HAL
open science

Tracking beats and microtiming in afro-latin american music using conditional random fields and deep learning

Magdalena Fuentes, Lucas S Maia, Martín Rocamora, Luiz W P Biscainho, Hélène C Crayencour, Slim Essid, Juan P. Bello

To cite this version:

Magdalena Fuentes, Lucas S Maia, Martín Rocamora, Luiz W P Biscainho, Hélène C Crayencour, et al.. Tracking beats and microtiming in afro-latin american music using conditional random fields and deep learning. ISMIR, Nov 2019, Delft, Netherlands. hal-02419361

HAL Id: hal-02419361

https://hal.science/hal-02419361v1

Submitted on 19 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

TRACKING BEATS AND MICROTIMING IN AFRO-LATIN AMERICAN MUSIC USING CONDITIONAL RANDOM FIELDS AND DEEP LEARNING

Magdalena Fuentes[1,2], **Lucas S. Maia**[2,3], **Martín Rocamora**[4], **Luiz W. P. Biscainho**[3]
Hélène C. Crayencour[1], **Slim Essid**[2], **Juan P. Bello**[5]
[1] L2S, CNRS–Univ.Paris-Sud–CentraleSupélec, France
[2] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[3] Federal University of Rio de Janeiro, Brazil
[4] Universidad de la República, Uruguay
[5] Music and Audio Research Laboratory, New York University, USA

ABSTRACT

Events in music frequently exhibit small-scale temporal deviations (microtiming), with respect to the underlying regular metrical grid. In some cases, as in music from the Afro-Latin American tradition, such deviations appear systematically, disclosing their structural importance in rhythmic and stylistic configuration. In this work we explore the idea of automatically and jointly tracking beats and microtiming in timekeeper instruments of Afro-Latin American music, in particular Brazilian *samba* and Uruguayan *candombe*. To that end, we propose a language model based on conditional random fields that integrates beat and onset likelihoods as observations. We derive those activations using deep neural networks and evaluate its performance on manually annotated data using a scheme adapted to this task. We assess our approach in controlled conditions suitable for these timekeeper instruments, and study the microtiming profiles' dependency on genre and performer, illustrating promising aspects of this technique towards a more comprehensive understanding of these music traditions.

1. INTRODUCTION

Across many different cultures, music is meter-based, i.e., it has a structured and hierarchical organization of pulsations. Within this metrical structure, the different pulsations interact with one another and produce the sensation of rhythm, inducing responses in the listeners such as foot tapping or hand clapping. In the so-called "Western" music tradition, that hierarchical structure often includes the beat and downbeat levels, where the former corresponds to the predominant perceived pulsation, and the latter has a longer time-span that groups several beats into bars. In some cases, the events in music present small-scale temporal deviations with respect to the underlying regular metrical grid, a phenomenon here referred to as microtiming. The interaction between microtiming deviations and other rhythmic dimensions contribute to what has been described as the sense of 'swing' or 'groove' [8, 9, 30]. The systematic use of these deviations is of structural importance in the rhythmic and stylistic configuration of many musical genres. This is the case of jazz [9, 10, 20, 38], Cuban *rumba* [2], Brazilian *samba* [32, 39] and Uruguayan *candombe* [22], among others. Consequently, the analysis of these music genres without considering microtiming leads to a limited understanding of their rhythm.

Samba and *candombe* are musical traditions from Brazil and Uruguay, respectively, that play a huge role in those countries' popular cultures. Both genres have deep African roots, partly evidenced by the fact that their rhythms result from the interaction of several rhythmic patterns played by large ensembles of characteristic percussive instruments. *Candombe* rhythm is structured in 4/4 meter, and is played on three types of drums of different sizes and pitches—*chico*, *repique* and *piano*—, each with a distinctive rhythmic pattern, the *chico* drum being the timekeeper. [1] *Samba* rhythm is structured in 2/4 meter, and comprises several types of instruments—*tamborim*, *pandeiro*, *chocalho*, *reco-reco*, *agogô*, and *surdo*, among others. Each instrument has a handful of distinct patterns [16], and more than one instrument may act as the timekeeper. Because of this combination of several timbres and pitches, the texture of a *samba* performance can become more complex than that of a *candombe* performance, where only three types of drums are present. Nevertheless, both rhythms have in common that they exhibit microtiming deviations at the sixteenth note level [22, 28, 32], with no deviations in the beat positions. [2] This is illustrated in Figure 1 for the recording of a *tamborim* playing in the *samba de enredo* style.

© Magdalena Fuentes, Lucas S. Maia, Martín Rocamora, Luiz W. P. Biscainho, Hélène C. Crayencour, Slim Essid, Juan P. Bello. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Magdalena Fuentes, Lucas S. Maia, Martín Rocamora, Luiz W. P. Biscainho, Hélène C. Crayencour, Slim Essid, Juan P. Bello. "Tracking Beats and Microtiming in Afro-Latin American Music Using Conditional Random Fields and Deep Learning", 20th International Society for Music Information Retrieval Conference, Delft, The Netherlands, 2019.

[1] In this musical context, the role of timekeeper is assigned to an instrument that plays an invariable rhythmic pattern (i.e., an *ostinato*) usually at a high rate, thus defining the subdivision of the beat.

[2] In other musical forms, such as waltz, microtiming may be mostly on beats.

1.1 Related Work

Microtiming has been studied in the context of Music Information Retrieval (MIR) for many years [2, 8, 17]. Besides the interest in characterizing microtiming for musicological studies, it is important for music synthesis applications, since it is a central component for "humanizing" computer generated performances [18]. Depending on the musical context, microtiming can take the form of tempo variations, like *rubato* or *accelerando*, or small-scale deviations of events with respect to an underlying regular metrical grid [22]. Therefore, in order to study microtiming deviations one has to know the expected position of the events within the metrical grid and the actual articulated positions—which can be inferred from information related to the onset positions, the tempo and/or the beats.

Most of the proposed methods for microtiming analysis are based on manually annotated data. Laroche et al. [27] proposed a method for the joint estimation of tempo, beat positions and swing in an ad-hoc fashion. The proposal exploits some simplifications: assuming constant tempo and swing ratio, and propagating beat positions based on the most likely position of the first beat. More recent works perform semi-automatic analysis still relying on informed tempo [9, 10], or using an external algorithm for its estimation [30]. Within the context of *candombe* and *samba*, microtiming characterization has also been addressed using either semi-automatic or heuristic methods [17, 22, 32].

In other rhythm-related MIR tasks such as beat and downbeat tracking, graphical models (GM) such as hidden Markov models or dynamic Bayesian networks are widely used [4, 19, 25, 36]. GMs are capable of encoding musical knowledge in a flexible and unified manner, providing structure to the estimations and usually a gain in performance for different models across genres [14]. In particular, Conditional Random Fields (CRFs) are discriminative undirected GMs for structured data prediction [37]. CRFs relax some conditional independence assumptions of Bayesian Networks, which allows for modeling complex and more general dependency structures, thus making them appealing for music modeling. CRFs have been applied in MIR tasks such as beat tracking [13] or audio-to-score alignment [21], and have been successfully combined with deep neural networks (DNNs) [11, 15, 24].

1.2 Our Contributions

This work takes a first step towards fully-automatic tracking of beats and microtiming deviations in a single formalism, applied to the analysis of two (usually underrepresented) Afro-Latin American music genres, namely Brazilian *samba* and Uruguayan *candombe*. More precisely, we introduce a CRF model that uses beat and onset activations derived from deep learning models as observations, and combines them to jointly track beats and microtiming profiles within rhythmic patterns at the sixteenth note level. To the best of our knowledge, this is the first work that explores the use of CRFs for tracking microtiming and beats jointly. This temporal granularity is in accordance with the type of microtiming deviations present in the music traditions under study. Following previous works [9], we derive microtiming labels from annotated onsets and use them to evaluate the proposed system, attaining promising results towards more holistic and descriptive models for rhythm analysis. We also study the usefulness of this approach in controlled conditions, as a first assessment of its capabilities. We explore our microtiming representation in some applications, namely the extraction of microtiming profiles of certain instruments, and the study of differences between musical genres based on their microtiming traits.

2. PROPOSED METHOD

2.1 Language Model

The proposed language model consists of a linear-chain CRF [26, 37]. Formally, the conditional probability of a label sequence $\mathbf{y} = (y_1, ..., y_T)$ of length T given an input sequence of observations $\mathbf{x} = (x_1, ..., x_T)$ is given by:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \psi(y_t, y_{t-1})\, \phi(y_t, x_t), \qquad (1)$$

where ψ is the transition potential and ϕ is the observation potential. They play a role similar to transition and observation probabilities in dynamic Bayesian networks or hidden Markov models, with the difference that the potentials in a CRF do not need to be proper probabilities, hence the need for the normalization factor $Z(\mathbf{x})$.

In our model, depicted in Figure 2, the output labels $\mathbf{y}$ are a function of three variables that describe the position inside the beat, the length of the beat interval in frames, and the microtiming within the beat-length pattern at the sixteenth note level. Formally,

$$y_t := (f_t, l_t, m_t), \qquad (2)$$

where f_t is the frame counter with $f_t \in \mathcal{F} = \{1, ..., l_t\}$, $l_t \in \mathcal{L} = \{l_{\min}, ..., l_{\max}\}$ is the number of frames per beat, which relates to the tempo of the piece; and the microtiming $\mathbf{m}_t \in \mathcal{M} = \{\mathbf{m}_1, ..., \mathbf{m}_N\}$. The observations $\mathbf{x}$ are based on estimated beat and onset likelihoods, as detailed later. The problem of obtaining the beat positions and microtiming profiles is then formulated as finding the sequence of labels $\mathbf{y}^*$ such that $\mathbf{y}^* = \arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x})$.

2.1.1 Microtiming Tracking

Both in *samba* and *candombe*, timekeeper instruments usually play a beat-length rhythmic pattern that articulates several sixteenth notes [16], as shown in Figures 1 and 3 for the *tamborim*. In order to provide a common framework for comparing both music genres, we focus our study on the microtiming deviations of beat-length rhythmic patterns articulated by timekeeper instruments in groups of four sixteenth notes. [3] To that end, we consider the following hypothesis, which we explain further below:

- The tempo is constant within a beat.

[3] Note that minor adjustments to the proposed model allow for the tracking of microtiming deviations in other kinds of rhythmic patterns.

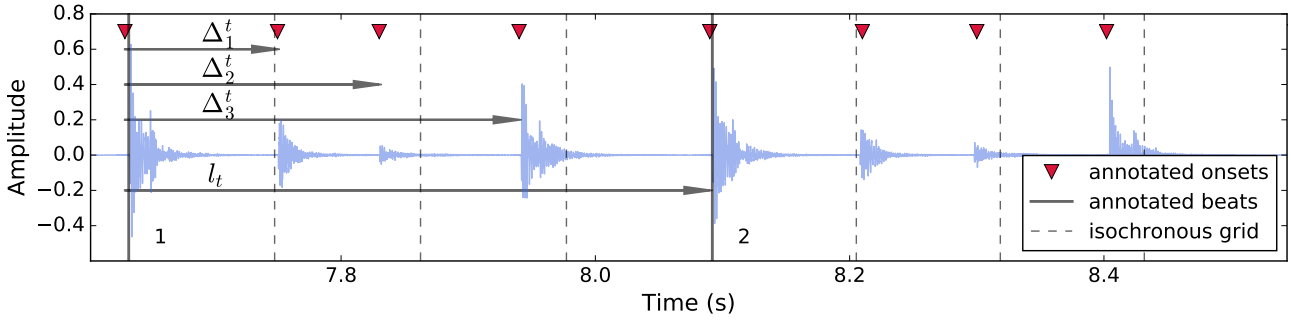

Figure 1. Example of microtiming deviations at the sixteenth note level for a beat-length rhythmic pattern from the *tamborim* in *samba de enredo*.

- The microtiming profile changes smoothly only at beat transitions.
- The tempo is between 120 and 135 BPM, to ensure an appropriate temporal resolution.

We define the microtiming descriptor $\mathbf{m}$ at frame t as:

$$\mathbf{m}_t := (m_t^1, m_t^2, m_t^3)$$

where $m_t^i := \frac{\Delta_t^i}{l_t} \in [\frac{i}{4}+\delta_L^i, \frac{i}{4}+\delta_U^i]$, and Δ_t^i is the distance in frames between an articulated sixteenth note and the beginning of the beat interval, as shown in Figure 1. Thus, each m_t^i models the position of the i-th sixteenth note with respect to the beginning of the beat, relative to the total beat length. For instance, the value of the microtiming descriptor for a rhythmic pattern of four isochronous sixteenth notes, i.e., located exactly on an equally-spaced metrical grid, is $\mathbf{m} = (0.25, 0.50, 0.75)$, indicating the articulation of events at $1/4, 1/2$ and $3/4$ of the beat interval respectively. To account for different microtiming profiles, the value of m_t^i is estimated within an interval determined by lower and upper deviations bounds, δ_L^i and δ_U^i, modeled as positive or negative percentages of the beat interval length. The proposed microtiming descriptor provides an intuitive idea of how the articulated sixteenth notes deviate within the rhythmic pattern from their isochronous expected positions. It is independent of tempo changes, since it is normalized by the estimated beat interval length, allowing for studies on microtiming–tempo dependencies.

The definition of the microtiming descriptor $\mathbf{m}_t$ can be related to the *swing-ratio*, s, proposed in previous work [9,30], though the two differ in various aspects. The *swing-ratio* is defined in terms of the inter-onset intervals (IOIs) of a long–short rhythmic pattern, such that $s \geq 1$ is the ratio between the *onbeat* IOI (longer interval) and the off-beat IOI (shorter interval). In contrast, the $\mathbf{m}_t$ descriptor is composed by three *microtiming–ratios*, m_t^i, whose IOIs are defined with respect to the beginning of the beat instead of the previous onset as in [9, 30]. However, it is possible to convert the model proposed here into the *swing-ratio* by redefining $\mathbf{m}$ as $m_s := m_t^1$, and then, from the estimated m_s, computing $s = \frac{m_t^1}{1-m_t^1}$. With such modifications, the model could be applied to the studies presented in [9, 30].

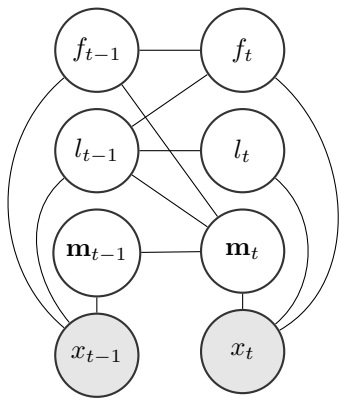

Figure 2. CRF graph. Observations and labels are indicated as gray and white nodes respectively.

2.1.2 *Transition Potential* ψ

The transition potential is given in terms of f_t, l_t, and m_t (see Equation 2) by:

$$\psi(y_t, y_{t-1}) := \psi_f(f_t, f_{t-1}, l_t, l_{t-1})\psi_m(\mathbf{m}_t, \mathbf{m}_{t-1}, f_{t-1}, l_{t-1})$$

Similar to [13,25], we force frame counter f_t to increase by one, at each step, up to the maximum beat length considered, and to switch to one at the end of the beat. Beat duration changes are unlikely (i.e., tempo changes are rare) and only allowed at the end of the beat. We constrain these changes to be smooth, giving inertia to tempo transitions. Those rules are formally expressed by:

$$\psi_f(f_t, f_{t-1}, l_t, l_{t-1}) := \begin{cases} 1 & \text{if } f_t = (f_{t-1} \bmod l_{t-1}) + 1, \\ & f_{t-1} \neq l_{t-1} \\ 1 - p_f & \text{if } l_t = l_{t-1}, \\ & f_t = 1,\ f_{t-1} = l_{t-1} \\ \frac{p_f}{2} & \text{if } l_t = l_{t-1} \pm 1,\ f_t = 1 \\ 0 & \text{otherwise} \end{cases}$$

The microtiming descriptor m_t changes smoothly and only at the end of the beat, that is:

$$\psi_m(\mathbf{m}_t, \mathbf{m}_{t-1}, f_{t-1}, l_{t-1}) := \begin{cases} 1 & \text{if } \mathbf{m}_t = \mathbf{m}_{t-1}, \\ & f_{t-1} \neq l_{t-1} \\ 1 - p_m & \text{if } \mathbf{m}_t = \mathbf{m}_{t-1}, \\ & f_{t-1} = l_{t-1} \\ \frac{p_m}{2} & \text{if } m_i^t = m_i^{t-1} \pm 0.02\ \forall i, \\ & f_{t-1} = l_{t-1} \\ 0 & \text{otherwise} \end{cases}$$

In the transition potential, p_f and p_m represent the probability of changing the beat interval length (i.e., tempo) and the probability of changing the microtiming profile at the end of the beat, respectively. The values of $1-p_f$ and $p_f/2$ were chosen following previous works, whereas $1-p_m$ and $p_m/2$ were similarly set in order to make the possible microtiming transitions equally likely.

Since m_t^i is given in percentage with respect to the inter-beat-interval (IBI), the resolution with which microtiming can be estimated in the model is also percentual, and it is given by the relation between the sampling rate SR of the features and the BPM: $\mathrm{res} = \frac{\mathrm{BPM}}{60\mathrm{SR}}$. It has been shown in the literature that a resolution of 0.02 of the IBI is sufficient for representing microtiming deviations [17, 32]. To keep computational complexity low but at the same time guaranteeing a resolution $\mathrm{res} = 0.02$, we use observation features sampled at 110 Hz and we study pieces whose tempo is within range of 120 to 135 BPM. Note that these assumptions are valid in the music under study, and they could be adapted to a different music genre, e.g. increasing sampling rate to increase the BPM interval.

2.1.3 Observation Potential ϕ

The observation potential depends on the beat and onset likelihoods, the frame counter f_t and the microtiming m_t:

$$\phi(f_t, \mathbf{m}_t, x_t) := \begin{cases} b_t & \text{if } f_t = 1 \\ o_t - b_t & \text{if } \frac{f_t}{l_t} \in \mathbf{m}_t \\ 1 - o_t & \text{otherwise} \end{cases}$$

where b_t and o_t are beat and onset likelihoods, respectively. The onset likelihood was estimated using the ensemble of recurrent neural networks for onset activation estimation from *madmom* [3]—we refer the interested reader to [5, 12] for further information. We designed a simple DNN for the beat likelihood estimation and trained it on *candombe* and *samba*.[4] It consists of 6 layers, namely: batch normalization, dropout of 0.4, bidirectional gated recurrent unit (Bi-GRU) [6] with 128 units, batch normalization, another identical Bi-GRU layer, and a dense layer with two units and a softmax activation.

We use a mel-spectrogram as input feature for the DNN. The short–time Fourier transform is computed using a window length of 2048 samples and a hop of 401 samples, to ensure a sampling rate of 110 Hz with audio sampled at 44.1 kHz. We use 80 mel filters, comprising a frequency range from 30 Hz to 17 kHz.

3. DATASETS

In our experiments we use a subset of the *candombe* dataset [35] and the BRID dataset [29] of Brazilian *samba*.

candombe dataset: it comprises audio recordings of Uruguayan *candombe* drumming performances in which ensembles of three to five musicians play the three different *candombe* drums: *chico*, *piano* and *repique*. It has separated stems of the different drums, which facilitates the microtiming analysis. We focus our study on the *chico* drum, which is the timekeeper of the ensemble. We select a subset of the recordings in the dataset, in which the *chico* drum plays a beat-length pattern of four sixteenth notes, for a total of 1788 beats and 7152 onsets.

BRID dataset: it consists of both solo and ensemble performances of Brazilian *samba*, comprising ten different instrument classes: *agogô*, *caixa* (snare drum), *cuíca*, *pandeiro* (frame drum), *reco-reco*, *repique*, *shaker*, *surdo*, *tamborim* and *tantã*. We focus our study on the *tamborim*, which is one of the timekeepers of the ensemble. We select a subset of the solo tracks, in which the *tamborim* plays a beat–length rhythmic pattern of four sixteenth notes (shown in music notation[5] in Figure 3), for a total of 396 beats and 1584 onsets.

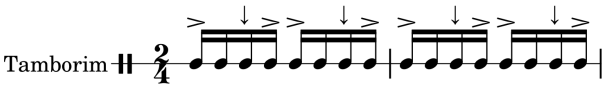

Figure 3. Example of the beat-length rhythmic pattern of the *tamborim* from the *samba* dataset in music notation.

3.1 Ground-Truth Generation

The microtiming ground-truth is inferred following the approach of [9], in which the onsets are used to derive the swing-ratio annotations. Analogously, we compute the microtiming ground-truth using the annotated onsets, obtaining one value of $\mathbf{m} = (m_1, m_2, m_3)$ for each beat. In order to mitigate the effect of onset annotation errors and sensori-motor noise, we use a moving-median filter to smooth the microtiming ground-truth, with a centered rectangular window of length 21 beats, as shown in Figure 4.

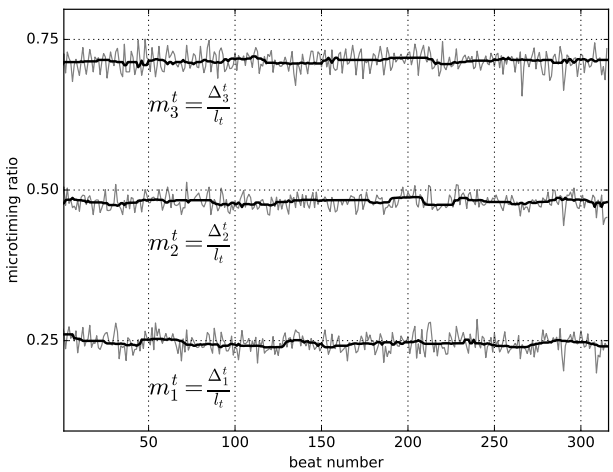

Figure 4. Example of the microtiming values for a *chico* drum recording in the *candombe* dataset. Dark and light lines represent the ground-truth with and without median filtering, respectively.

[4] The training proved necessary because the timekeeper pattern of *candombe* rhythm has a distinctive accent displaced with respect to the beat that misleads beat–tracking models trained on "Western" music [34].

[5] The symbol '>' refers to an accent, and '↓' implies to turn the *tamborim* upside down and execute the strike backwards.

4. EXPERIMENTS

4.1 Experimental Setup

We investigate the performance of the model and whether the microtiming descriptor is useful for analyzing the music at hand. To that end, we scale the *candombe* dataset to match the size of the *samba* dataset at test time by selecting excerpts in each track. We assess the model's performance using manually annotated onsets and beats, from which we derive our ground-truth as explained in Section 3.1. To evaluate if the microtiming estimation affects the beat tracking, we compare the model's performance with a simplified version of it that only tracks beats. This version has only variables f_t and l_t (see Figure 2); the same potential ψ_f is used, and the observation potential is simply b_t (the beat likelihood) at beat positions and $1 - b_t$ otherwise. We assess the microtiming estimation by: varying the p_m microtiming transition parameter—allowing smooth changes within the piece or no changes at all; and varying the tolerance used on the F-measure (F1) score. Finally, we discuss our main findings on the potential of jointly tracking beats and microtiming.

4.1.1 Implementation, Training and Evaluation Metrics

The DNN beat likelihood model is implemented in *Keras* 2.2.4 and *Tensorflow* 1.13.1 [1, 7]. We use the *Adam* optimizer [23] with default parameters. Training is stopped after 10 epochs without improvement in the validation loss, to a maximum of 100 epochs. We train the network with patches of 500 frames and a batch size of 64, leaving one track out and training with the rest, which we split in 30% and 70% for validation and training respectively, among the same genre. The onset activation was obtained with *madmom* version 0.16.1 [3], and the mel-spectra was computed using *librosa* 0.6.3 [31].

We evaluate the model using the F1 score for beat tracking with a tolerance window of 70 ms, as implemented in *mir_eval* 0.5 [33]. To evaluate the microtiming estimation, we first select the correctly estimated beats, then compute F1 for each estimated m_t^i with tolerance windows of different lengths, and the overall score as the mean F1 ($\text{F1}_{\mathbf{m}_t} = \sum_i \text{F1}_{m_t^i}/3$).

4.2 Results and Discussion

The results on the microtiming tracking are depicted in Figure 5, which shows the F1 scores as a function of the tolerance. The different colors represent the different p_m values. We evaluate the model for the set of values $p_m = \{0, 0.001, 0.06\}$, that is no, very unlikely and more likely microtiming changes respectively. Those values were obtained from statistics on the data in preliminary experiments. We searched for microtiming ratios within the interval $[0.25, 0.29] \times [0.42, 0.5] \times [0.67, 0.75]$, for microtiming dimensions $i = 1, 2, 3$ respectively. This corresponds to $\boldsymbol{\delta}_L = (0, -0.03, -0.08)$ and $\boldsymbol{\delta}_U = (0.04, 0, 0)$.[6] As illustrated in Figure 5, we found that the restriction of a constant microtiming profile ($p_m = 0$) along the piece relates to a worse performance, specially with small tolerances. We hypothesize that this occurs because in some *samba* excerpts the microtiming ratio changes are percentually bigger than those tolerances, leading to an inaccurate estimation. From considering the dependency of the F1 score with respect to the tolerance, we observe that is possible to achieve a reasonable F1 score from 0.025 on. The results with the best compromise in terms of variance and median are achieved with $p_m = 0.001$, which aligns with the hypothesis that microtiming profiles change very smoothly over time. We explored different tolerances since we are working with frames which are noisy, and the comparison with the smoothed ground truth still makes sense with large tolerances.

[6] Symmetric windows around the isochronus sixteenth note positions were used for the microtiming ratios in preliminary experiments with no gain in tracking performance and a higher computational burden.

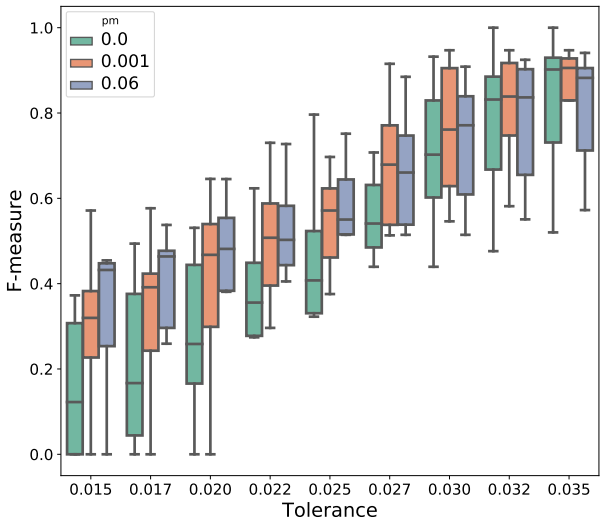

Figure 5. Mean microtiming F-measure score on the two datasets.

We found that the beat tracking performance of the model reaches a 95.7% F1 score, being equivalent to the beat tracking only version. The high F1 score in beat tracking is not surprising given that the DNN was trained using data of the same nature (acoustic conditions and genre) and the sets are homogeneous. As mentioned before, state-of-the-art beat tracking systems based on DNNs fail dramatically in this specific scenario [34], particularly tracking the beats in time-keeper instruments in *candombe*, because the data is too different from what was used in their training. We do not consider this as a challenging beat tracking case, but a training stage was needed to perform adequately.

During our experiments we observed that the microtiming descriptor $\mathbf{m}_t$ could be used to help beat tracking in some cases. Informing the microtiming profile a priori, by setting δ_L^i and δ_U^i, can disambiguate beat positions by helping the joint inference. This could allow to apply non pre-trained beat tracking models to *candombe* recordings, which usually fail in estimating the beat location by displacing it one sixteenth note (due to an accent in the rhythmic pattern). Aligned to that idea, the model could be use-

ful in scenarios where onsets from other instruments are present. Besides, when the beat tracking is incorrect, the obtained microtiming profile can be descriptive of the type of mistake that occurs by contrasting the obtained profile with the expected one. The same case mentioned before—a lag of a sixteenth note in the beat estimation—shows in the microtiming estimation as unexpected forward positions in the second and third sixteenth notes, with a synchronous fourth one, which is the *candombe* microtiming profile lagged by a sixteenth note position.

Microtiming description and insights: Figure 4 illustrates an example of microtiming profile for an excerpt of the *candombe* dataset. This example shows the microtiming variations per beat interval along the complete recording. In the performance of the example, the rhythmic pattern is played with the same microtiming profile in the whole track. This microtiming template is characteristic of some patterns of *candombe* drumming [34], and it is present in several recordings of the dataset. We noticed that microtiming profiles do not present significant variations within tempo changes in the *candombe* dataset. However, the presented method can be used to characterize curves of microtiming vs. tempo that could be informative of musical phenomena for other music genres or datasets.

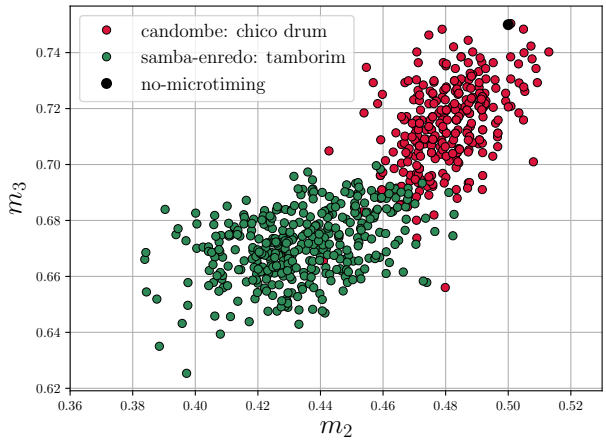

Figure 6. Microtiming distribution depending on style, view of the plane $m_t^2 m_t^3$, denoted as $m_2 m_3$ for simplicity. A dot at $(0.50, 0.75)$ indicates the expected position of a beat that shows no microtiming, that is, where all onsets are evenly spaced.

As shown in Figures 6 and 7, the microtiming descriptor $\mathbf{m}_t$ encodes musical features that are informative about the music genre, instrument type or performer. These two figures show the microtiming profile for all beats from *tamborim* and *chico* recordings, using the annotations for better visualization. Firstly, by observing the 'no-microtiming' reference in the figures that corresponds to $\mathbf{m}_t = (0.25, 0.50, 0.75)$, it becomes clear that both *samba* and *candombe* present considerable microtiming deviations in their time-keeper instruments. Even though the rhythmic patterns from both instruments present deviations that tend to compress the IOI in a similar manner, the microtiming profile differs for each music style, being more drastic in the case of the *tamborim*. This analysis should be extended to other *samba* instruments in order to determine if differences are due to the rhythmic pattern of a particular instrument; or if different patterns within the same genre tend to follow the same microtiming profile (characteristic of the genre). Figure 7 shows the microtiming profiles of each performer. It is quite clear that performers tend to be consistent with their microtiming, opening the perspective of studying microtiming profiles for performer characterization, as was done for jazz [9].

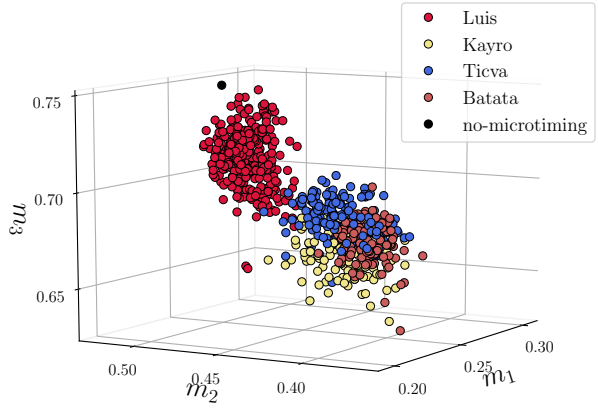

Figure 7. Microtiming distribution depending on performer (top musician plays *candombe* and the others play *samba*). A dot at $(0.25, 0.50, 0.75)$ indicates the point of no microtiming.

5. CONCLUSIONS AND FUTURE WORK

In this work, we introduced a language model that performs automatic tracking of beats and microtiming deviations in a single formalism. We applied this model to Afro-Latin American music, particularly Brazilian *samba* and Uruguayan *candombe*, and we focused our study on beat–length rhythmic patterns of timekeeper instruments, with four articulated sixteenth notes. The promising results we obtained with our method using a ground-truth derived from annotated onsets indicate it can facilitate automatic studies of these rhythms. This work intends to take a further step towards holistic systems that produce consistent and coherent estimations of music content.

As future work, we plan to extend our model to describe the microtiming profile depending on the nature of the rhythmic pattern being played, i.e., whether they articulate 2, 3, 4 or more notes, and to explore the usefulness of our model in challenging scenarios in comparison to heuristic methods.

6. ACKNOWLEDGEMENTS

This work was partially funded by CAPES, CNPq, ANII, CNRS, and STIC-AmSud program project 18-STIC-08. The authors would like to thank the reviewers for their valuable feedback.

7. REFERENCES

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] J. A. Bilmes. *Timing is of the essence: Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm*. Master's thesis, Massachusetts Institute of Technology, Cambridge, USA, 1993.

[3] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer. madmom: a new python audio and music signal processing library. In *24th ACM International Conference on Multimedia*, pages 1174–1178, Amsterdam, The Netherlands, October 2016.

[4] S. Böck, F. Krebs, and G. Widmer. Joint beat and downbeat tracking with recurrent neural networks. In *17th Int. Society for Music Information Retrieval Conf. (ISMIR)*, pages 255–261, New York, USA, August 2016.

[5] S. Böck, A. Arzt, F. Krebs, and M. Schedl. Online real-time onset detection with recurrent neural networks. In *15th Int. Conf. on Digital Audio Effects (DAFx)*, York, UK, September 2012.

[6] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[7] F. Chollet. Keras. `https://github.com/fchollet/keras`, 2015.

[8] M. E. P. Davies, G. Madison, P. Silva, and F. Gouyon. The effect of microtiming deviations on the perception of groove in short rhythms. *Music Perception*, 30(5):497–510, June 2013.

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

[10] C. Dittmar, M. Pfleiderer, and M. Müller. Automated estimation of ride cymbal swing ratios in jazz recordings. In *16th Int. Society for Music Information Retrieval Conf. (ISMIR)*, pages 271–277, Málaga, Spain, October 2015.

[11] S. Durand and S. Essid. Downbeat detection with conditional random fields and deep learned features. In *17th Int. Society for Music Information Retrieval Conf. (ISMIR)*, pages 386–392, New York, USA, August 2016.

[12] F. Eyben, S. Böck, B. Schuller, and A. Graves. Universal onset detection with bidirectional long-short term memory neural networks. In *11th Int. Society for Music Information Retrieval Conf. (ISMIR)*, pages 589–594, Utrecht, The Netherlands, August 2010.

[13] T. Fillon, C. Joder, S. Durand, and S. Essid. A conditional random field system for beat tracking. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 424–428, South Brisbane, Australia, April 2015.

[14] M. Fuentes, B. McFee, H. Crayencour, S. Essid, and J. Bello. Analysis of common design choices in deep learning systems for downbeat tracking. In *19th Int. Society for Music Information Retrieval Conf. (ISMIR)*, pages 106–112, Paris, France, September 2018.

[15] M. Fuentes, B. McFee, H. Crayencour, S. Essid, and J. Bello. A music structure informed downbeat tracking system using skip-chain conditional random fields and deep learning. In *44th Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 481–485, Brighton, UK, May 2019.

[16] G. Gonçalves and O. Costa. *The Carioca Groove: The Rio de Janeiro's Samba Schools Drum Sections*. Groove, Rio de Janeiro, Brazil, 2000.

[17] F. Gouyon. Microtiming in 'Samba de Roda' — preliminary experiments with polyphonic audio. In *11th Brazilian Symposium on Computer Music (SBCM)*, pages 197–203, São Paulo, Brazil, September 2007.

[18] H. Hennig, R. Fleischmann, A. Fredebohm, Y. Hagmayer, J. Nagler, A. Witt, F. J. Theis, and T. Geisel. The nature and perception of fluctuations in human musical rhythms. *PloS one*, 6(10):e26457, October 2011.

[19] A. Holzapfel, F. Krebs, and A. Srinivasamurthy. Tracking the 'odd': Meter inference in a culturally diverse music corpus. In *15th Int. Society for Music Information Retrieval Conf. (ISMIR)*, pages 425–430, Taipei, Taiwan, October 2014.

[20] V. Iyer. Embodied mind, situated cognition, and expressive microtiming in african-american music. *Music Perception*, 19(3):387–414, 2002.

[21] C. Joder, S. Essid, and G. Richard. A conditional random field framework for robust and scalable audio-to-score matching. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 19(8):2385–2397, November 2011.

[22] L. Jure and M. Rocamora. Microtiming in the rhythmic structure of Candombe drumming patterns. In *4th Int. Conf. on Analytical Approaches to World Music (AAWM)*, New York, USA, June 2016.

[23] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] F. Korzeniowski, S. Böck, and G. Widmer. Probabilistic extraction of beat positions from a beat activation function. In *15th Int. Society for Music Information Retrieval Conf. (ISMIR)*, pages 513–518, Taipei, Taiwan, October 2014.

[25] F. Krebs, S. Böck, M. Dorfer, and G. Widmer. Downbeat tracking using beat synchronous features with recurrent neural networks. In *17th Int. Society for Music Information Retrieval Conf. (ISMIR)*, pages 129–135, New York, USA, August 2016.

[26] J. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289, June 2001.

[27] J. Laroche. Estimating tempo, swing and beat locations in audio recordings. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 135–138, New Paltz, USA, October 2001.

[28] K. A. Lindsay and P. R. Nordquist. More than a feeling: Some technical details of swing rhythm in music. *Acoustics Today*, 3(3):31–42, July 2007.

[29] L. S. Maia, P. D. T. Jr., M. Fuentes, M. Rocamora, L. W. P. Biscainho, M. V. M. Costa, and S. Cohen. A novel dataset of Brazilian rhythmic instruments and some experiments in computational rhythm analysis. In *2018 AES Latin American Congress of Audio Engineering (AES LAC)*, pages 53–60, Montevideo, Uruguay, September 2018.

[30] U. Marchand and G. Peeters. Swing ratio estimation. In *18th Int. Conf. on Digital Audio Effects (DAFx)*, pages 423–428, Trondheim, Norway, December 2015.

[31] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenbergk, and O. Nieto. librosa: Audio and music signal analysis in Python. In *14th Python in Science Conf. (SciPy)*, pages 18–24, Austin, USA, July 2015.

[32] L. Naveda, F. Gouyon, C. Guedes, and M. Leman. Microtiming patterns and interactions with musical properties in samba music. *Journal of New Music Research*, 40(3):225–238, 2011.

[33] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis. mir_eval: A transparent implementation of common MIR metrics. In *15th Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, pages 367–372, Taipei, Taiwan, October 2014.

[34] M. Rocamora. *Computational methods for percussion music analysis : The Afro-Uruguayan Candombe drumming as a case study*. PhD thesis, Universidad de la República (Uruguay). Facultad de Ingeniería. IIE, April 2018.

[35] 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 *II Congreso Int. de Ciencia y Tecnología Musical (CICTeM)*, pages 17–24, Buenos Aires, Argentina, September 2015.

[36] A. Srinivasamurthy, A. Holzapfel, A. T. Cemgil, and X. Serra. A generalized bayesian model for tracking long metrical cycles in acoustic music signals. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 76–80, Shanghai, China, March 2016.

[37] C. Sutton and A. McCallum. An introduction to conditional random fields for relational learning. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*, chapter 4, pages 93–128. MIT Press, Cambridge, USA, 2006.

[38] C. H. Waadeland. "It dont mean a thing if it aint got that swing"—Simulating expressive timing by modulated movements. *Journal of New Music Research*, 30(1):23–37, 2001.

[39] M. Wright and E. Berdahl. Towards machine learning of expressive microtiming in Brazilian drumming. In *ICMC*. Citeseer, 2006.