone signals, when using fully automatic onset detection. To this end, it is not surprising that once we move away from this most constrained signal acquisition and analysis scenario that performance degrades. Indeed, while it was not possible to obtain equivalent performance from the stereo mixtures or separated sources, we believe this kind of "negative" result constitutes an important first step to motivate future research in which some of these constraints can be relaxed. As we noted earlier in the dissertation, while musical audio source separation has improved greater due to the advent of deep neural networks, the publicly available approaches could not be applied here, both due to the need to individual separate percussion instruments rather than isolate all of them together from other musical instruments, as well as the lack of available fully separated data in which such models could be retrained. In this kind of musicological work which ultimately focuses on a small set of recordings from single set of performers on fixed instruments, we cannot hope to obtain "big data" in the sense to which the term is applied in other domains using DNNs. Thus, this raises important questions about how we can develop techniques that can learn from small data, or adapt existing trained models so they can be effective in highly constrained circumstances.

References

[1] Steve Swallow and Paul Bley. Le real book - partition gratuite. *Adapted from:* https://www.swiss-jazz.ch/standards-jazz/BlueInGren.pdf.

[2] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on speech and audio processing*, 13(5):1035–1047, 2005.

[3] Nicholas Bryan and Dennis Sun. Ccrma, stanford university - source separation tutorial mini-series ii: Introduction to non-negative matrix factorization. *Adapted from:* https://ccrma.stanford.edu/~njb/teaching/sstutorial/part2.pdf.

[4] Thomas Turino. *Music as social life: The politics of participation*. University of Chicago Press, 2008.

[5] Markus Schedl, Emilia Gómez, Julián Urbano, et al. Music information retrieval: Recent developments and applications. *Foundations and Trends in Information Retrieval*, 8(2-3):127–261, 2014.

[6] Avery Wang et al. An industrial strength audio search algorithm. In *Ismir*, volume 2003, pages 7–13. Washington, DC, 2003.

[7] Bas de Haas, José Pedro Magalhaes, Dion Ten Heggeler, Gijs Bekenkamp, and Tijmen Ruizendaal. Chordify: Chord transcription for the masses. *in International Society for Music Information Retrieval Conference*, page 8–12, 2012.

[8] Instituto de Etnomusicologia Centro de Estudos em Música e Dança. Help-md - the healing and emotional power of music and dance. *Adapted from:* http://www.inetmd.pt/index.php/en/investigacao/projects/9609-help-md-the-healing-and-emotional-power-of-music-and-dance.

[9] Journal of the Portuguese Republic. Diário da república n.º 53/2019, série ii de 2019-03-15. *Diário da República link:* https://dre.pt/web/guest/pesquisa/-/search/121075755/details/.

[10] Matthew E. P. Davies, Magdalena Fuentes, João Fonseca, Luís Aly, Marco Jerónimo, and Filippo Bonini Baraldi. Moving in time: Computational analysis of microtiming in maracatu de baque solto. In *21st International Society for Music Information Retrieval Conference, ISMIR*, 2020.

[11] Matthew Davies, Guy Madison, Pedro Silva, and Fabien Gouyon. The effect of microtiming deviations on the perception of groove in short rhythms. *Music Perception: An Interdisciplinary Journal*, 30(5):497–510, 2012.

[12] Jeffrey Adam Bilmes. *Timing is of the essence: Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm*. PhD thesis, Massachusetts Institute of Technology, 1993.

[13] Anders Friberg and Andreas Sundströöm. Swing ratios and ensemble timing in jazz performance: Evidence for a common rhythmic pattern. *Music Perception: An Interdisciplinary Journal*, 19(3):333–349, 2002.

[14] Luis Jure and Martín Rocamora. Microtiming in the rhythmic structure of candombe drumming patterns. In *Fourth International Conference on Analytical Approaches to World Music, New York, USA*, pages 8–11, 2016.

[15] Luiz Naveda, Fabien Gouyon, Carlos Guedes, and Marc Leman. Microtiming patterns and interactions with musical properties in samba music. *Journal of New Music Research*, 40(3):225–238, 2011.

[16] Fabien Gouyon. Microtiming in "samba de roda"—preliminary experiments with polyphonic audio. *XII Simpósio da Sociedade Brasileira de Computação Musical São Paulo*, 2007.

[17] Leonardo O Nunes, Martín Rocamora, Luis Jure, and Luiz WP Biscainho. Beat and downbeat tracking based on rhythmic patterns applied to the uruguayan candombe drumming. In *ISMIR*, pages 264–270, 2015.

[18] Magdalena Fuentes, Lucas S Maia, Martín Rocamora, Luiz WP Biscainho, Hélene C Crayencour, Slim Essid, and Juan P Bello. Tracking beats and microtiming in afro-latin american music using conditional random fields and deep learning. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR, Delft, The Netherlands*, pages 251–258, 2019.

[19] Christian Dittmar, Martin Pfleiderer, Stefan Balke, and Meinard Müller. A swingogram representation for tracking micro-rhythmic variation in jazz performances. *Journal of New Music Research*, 47(2):97–113, 2018.

[20] Jean Laroche. Estimating tempo, swing and beat locations in audio recordings. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, pages 135–138. IEEE, 2001.

[21] Magdalena Fuentes. *Multi-scale computational rhythm analysis: a framework for sections, downbeats, beats, and microtiming*. PhD thesis, Université Paris-Saclay, 2019.

[22] Martín Rocamora, Luis Jure, Magdalena Fuentes, Lucas Maia, and Luiz Biscainho. Carat: computer-aided rhythmic analysis toolbox. In *20th Conference of the International Society for Music Information Retrieval, Delft, Netherlands, 4-8 Nov*. International Society for Music Information Retrieval., 2019.

[23] Florian Eyben, Sebastian Böck, Björn Schuller, and Alex Graves. Universal onset detection with bidirectional long-short term memory neural networks. In *Proc. 11th Intern. Soc. for Music Information Retrieval Conference, ISMIR, Utrecht, The Netherlands*, pages 589–594, 2010.

[24] Sebastian Böck, Andreas Arzt, Florian Krebs, and Markus Schedl. Online real-time onset detection with recurrent neural networks. In *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12), York, UK*, 2012.

[25] Sebastian Böck, Florian Krebs, and Markus Schedl. Evaluating the online capabilities of onset detection methods. In *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR, Porto, Portugal*, pages 49–54, 2012.

[26] A. W. Schloss. *On the Automatic Transcription of Percussive Music—From Acoustic Signal to High-Level Analysis*. PhD thesis, Stanford University, 1985.

[27] Juan Pablo Bello, Chris Duxbury, Mike Davies, and Mark Sandler. On the use of phase and energy for musical onset detection in the complex domain. *IEEE Signal Processing Letters*, 11(6):553–556, 2004.

[28] Masataka Goto and Yoichi Muraoka. Beat tracking based on multiple-agent architecture-a real-time beat tracking system for audio signals. In *Proceedings of the Second International Conference on Multiagent Systems*, pages 103–110, 1996.

[29] Paul Masri. *Computer modelling of sound for transformation and synthesis of musical signals.* PhD thesis, University of Bristol, 1996.

[30] Simon Dixon. Simple spectrum-based onset detection. *MIREX 2006*, page 62, 2006.

[31] Chris Duxbury, Mark Sandler, and Mike Davies. A hybrid approach to musical note onset detection. In *Proc. Digital Audio Effects Conf.(DAFX,'02)*, pages 33–38, 2002.

[32] Juan Pablo Bello and Mark Sandler. Phase-based note onset detection for music signals. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 5, pages V–441. IEEE, 2003.

[33] Ryan Stables, Jason Hockman, and Carl Southall. Automatic drum transcription using bi-directional recurrent neural networks. In *17th International Society for Music Information Retrieval Conference, ISMIR*. dblp, 2016.

[34] Richard Vogl, Matthias Dorfer, and Peter Knees. Recurrent neural networks for drum transcription. In *ISMIR*, pages 730–736, 2016.

[35] Simon Dixon. Onset detection revisited. In *Proceedings of the 9th International Conference on Digital Audio Effects*, volume 120, pages 133–137. Citeseer, 2006.

[36] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. mir_eval: A transparent implementation of common mir metrics. In *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer, 2014.

[37] Nick Collins. A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions. In *Audio Engineering Society Convention 118*. Audio Engineering Society, 2005.

[38] Jen-Tzung Chien. *Source Separation and Machine Learning*, chapter 1. Academic Press, 2018.

[39] E Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979, 1953.

[40] Ruolun Liu and Suping Li. A review on music source separation. In *2009 IEEE Youth Conference on Information, Computing and Telecommunication*, pages 343–346. IEEE, 2009.

[41] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, Derry FitzGerald, and Bryan Pardo. An overview of lead and accompaniment separation in music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(8):1307–1335, 2018.

[42] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep u-net convolutional networks. *18th International Society for Music Infor-mation Retrieval Conferenceng, Suzhou, China, 2017*, 2017.

[43] Derry Fitzgerald. Upmixing from mono-a source separation approach. In *2011 17th International Conference on Digital Signal Processing (DSP)*, pages 1–7. IEEE, 2011.

[44] Elias K Kokkinis, Joshua D Reiss, and John Mourjopoulos. A wiener filter approach to microphone leakage reduction in close-microphone applications. *IEEE transactions on audio, speech, and language processing*, 20(3):767–779, 2011.

[45] Emmanuel Vincent, Nancy Bertin, Rémi Gribonval, and Frédéric Bimbot. From blind to guided audio source separation: How models and side information can improve the separation of sound. *IEEE Signal Processing Magazine*, 31(3):107–115, 2014.

[46] Emmanuel Vincent, Shoko Araki, Fabian Theis, Guido Nolte, Pau Bofill, Hiroshi Sawada, Alexey Ozerov, Vikrham Gowreesunker, Dominik Lutter, and Ngoc QK Duong. The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges. *Signal Processing*, 92(8):1928–1936, 2012.

[47] Estefania Cano, Derry FitzGerald, Antoine Liutkus, Mark D Plumbley, and Fabian-Robert Stöter. Musical source separation: An introduction. *IEEE Signal Processing Magazine*, 36(1):31–40, 2018.

[48] Scott Rickard and Ozgiir Yilmaz. On the approximate w-disjoint orthogonality of speech. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–529. IEEE, 2002.

[49] Robert McAulay and Thomas Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):744–754, 1986.

[50] Antoine Liutkus and Roland Badeau. Generalized wiener filtering with fractional power spectrograms. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 266–270. IEEE, 2015.

[51] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[52] Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications, finland. *Neural Networks Research Centre, Helsinki University of Technology*, pages 1–3, 2000.

[53] Emmanuel Vincent. Musical source separation using time-frequency source priors. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):91–98, 2005.

[54] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[55] Sunnydayal Vanambathina. Intechopen - speech enhancement using an iterative posterior nmf. *Adapted from:* https://www.intechopen.com/online-first/speech-enhancement-using-an-iterative-posterior-nmf.

[56] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[57] Marko Helen and Tuomas Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *2005 13th European Signal Processing Conference*, pages 1–4. IEEE, 2005.

[58] Christian Dittmar and Daniel Gärtner. Real-time transcription and separation of drum recordings based on nmf decomposition. In *DAFx*, pages 187–194, 2014.

[59] Paris Smaragdis and Bhiksha Raj. Shift-invariant probabilistic latent component analysis. *Journal of Machine Learning Research - JMLR*, 01 2008.

[60] Paris Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *International Conference on Independent Component Analysis and Signal Separation*, pages 494–499. Springer, 2004.

[61] Stefan Uhlich, Franck Giron, and Yuki Mitsufuji. Deep neural network based instrument extraction from music. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139, 2015.

[62] Aditya Nugraha, Antoine Liutkus, and Emmanuel Vincent. Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24, 06 2015.

[63] Marius Miron. *Source separation methods for orchestral music: timbre-informed and score-informed strategies*. PhD thesis, Universitat Pompeu Fabra, 2018.

[64] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. The MUSDB18 corpus for music separation, December 2017.

[65] Rachel M Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. Medleydb: A multitrack dataset for annotation-intensive mir research. In *ISMIR*, volume 14, pages 155–160, 2014.

[66] Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji. Open-unmix - a reference implementation for music source separation. *Journal of Open Source Software*, 4:1667, 09 2019.

[67] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussalam. Spleeter: A fast and state-of-the art music source separation tool with pre-trained models. In *Proc. International Society for Music Information Retrieval Conference*, 2019.

[68] Emmanuel Vincent, Cédric Févotte, Rémi Gribonval, Laurent Benaroya, Xavier Rodet, Axel Röbel, Eric Le Carpentier, and Frédéric Bimbot. A tentative typology of audio source separation tasks. *in 4th Int. Symp. on Independent Component Analysisand Blind Signal Separation (ICA), 2003*, page 715–720, 2003.

[69] M. Rocamora, L. Jure, B. Marenco, M. Fuentes, F. Lanzaro, and A. Gómez. An audio-visual database of candombe performances for computational musicological studies. In *in Proceedings of the International Congress on Science and Music Technology 2015*, CICTeM, 2015.

[70] Sebastian Böck and Gerhard Widmer. Maximum filter vibrato suppression for onset detection. In *Proc. of the 16th Int. Conf. on Digital Audio Effects (DAFx). Maynooth, Ireland (Sept 2013)*, volume 7, 2013.

[71] Chris Cannam, Christian Landone, Mark B Sandler, and Juan Pablo Bello. The sonic visualiser: A visualisation platform for semantic descriptors from musical signals. In *ISMIR*, pages 324–327, 2006.

[72] Climéro de Oliveira Santos, Tarcísio Soares Resende, and Peter Malcom Keays. *Batuque Book: Maracatu Baque Virado e Baque Solto*. Recife: Edição do autor, 2009.

[73] Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer. madmom: a new Python Audio and Music Signal Processing Library. In *Proceedings of the 24th ACM International Conference on Multimedia*, pages 1174–1178, Amsterdam, The Netherlands, 10 2016.

[74] Matthew E. P. Davies and Sebastian Böck. Temporal convolutional networks for musical audio beat tracking. In *Proc. of the 27th European Signal Processing Conf.*, 2019.

[75] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka. Supervised and semi-supervised separation of sounds from single-channel mixtures. In *International Conference on Independent Component Analysis and Signal Separation*, pages 414–421. Springer, 2007.