Modeling Onset Spectral Features for Discrimination of Drum Sounds

Martín Rocamora[1(✉)] and Luiz W.P. Biscainho[2]

[1] Universidad de la República, Montevideo, Uruguay
rocamora@fing.edu.uy
[2] Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
wagner@smt.ufrj.br

Abstract. Motivated by practical problems related to ongoing research on Candombe drumming (a popular afro-rooted rhythm from Uruguay), this paper proposes an approach for recognizing drum sounds in audio signals that models for sound classification the same audio spectral features employed in onset detection. Among the reported experiments involving recordings of real performances, one aims at finding the predominant Candombe drum heard in an audio file, while the other attempts to identify those temporal segments within a performance when a given sound pattern is played. The attained results are promising and suggest many ideas for future research.

Keywords: Audio signal processing · Machine learning applications · Musical instrument recognition · Percussion music · Candombe drumming

1 Introduction

The extraction of musically meaningful content information via automatic analysis of audio recordings has become an important research field in audio signal processing. It encompasses a wide scope of applications, ranging from computer-aided musicology to automatic music transcription and recommendation. Research on automatic music transcription has concentrated mainly on pitched instruments, and only in the past decade percussion instruments have gained interest, most of the work focusing on the standard pop/rock drum kit [4]. The striking of a drum membrane produces a very short waveform that can be modeled as an impulsive function with broad-band spectrum [5], whose accurate characterization and analysis is a challenging problem in signal processing. In this context, the goal of automatic transcription is to determine the type of percussion instrument played (instrument recognition) and the temporal location of the event. Even if the problem of isolated sound classification is widely studied [7], the performance of available methods largely decreases when simultaneous sounds and real performances are considered [11].

This work was supported by CAPES and CNPq from Brazil, and ANII from Uruguay.

A. Pardo and J. Kittler (Eds.): CIARP 2015, LNCS 9423, pp. 100–107, 2015.
DOI: 10.1007/978-3-319-25751-8_13

Existing approaches for percussion transcription can be roughly divided into two types [4]. Most of the proposed solutions apply a pattern recognition approach to sound events. Firstly the audio signal is segmented into meaningful events, either by detecting onsets or by building a pulse grid. Then, audio features are computed for each segment, usually to describe spectral content and its temporal evolution [4, 7]. Finally, the segments are classified using pattern recognition methods. The other usual approach is based on segregating the audio input into streams which supposedly contain events from a single percussion sound class, by means of signal separation techniques [6]. After that, a class is assigned and an onset detection procedure is applied to each stream. Other distinctions can be made, such as if the classification is supervised or not, and whether it takes high-level musicological information into account [4].

In this paper, automatic percussion instrument recognition addresses audio files in which a predominant instrument suffers the interference from some others, aiming to determine the prevailing one. This type of audio file could be either the result of a signal separation technique as previously described, or coming from a microphone placed close to an instrument when a poly-instrument performance is recorded. The latter situation is common practice in some music productions or musicological field studies [10], and is the case of the dataset considered in the reported experiments.

The present work is part of an interdisciplinary collaboration that pursues the development of automatic tools for computer-aided analysis and transcription of Candombe drumming, one of the most defining traits of popular culture in Uruguay. Part of a tradition that has its roots in the culture brought by the African slaves in the 18th century, it evolved during a long historical process and is nowadays practiced by thousands of people and influenced various genres of popular music [1]. The rhythm is produced by groups of people marching in the streets playing drums [3]. There are three drum sizes (see Fig. 1) with respective registers: *chico* (small/high), *piano* (big/low) and *repique* (medium). The minimal ensemble of drums must have at least one of each type. All the drums are played with one hand hitting the drumhead bare and the other holding a stick. The stick is also used to hit the wooden shell of the drum, producing a sound called *madera*, when playing the *clave* pattern. This pattern serves as a mean of temporal organization and synchronization, and is played by all the drums before the rhythm patterns are initiated, and also by the *repique* drum in between phrases. A *repique* performance can also include occasional *madera* sounds as part of the repertoire of strokes used when improvising.

Two types of experiments are conducted in this work, one aiming to recognize the predominant Candombe drum in an audio file, and the other attempting to identify those temporal segments of a *repique* performance when the *clave* pattern is played. The classification is addressed by modeling the same audio features used for onset detection.

The remaining of the document is organized as follows. The dataset of audio recordings is introduced in the next section. Then, Section 3 is devoted to the extraction of audio features. The clustering and classification methods applied

are described in Sections 4 and 5 respectively. Experiments and results are presented in Section 6. The paper ends with some critical discussion of the present work and ideas for future research.

2 Datasets

A training dataset containing isolated sounds of Candombe drums was compiled and annotated for this work. To this end, a studio recording session was conducted in which five percussionists played in turns one among a set of three drums (one of each type) called `drums-1` hereafter. Automatic onset detection was performed over each audio track, and the resulting events were manually checked and labeled as of a certain sound type. A different class was attributed to each drum type (i.e. *chico*, *repique*, *piano*) besides an additional one to *madera* strokes (which sound very similar for all drums). Recording each type of drum separately greatly simplified the manual labeling process, since once *madera* sounds had been identified and labeled in a given track, all remaining events could be assigned to its (known) drum type. Finally, a training dataset of 2000 patterns was built through a stratified random sampling (i.e., 500 of each class).

Fig. 1. Testing dataset recording session. Drums on the left are also used for training (`drums-1`), while drums on the right belong to the set used only for testing (`drums-2`).

Another dataset of real performances of drum ensembles was used for testing. This data was collected in other recording session, in which five renowned Candombe drummers took part, playing in groups of three to five. Two of these configurations are depicted in Fig. 1. Audio recordings were done using spot microphones close to each drum.[1] This provides synchronized audio tracks in which a certain drum is predominant, whilst there is interference from the other drums. Complete performances of variable lengths were recorded, approximately from two to four minutes each. The same set of drums, `drums-1`, used for recording the training samples was used in all three-player performances. Another set of drums, called `drums-2`, was involved in the four- and five-player recordings. This setup allows for two different types of experiment regarding the generalization ability of the classification system: one in which training and testing drums

[1] Except for the *chico* drum in ensembles of five players due to equipment constraints.

are the same, but recording conditions (e.g. room acoustics, microphones) and performance configuration (e.g. drum tuning, percussionist) change; and another in which the instruments are also changed.

3 Extraction of Audio Features

In order to find the occurrence of sound events in recorded audio, usually one implements a detection function that emphasizes note onsets by detecting changes in some signal properties, such as the energy content in different frequency bands [2]. This work adopts a typical approach, the Spectral Flux: first, the Short-Time Fourier Transform of the signal is computed for sequential 80 -ms duration windows in hops of 20 ms, and mapped to the MEL scale (approximately linear in log frequency); the resulting sequences are time-differentiated (via first-order difference) and half-wave rectified. To produce the detection function the obtained feature values are summed along all MEL sub-bands. Any peak above 20% of the maximum value in this function is taken as a true onset.

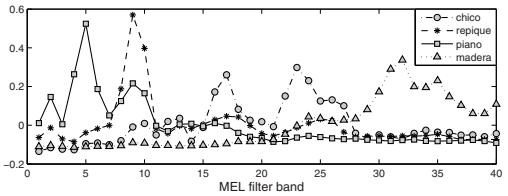
For drum sound classification, this work adopts the same spectral features, specifically the vector containing the first 40 MEL bands (corresponding to frequencies up to 1000 Hz). This value was chosen based on some feature selection experiments.

4 Clustering for Data Exploration

In order to explore the training data, a clustering analysis using the K-means algorithm [8] was carried out. The distance measure for the analysis should reflect the similarity in shape between two spectral feature profiles, and turned out to be a key issue since several measures considered were not appropriate. The Pearson correlation $\mathrm{PearsonCorr}(x, y)$, computed as the inner product of two sequences x and y normalized to zero mean and unit standard deviation, can be seen as a shift-invariant cosine similarity. By treating the data points as the correlated sequences, their distances can be measured as $D(x, y) = 1 - \mathrm{PearsonCorr}(x, y) \in [0, 2]$. The component-wise mean of its points is the centroid of each cluster.

The results of this clustering analysis applied to the training data when setting the number of clusters K=4 is presented in Fig. 2. The confusion matrix of a cluster-to-class evaluation (top left part) shows that *madera* and *chico* classes are correctly grouped, while *piano* and *repique* exhibit a higher rate of misclassification. A three-dimensional representation computed with Multidimensional Scaling (MDS) using the same distance measure is included (right part) for data visualization, and highlights the overlapping of classes. In particular, *repique* is the most troublesome class, which is not surprising since this is the drum of medium size and register, and thus expected to overlap the other drums' spectra. This issue is confirmed by the cluster centroids (bottom left part), whose shape is consistent with the spectral content of each sound class. The centroid of the *piano* drum class has a clear predominance at low frequencies, whereas the centroid of the *madera* class is dominant at high frequencies. At medium

	chico	repique	piano	madera
chico	474	23	0	3
repique	78	403	3	16
piano	1	79	420	0
madera	2	2	0	496
%	94.8	80.6	84.0	99.2



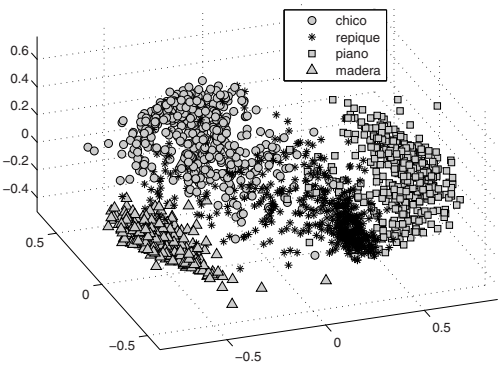


Fig. 2. Clustering analysis of the training data. Confusion matrix of a cluster-to-class evaluation and cluster centroids (left). Three-dimensional MDS representation (right).

frequencies, the centroid of the *repique* class exhibits a maximum towards the lower range, while the centroid of the *chico* class has higher frequency content.

5 Classification Methods

Results of the clustering analysis motivated the idea of testing a very simple classifier based on the obtained centroids: each centroid was considered as a single class prototype in a 1-NN classifier, using the previously introduced Pearson correlation distance. Such a classification scheme can simplify the process of building the training database, since unsupervised clustering can substitute for manual labeling. Furthermore, data coming from different sources, for instance different sets of drums or recording conditions, may be clustered independently so as to better describe classes with more than a single prototype. A k-NN and an RBF-SVM using the same distance measure were also implemented for comparison. SVM parameters were grid-searched in a cross-validation scheme.

6 Experiments and Results

6.1 Predominant Drum Recognition

The predominant drum recognition of a given audio track is tackled in a straightforward manner. First, the Spectral Flux feature is computed, followed by onsets detection, and classification of each detected event into one of the four defined classes. The proportion of onsets in each class gives an indication of the predominant instrument in the audio file. A simple but effective strategy was adopted to improve the detection of the *repique* drum, already identified in the training phase as the most difficult one. Considering that in a real performance, after the rhythm patterns have been initiated (i.e. after the first few seconds), *madera*

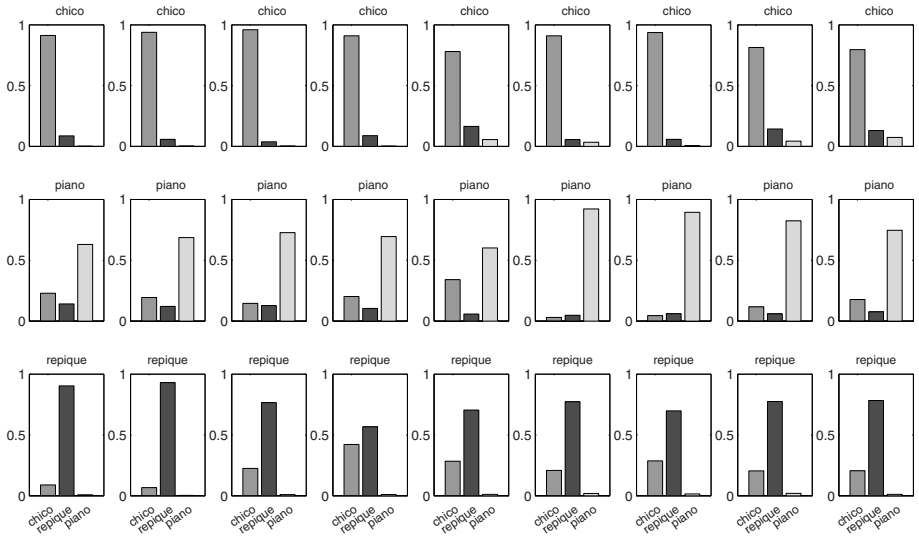
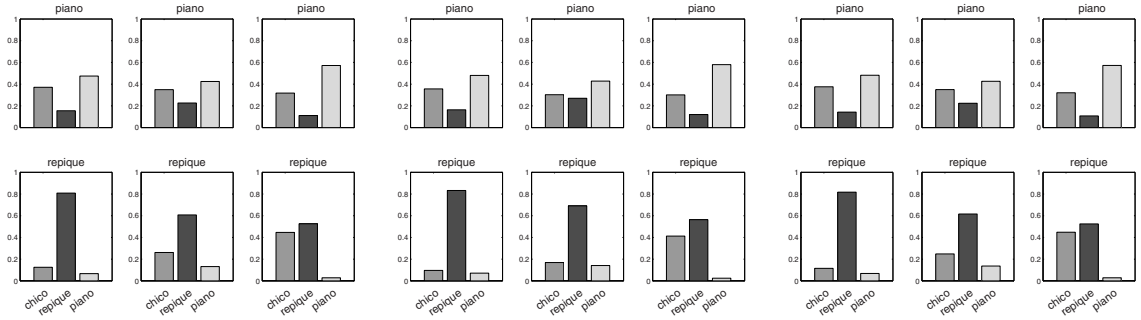


Fig. 3. Results of predominant drum recognition for the three-drum recordings using a 1-NN classifier of training dataset K-means centroids (■ *chico*, ■ *repique*, ■ *piano*).



(a) 1-NN with centroids (b) 5-NN (c) RBF-SVM

Fig. 4. Predominant drum recognition for `drums-2` set (■ *chico*, ■ *repique*, ■ *piano*).

sounds are played only by the *repique* drum, the onsets in the *madera* class were included in the *repique* drum class before computing the proportions.

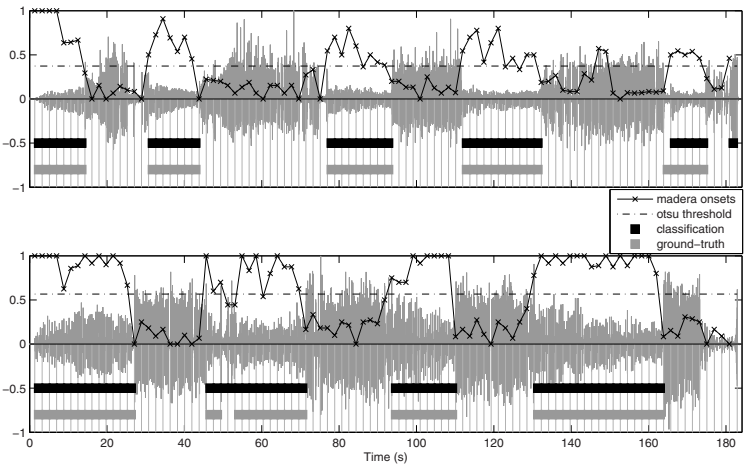
In the first experiment setup all three-drum performances were considered. There are 9 recordings of 3 tracks, totaling 75 minutes and 27 audio files. Note that in this case, the same set of drums of the training samples (`drums-1`) was used. The estimated proportion of onsets for each audio file is shown in Fig. 3, for the 1-NN classifier based on the K-means centroid prototypes. It can be seen that the majority class always indicates the predominant drum. Similar results were obtained with k-NN and RBF-SVM, as shown in the next experiments.

The other set of drums (`drums-2`), not used for training, was employed in another experiment. There are 6 different drums, 3 *piano* and 3 *repique* (no *chico*). A track was processed for each drum, totaling 22 minutes of audio. Classification results are presented in Fig. 4 for a 1-NN of centroid prototypes, a 5-NN, and an RBF-SVM. Although the majority class always reveals the correct

drum type, there is a noticeable difference in the disparity among classes w.r.t. the previous experiment. This seems to disclose some lack of generalization ability to handle different sets of drums. However, it has to be taken into account that these recordings involve more than three drums, which reduces the distance between performers (as seen in Fig. 1) and therefore increases the interference (e.g. *chico* in the *piano* tracks for five-player recordings). Differences among classifiers are marginal, and results are very similar for different choices of k-NN neighbors.

6.2 Detection of *Clave* Pattern Sections

A similar approach was followed for detecting those sections when a *repique* drum plays the *clave* pattern. Five performances in which two *repique* drums take part were chosen for this experiment, totaling 10 tracks and 33 minutes of audio. A *clave* pattern lasts for a whole musical bar; therefore, the recordings were manually labeled indicating all bar locations as well as which of them contained the *clave* pattern. The onsets in each track were detected and classified. Then, the proportion of *madera* onsets to the total detected events within each bar was computed as an indication of the presence of the *clave* pattern. A two-state classification was performed according to a threshold computed using Otsu's method [9]. Finally, to avoid spurious transitions, an hysteresis post-processing was implemented in which a change of state is validated only if it is confirmed by the following two points of the sequence. The segmentation process is illustrated in Fig. 5-left for two of the audio tracks. The performance error attained by the three classifier schemes for each audio track, computed as the percentage of bars in which annotation and classification are different, is presented in Fig. 5-right.



1-NN	5-NN	SVM
8.4	5.9	8.4
1.7	5.4	4.2
2.0	1.0	3.1
2.0	2.0	2.0
9.3	6.5	6.5
1.0	1.0	1.0
2.7	1.4	2.7
2.7	1.4	2.7
0.8	4.7	0.8
11.6	10.9	10.9
4.2	4.0	4.2

Fig. 5. Detection of *clave* pattern for two *repique* tracks of the same performance (left) and classification error for each track of the dataset (right). For each waveform plot: in the upper part, the proportion of *madera* onsets detected within each bar is depicted along with the Otsu threshold; in the lower part, vertical lines indicate the labeled bars, while horizontal thick lines show classification and ground-truth labels.

7 Discussion and Future Research

In this work an approach for predominant drum recognition in audio signals based on modeling onset spectral features was described. It is motivated by practical applications related to ongoing research on Candombe drumming from audio recordings. The reported experiments yielded promising results, even for the 1-NN classifier of centroid prototypes. To this regard, the Pearson correlation measure–which captures the similarity in shape between two spectral profiles–plays an essential role, which will be further assessed in future work.

Automatically detecting *clave* patterns from audio recordings, as proposed in this work, is a valuable tool for studying performance in musicological research. For instance, the interaction of two *repique* drums playing together is clearly visible in Fig. 5. Sections in which a performer plays the *clave* pattern show an almost perfect anti-symmetry between the two tracks. Besides, there exist several variations of the *clave* pattern that deserve a thorough study. To do that, the automatic detection of *clave* sections in a recording could allow dealing with large audio collections. In addition, *clave* pattern serves as a mean of temporal synchronization and can be exploited by automatic rhythm analysis algorithms.

References

1. Andrews, G.: Blackness in the White Nation: A History of Afro-Uruguay. The University of North Carolina Press, Chapel Hill (2010)
2. Dixon, S.: Onset detection revisited. In: Proc. of the 9th International Conference on Digital Audio Effects, Montreal, Canada, pp. 133–137, September 2006
3. Ferreira, L.: An afrocentric approach to musical performance in the black south atlantic: The candombe drumming in Uruguay. TRANS-Transcultural Music Review **11**, 1–15 (2007)
4. Fitzgerald, D., Paulus, J.: Unpitched percussion transcription. In: Klapuri, A., Davy, M. (eds.) Signal Processing Methods for Music Transcription, pp. 131–162. Springer, New York (2006)
5. Fletcher, N.H., Rossing, T.D.: The Physics of Musical Instruments, 2nd edn. Springer, New York (2010)
6. Gillet, O., Richard, G.: Transcription and separation of drum signals from polyphonic music. IEEE Transactions on Audio, Speech, and Language Processing **16**(3), 529–540 (2008)
7. Herrera, P., Dehamel, A., Gouyon, F.: Automatic labeling of unpitched percussion sounds. In: AES 114th Convention, Amsterdam, The Netherlands, pp. 1–14, March 2003. Convention Paper 5806
8. Jain, A.K.: Data Clustering: 50 Years Beyond K-Means. Pattern Recognition Letters **31**(8), 651–666 (2010)
9. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9**(1), 62–66 (1979)
10. Polak, R., London, J.: Timing and Meter in Mande Drumming from Mali. Music Theory Online **20**(1), 1–22 (2014)
11. Sillanpää, J.: Drum stroke recognition. Technical report, Tampere University of Technology, Tampere, Finland (2000)