Towards automatic transcription of expressive oral percussive performances

Amaury Hazan
Music Technology Group
Pompeu Fabra University
Ocata 1, 08003 Barcelona, Spain
ahazan@iua.upf.es

Rafael Ramirez
Music Technology Group
Pompeu Fabra University
Ocata 1, 08003 Barcelona, Spain
rramirez@iua.upf.es

ABSTRACT

We describe a tool for transcribing voice generated percussive rhythms. The system consists of: (a) a segmentation component which separates the monophonic input stream into percussive events (b) a descriptors generation component that computes a set of acoustic features from each of the extracted segments, (c) a machine learning component which assigns to each of the segmented sounds of the input stream a symbolic class. We describe each of these components and compare different machine learning strategies that can be used to obtain a symbolic representation of the oral percussive performance.

Keywords

Knowledge-based approaches, speech processing, performance transcription

INTRODUCTION

The paper presented here focuses on information retrieval in voice generated rhythms. The aim of this work is to develop a system able to reduce the gap between the user and a device (namely keyboard, drum pad or GUI) in order to get a symbolic rhythmic representation. This is relevant as many musicians who just have an intuitive notion of rhythm and groove can not easily transcript a beat they have in mind. Furthermore, in both non western civilizations music and recent western urban genres, the oral tradition of music and especially rhythm is predominant. Few works have focussed in indexing automatically non standard drum based rhythms and an effort in finding a representation that can apply to a whole range of acoustic oral rhythms from beat boxing to Indian tabla oral recitals has to be done. This can be achieved in several ways, from finding an appropriate instrument taxonomy asserting that any performer tries to imitate some drum percussion, to using a completely data-driven model. Between theses extremes, one can consider the taxonomy of human phonemes as a starting point to identify recurrent oral drum clusters. Recent works in non standard drum percussive signals transcription and sax expressive performance analysis lead to believe that we can take advantage of performing some contextual analysis of the oral drum part instead of just considering the description of each isolated percussive event. As starting point we present a system that performs automatic transcription of oral percussive performance into a 4 drum-class score. The system performs percussive events segmentation of the input audio stream, descriptors generation of the percussive segments and classification regarding the generated descriptors. The system components related to user interface and signal processing were designed and implemented within the Object Oriented CLAM framework [1] and are presented below. In continuation we test different Machine Learning techniques on multi-user training and test sets and discuss the results. Finally some related work is presented.

[Leave blank the last 2.5 cm (1 inch) of the left-hand column on the first page for the copyright notice.]

SOUND SEGMENTATION

In order to segmentate the input audio stream, that is, to index each percussive event with an onset time and a duration, we used an energy based algorithm that decomposes the input stream into successive and possibly overlapping frames, compute each frame energy, and indexes successive frames with an overall energy greater than a user-defined threshold. Although being basic, this method fits well to monophonic oral percussive recordings and requires few computational resources, compared to more complex and time consuming method, as band wise processing using psychoacoustic knowledge used in [1].

[1] http://www.iua.upf.es/mtg/clam

ORAL PERCUSSIVE DESCRIPTORS

We defined a compact representation of oral percussive sounds, based on [2] which focussed on classifying acoustic and synthetic drum sounds and which highlighted the importance of some descriptors to classify unpitched sounds. We believe that the noisy nature of oral percussive can lead to good results using similar features. Each percussive event was split into an attack part and a release part with another segmentation algorithm that finds the maximum of the sound envelope to detect the attack/decay boundary. In the case of oral percussive sounds the attack part is considerably shorter than the release. Therefore, we only extracted temporal descriptors from the attack and both temporal and spectral descriptors from the decay part.

Temporal Features: We included in our system the following temporal features. **Duration** refers to the duration of the segmented attack or decay part and **Log-Duration** to its natural logarithm, **Energy** refers to the mean energy of the segmented part. **Zero-crossing rate** is calculated as the rate the signal changes sign in a given period of time, that is, the duration of the segmented part. **Temporal centroid** refers to the centroid of the waveform.

Spectral Features: To compute the spectral features we did the following. We performed a Fast Fourier Transform of the successive frames of the decay part of the oral percussive event, and for each of the frames, we computed the descriptors presented below. **Energy** refers the calculated energy of the spectrum at a given frame, as **Centroid** refers the the center of gravity of the spectrum. **Flatness** is the ratio between the geometrical mean and the arithmetical mean of the spectrum. A flat spectrum sound tends to be perceived as noisy as low flatness refers to more harmonic sounds. **Kurtosis** is the forth-order central moment and gives the information if the spectrum is peaky or not. Finally, **Mel Frequency Cepstrum Coefficients** (MFCC) have been used in speech recognition and musical application due to their capacity to give a compact representation of the spectrum. We retained the first five MFCC. These descriptors were computed for each frame of the decay part and and we retained as a compact representation of the decay spectrum the respective mean and variance of each of them among these frames. Consequently we generated a training set which was composed by instances of 28 features, namely 10 temporal features and 18 spectral features.

MACHINE LEARNING TECHNIQUES

A taxonomy of 4 standard drum sounds was defined i.e the classes were *Bass Drum*, *Snare Drum*, *Closed Hat* and *Open Hat*. A training set formed with the recordings of 4 performers, 2 men and 2 women, was used (they were asked to imitate the sound of each of the drum classes), totalizing 242 training instances. Different machine learning techniques that were tested on this dataset are presented below.

Tree induction algorithms build a tree model by selecting at each node the most relevant attribute. We compare the results of C4.5 [6], C4.5 with boosting, and C4.5 with bagging. Boosting refers to a meta algorithm that can improve the results of any classification algorithm by giving to each instance of the training set a particular weight proportional with the difficulty to classify such instance. That is, a first classification model is proposed giving the same weight to all the training instances. Misclassified instances with this model are then given a greater weight, and so on. After a user defined number of iterations (in our case 10 iterations) the resulting model is able to deal with "difficult" training instances. This boosting method can drastically improve the results of an inaccurate model, thought overfitting can occur. Bagging helps improving classification by sampling the training set in smaller training sets, and calling the induction tree to build a model for each of these subsets. The classification decision is then taken combining the votes of each of the sub-models.

Lazy Methods: The notion of lazy learning subsumes a family of algorithms that store the complete set of given (classified) examples of an underlying example language and delay all further calculations until requests for classifying yet unseen instances are received. We included in our experiment K-Nearest Neighbor algorithm, is one of the most popular instanced-based algorithm, which handles well noisy data if the training set has an acceptable size. The main idea of this algorithm is to compare a test set with its nearest neighbors (which number is determined by the user, we empirically found the best accuracy for K=1).

CLASSIFICATION RESULTS

The presented classification results of our system come from 2 experiments. In the first one, we performed a 10 fold cross validation of the training set, that is, the model was built 10 times, putting apart one training instance in order to test it against this model. This methods gives a preliminary idea of the the classification accuracy of the system, and the results are presented in Table 1. In the second experiment we simulated a real world situation of the system, as we tested the classification model with 2 new performers recordings that were not used in the training set. The results of the experiments are presented in Table 2. The classification accuracy, denoted C.A, has to be compared with the accuracy of a random classification i.e. 25%.

Algorithm	C.A(%)
kNN	75.7
C4.5	81.6
C4.5 w/Boosting	89.3
C4.5 w/Bagging	**86.9**

Table 1.Cross validation results of the training set

Algorithm	C.A(%)
kNN	79.0
C4.5	79.0
C4.5 w/Boosting	87.0
C4.5 w/Bagging	**90.0**

Table 2. Classification accuracy for a test set with recording from unseen performers

DISCUSSION AND FUTURE WORK

A classification accuracy of 90% has been obtained in a test involving 2 performers whose recordings were not used in the training set, the test set being composed of 62 instances. This last experiment gives an idea of the accuracy that can be reached in a real world application, and the results are encouraging. The percussive event descriptors presented before have shown to be useful, especially attack and decay durations and zero crossing rate among the temporal descriptors, kurtosis, second and forth MFCC means among the spectral descriptors. Nevertheless the system performs classification into a fixed 4 class taxonomy, and the results fall quickly if we use a test set in which the user produces the rhythm using quite different phonemes compared to these that were used during the training. To improve the robustness of the system, we plan to consider a rhythmic role based analysis which would use the human phoneme taxonomy, considering that 2 different successions of phonemes can describe the same rhythmic pattern. To achieve this aim, we plan to increase the amount of training data and the diversity of the sources. We also plan to study the results of a fast recognition system, although being more limited, in order to design a real time application.

RELATED WORK

Automatic audio analysis has been widely studied during the past years, and a growing interest has been shown in developing high level content extraction systems able to process automatically large amounts of data. Some works have largely helped the developments of the automatic drum classification branch. In [1] a system that performs sound segmentation using psychoacoustic knowledge (i.e using the characteristics of human auditive perception) has been designed. Automatic drum transcription has been studied in [2] using the standard drum taxonomy at 3 levels of abstraction. [3] deals with the problem of beat induction and meter detection of segmented signals. The problem of analyzing audio percussive excerpts which are not generated using standard drum sounds is studied in [4, 5] when some context analysis of the percussive events is considered. In [4], which focuses on the automatic labelling of tabla signals into tabla recital vocabulary, a Hidden Markov Model is considered to represent the contextual dependencies between percussive strokes and because the representation symbols may be context dependent. In [5], arbitrary sounds such as voice, or hand claps can be segregated into 3 labels that represent a rhythmic role rather than an instrument, and thus concentrate on the pattern context of each rhythmical event. Melody context analysis has also been used in [7] to study expressive transformations in monophonic sax performances of jazz standards. In this work, an Inductive Logic Programming approach is being considered i.e. the system tries to induce first order logic rules that take into account the melodic context of the notes to induce transformations rules, in which the tradeoff between precision and generality can be easily controlled.

ACKNOWLEDGMENTS

This work is being supported by the Spanish TIC project ProMusic (TIC 2003-07776-C02-01).

REFERENCES

1. Klapuri, A. (1999). Sound Onset Detection by Applying Psychoacoustic Knowledge, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP.
2. Herrera, P. and Dehamel, A. and Gouyon, F. (2003). Automatic labeling of unpitched percussion Sounds, Proceedings of Audio Engineering Society, 114th Convention, Amsterdam, The Netherlands.
3. Gouyon, F. Herrera, P. (2003). Determination of the Meter of musical audio signals: Seeking recurrences in beat segment descriptors. Proceedings of Audio Engineering Society, 114th Convention Amsterdam, The Netherlands
4. Gillet, O. Richard,G. (2003). Automatic Labelling of Tabla Signals, Proc of ISMIR 2003, Baltimore, USA Oct. 2003.
5. Jouni, P., Klapuri, A. (2003) Model-Based Event Labeling in the Transcription Of Percussive Audio Signals Proc. of the 6th Int. Conference on Digital Audio Effects (DAFX-03), London, UK, September 8-11.
6. Quinlan, J.R. (1993). C4.5: Programs for Machine Learning, San Francisco, Morgan Kaufmann.
7. Ramirez, R. Hazan, A. Gómez, E. Maestre, E. (2004). Understanding Expressive Transformations in Monophonic Musical Phrases. Proc. of SMC'04, Paris, October 20-22.