AUTOMATIC TRANSCRIPTION OF DRUM LOOPS

Olivier Gillet and Gaël Richard

GET-ENST (TELECOM Paris)
Signal and Image Processing department
46, rue Barrault, 75013 Paris, France
olivier.gillet,gael.richard@enst.fr

ABSTRACT

Recent efforts in audio indexing and retrieval in music databases mostly focus on melody. If this is appropriate for polyphonic music signals, specific approaches are needed for systems dealing with percussive audio signals such as those produced by drums, tabla or djembé. Most studies of drum signals transcription focus on sounds taken in isolation. In this paper, we propose several methods for drum loops transcription where the drums signals dataset reflects the variability encountered in modern audio recordings (real and natural drum kits, audio effects, simultaneous instruments, ...). The approaches described are based on Hidden Markov Models (HMM) and Support Vector Machines (SVM). Promising results are obtained with a 83.9% correct recognition rate for a simplified taxonomy.

1. INTRODUCTION

Pre-recorded audio databases of drum loops are becoming very popular and are now widely used in modern music compositions. Such databases typically gather a large number of short drum signals (called loops) referenced (in the best case) by their tempo and general style. Due to the continuously growing size of such databases, searching an appropriate drum loop only based on tempo and style becomes rather tedious. There is, therefore, a need for content-based methods that would allow to search in these databases more efficiently, that is with more natural or specific queries. An essential aspect of such a searching tool is the necessity to obtain beforehand an automatic transcription of drum loop signals.

The transcription of drum signals has gained much interest in the past few years. For example, McDonald & al. [1] identified isolated percussive sounds based on spectral centroid trajectories and Sillanpää & al., [2], presented a classification system in five broad categories (Bass drum, snare drum, hi-hat, cymbal, and toms). More recently, Gouyon & al. [3] evaluated several methods for natural and synthetic drum signals recognition. These technics proved to be successful but were limited to isolated sounds.

Other works deal with more complex signals and aim at extracting the drum tracks from polyphonic music signals [4], or use source separation approaches to pre-process drum loops signals [5]. A particularity of drum loops signals is that each event can be produced by simultaneous strokes on different instruments (for example bass drum and hi-hat). Lawlor & al. [5] showed very promising results but this work was limited to three instruments (snare drum, kick drum and hi-hats) and tested on only fifteen manually selected loops.

Another particularity of drum loops is that they contain a succession of events (or strokes). As a consequence, drum loop signals or drum tracks often exhibit a temporal structure. Two concurrent studies have exploited such a structure by means of a sequence model, or ”language model” by analogy with large vocabulary speech recognition systems ([6] for drum sequences transcription, or [7] for the transcription of tabla signals).

The objective of this paper is to propose and to evaluate two novel approaches for the transcription of drum loops signals. The work described in this paper is a following, to some extent, of a previous study conducted on tabla signals ([7] where specific sequence models were successfully used). It is important to emphasize that this work is conducted on a rather large database of drum loops (315 drum loops containing 5327 strokes). Moreover, this database reflects various aspects of variability encountered in modern audio recordings (natural and synthetic drum kit, audio effects such as flanger or reverberation, ...) and includes complex signals resulting from simultaneous strokes on several instruments.

The paper is organized as follows. Next section describes the overall system architecture. Then, section 3 is dedicated to the description of the database used and of the statistical approaches followed for the automatic transcription of drum loops. Section 4 discusses the results obtained and, finally, section 5 suggests some conclusions.

2. SYSTEM ARCHITECTURE

The aim of this system is to transcribe drum loops signal into a higher level of representation for indexing and retrieval applications. The information automatically extracted from the signal includes the instrument (or the combination of instruments) played on each stroke, the onset time of each event and the overall tempo of the drum loop. The system architecture is then based on three major parts:

1. a segmentation and tempo extraction module (described in section 3.2)
2. a features extraction module (see section 3.3)
3. and a classification module for which three different approaches were tested (see section 3.4)

The overall architecture of the system is depicted in figure 1.

0-7803-8484-9/04/$20.00 ©2004 IEEE

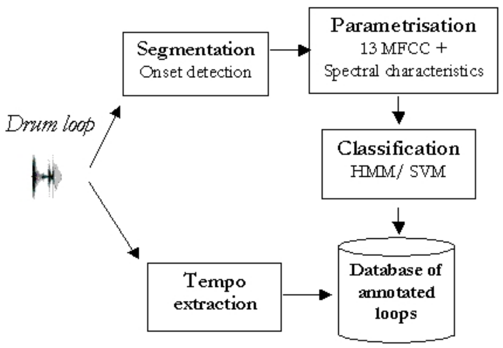

Fig. 1. Architecture of the drum loop transcription system.

3. TRANSCRIPTION OF DRUM LOOPS

3.1. Drum loops database

The database used for this study consists of 315 drum loops containing 5327 strokes. This database was manually annotated using eight basic categories: *bd* for bass drum, *sd* for snare drum , *hh* for hi-hat, *clap* for hands clap, *cym* for cymbal, *rs* for rim shot, *tom* for any other tom of a drum and *perc* for all other percussive instruments with more definite pitch such as congas, djembé or tabla. When two or more instruments are played at the same time, the event is labelled by all corresponding categories (for example if bass drum and cymbal are hit simultaneously, both labels are attached to the corresponding stroke). Combinations of up to four simultaneous instruments exist in the database (although they are not frequent).

All drum loops were extracted from commercial samples CDs. The loops are representatives of different styles including rock, funk, jazz, hip-hop, drum'n'bass and techno and are played on different drum kits including electronic kits. Some loops also have special effects such as flanger, reverberation, distortion or compression. The loop duration is between two and fifty seconds. If our database is comparable (or larger) in size to the dataset used in most other studies, it is important to emphasize that it contains more important situations commonly encountered in modern audio recordings including simultaneous percussive instruments and audio effects. A compressed version of a few drum loops along with their annotation is given on our web site ([8].

3.2. Segmentation and tempo extraction

Due to the impulsiveness of the drum loops signals, it seems appropriate to segment the signal into individual events. Each segment then corresponds to a stroke on a given instrument or to simultaneous strokes on several instruments and can be labelled accordingly. To segment the drum loops signals, an onset detection algorithm based on sub-band decomposition was used [9]. Since the drum loops signals consist in localized events with abrupt onsets, this algorithm obtains very satisfying results. Concurrently, the overall tempo is estimated using a slightly modified version of Scheirer's algorithm [10]. It consists in associating a filter bank with an onset detector in each band and with a robust pitch detection algorithm such as the spectral sum or spectral product [11].

3.3. Features extraction

To select an appropriate features set, a simple classifier (k-Nearest Neighbors) was used. The recognition rates on the different feature sets envisaged were compared and the results obtained have, for a large part, confirmed those obtained by [3]. Finally, our features set includes:

- **Mean of 13 MFCC** The Mel Frequency Cepstral Coefficients (MFCC) including c_0 are calculated on 20 ms frames with an overlap of 50 %. The mean is then obtained by averaging the coefficients over the stroke duration. In our work, c_0 is not excluded since it led to better classification performance.
- **4 Spectral shape parameters** defined from the first four order moments:
  - the spectral centroïd given by $S_c = \mu_1$,
  - the spectral width given by $S_w = \sqrt{\mu_2 - \mu_1^2}$,
  - the spectral asymmetry S_a defined by the spectral skewness :$S_a = \frac{2(\mu_1)^3 - 3\mu_1\mu 2 + \mu_3}{(S_w)^3}$
  - and the spectral flatness S_f defined from the spectral kurtosis $S_f = \frac{-3\mu_1^4 + 6\mu_1\mu_2 - 4\mu_1\mu_3 + \mu_4}{(S_w)^4} - 3$

  where $\mu_i = \frac{\sum_{k=0}^{N-1} k^i.A(k)}{\sum_{k=0}^{N-1} A(k)}$ and where $A(k)$ is the amplitude of the k^{th} component of the Fourier transform of the input signal.
- **6 Band-wise Frequency content parameters** These parameters correspond to the log-energy in six pre-defined bands (in Hertz: [10-70] Hz, [70-130] Hz, [130-300] Hz, [300-800] Hz, [800-1500] Hz, [1500-5000] Hz). These bands were chosen according to a meticulous observation of the frequency content of each drum instrument. Such a choice led to better performance compared to a more classical Bark scale filterbank (as used in [3]).

3.4. Classification approaches

3.4.1. Hidden Markov Models

Drum signals exhibit some kind of context dependencies. In fact, the sound produced by a given stroke (and especially if it is resonant) may continue while the following stroke happens and thus may have an impact on the spectral characteristics of the following event. Also, some typical sequences of instruments are often played (i.e succession of bass drum and cymbal,...).

An efficient approach that integrates context (or time) dependencies is given by the Hidden Markov Model (HMM). This class of models is particularly suitable for modelling short term time-dependencies and it has been successfully used for a wide variety of problems ranging from speech recognition to tabla signals transcription [7]. In such a framework, the sequence of feature vectors $\overline{o_t}$ is represented as the output of a Hidden Markov Model. The recognition is performed by searching the most likely states sequence, given the output sequence of feature vectors. In this model, a succession of strokes $S_{k-m}, ..S_k$ is associated to each state q_t. Intuitively, the state q_t represents the stroke S_k in the context of $S_{k-m}..S_{k-1}$ at time t. The model is thus clearly context dependent. The transition probabilities from state i to state j

is given by (in the case of 3-grams):

$$\begin{aligned} a_{ij} &= p(q_t = j | q_{t-1} = i) \\ &= p(s_t = S_3 | s_{t-1} = S_2, s_{t-2} = S_1) \end{aligned}$$

where $p(s_t = S_3)$ is the probability density of observing the instrument S_3 at time t. The observation probability distribution associated to each state is given by:

$$\begin{aligned} b_i(x) &= p(\overline{o_t} = x | q_t = i) \\ &= p(\overline{o_t} = x | s_t = S_2, s_{t-1} = S_1) \end{aligned}$$

In this work $b_i(x)$ is either modelled by a single mixture (a Gaussian vector distribution with diagonal covariance matrix) or a mixture of two Gaussian distributions. For example, in the single mixture case, the feature vectors are modelled with a single vector distribution of 23 Gaussian distributions (where each Gaussian characterizes the mean, variance of each parameter of the features set). In the case of several Gaussian mixtures, the EM algorithm is used. The decoding is carried out using the traditional Viterbi algorithm.

3.4.2. Support Vector Machines

The other classification approach used in this study is known as Support Vectors Machines (SVM) which are well designed for binary problems classification. Support Vector Machines non-linearly map (using a Kernel function) their n-dimensional input space into a higher dimensional feature space where the two classes are linearly separable with an optimal margin. Such classifiers can perform binary classification and regression estimation tasks but can also be adapted to perform n-class classification [12],[13]. SVM have very interesting generalization properties since the decision surface in the data space can be well defined even in the case where a complex surface would be necessary to separate the data. Several kernels can be used. For this study, the library LibSVM [14] was used and a radial basis kernel was chosen ($K(x, y) = \exp^{-\frac{(x-y)^2}{\lambda}}$ with $\lambda = \frac{1}{N}$ and where N is the number of features).

Note that with SVM the data are directly the features vectors obtained for each strokes regardless of their left context (i.e. there is no sequence model in this case).

Since we are interested in labelling each segment by one or many labels among the n instruments in the kit, two different approaches are possible :

One 2^n -ary classifier. In a first approach, only one classifier is used, in which each possible combination of strokes is represented by a distinct class. Our study uses 8 instruments, implying thus the use of a 255 classes classifier. It is important to notice that among the 255 possible combinations of strokes only 45 of them were present in the database.

n binary classifiers In a second approach, one binary classifier per instrument is trained. This binary classifier is used to decide whether the instrument is played or not in each segment.

3.5. Drum kit dependent approach

Due to the high variability of the data, a drum kit dependent approach was also tested. Instead of using one generic classifier, four classifiers "specialized" in four different kinds of drumkits were trained by splitting the training database according to style / drumkit criteria. The four categories roughly correspond to four types of drum kits:

Instrument alone or prominent	
Snare Drum, Rim-Shot or Clap	1440
Bass drum	1652
Hit-hat or Cymbal (alone)	1558
Conga, tom, djembé, Tabla	462
Combinations	
Bass Drum + (snare drum, Rim-Shot or Clap)	53
Snare Drum + (Tom or Congas)	44
Bass drum + snare drum + (tom or Congas)	12
Bass drum + (Tom or Congas)	106

Table 1. Number of occurrences in the database of each label for the simplified taxonomy

Electro style which mostly includes sounds generated by electronic drums such as Roland TR-808 or TR-909 (41 loops - techno, hip hop) ,

Light style which is representative of traditional acoustic drums eventually with light effects (125 loops - jazz, funk),

Heavy style which includes sounds with heavy and long reverberation (67 loops - rock, industrial),

Hip-hop style which includes sounds often compressed with various audio effects such as flanger (82 loops - drum'n'bass, hip hop).

We use as a transcription the output of the classifier which gives the best likelihood score. Note that this approach can only be used with HMM-based classifiers, since the SVM classifiers perform a "hard" decision.

4. TRANSCRIPTION RESULTS

4.1. Taxonomy

In theory, all instruments from the eight basic categories can be played simultaneously leading to 2^n possible combinations. In practice (i.e. in our database) only 45 out of 255 combinations are observed. As a consequence, the first taxonomy (*detailed taxonomy*) is defined where each combination is characterized by a label.

To better analyse the results, another taxonomy is also used. The so-called *simplified taxonomy* gathers some instruments in a reduced number of categories and only keeps the label of the prominent instrument for each stroke with a few exceptions for frequent combinations or for combination where there is no salient instruments (see table 1).

Note that the simplified taxonomy is only used to provide an additional interpretation of the results but that the same models have been used for both (i.e. same training and decoding).

4.2. Evaluation protocol

For evaluation, the usual cross-validation approach was followed (often called ten-fold procedure in the literature [15]). It consists in splitting the whole database in 10 subsets randomly selected and in using nine of them for training and the last subset (i.e. 10 % of the data) for testing. The procedure is then iterated by rotating the

Taxonomy	Detailed	simplified
one 2^n-ary classifier		
HMM, 3-grams, 1 mixture	59.1%	78.7%
HMM, 3-grams, 2 mixtures	58.7%	78.3%
HMM, 4-grams, 1 mixture	59.3%	77.3%
SVM	65.1%	83.1%
n binary classifiers		
HMM, 3-grams, 1 mixture	45.6%	68.6%
HMM, 3-grams, 2 mixtures	41.5%	65.2%
HMM, 4-grams, 1 mixture	34.0%	53.1%
SVM	64.8%	83.9%
Drum kit dependent approach		
HMM, 3-grams, 1 mixture	62.5%	82.2%
HMM, 3-grams, 2 mixtures	58.4%	83.4%
HMM, 4-grams, 1 mixture	60.8%	77.3%

Table 2. Drum instruments recognition results

10 subsets used for training and testing. The results are computed as the average values for the ten runs.

4.3. Results and discussion

The results obtained on our dataset are summarised in table 2. It can be observed that SVM clearly outperforms the HMM approach for both taxonomies when the models are trained on all data. This may be explained by the fact that the rather simple acoustic model used with HMM cannot cope with the high variability of the dataset.

This is confirmed by the experiment implementing a drum dependent approach. When a drum kit dependent model is used for HMM, performances of both approaches (SVM and HMM) are comparable. In fact, this approach permits to split the data according to the drum kit used and thus to decrease the variability of data within a given class which is appropriate for HMM.

Still, it is surprising that the SVM classification that does not include any sequence modelling outperforms the HMM approach. In fact, the sequence modelling was very efficient with tabla signals where time dependencies can be observed at the label-level (one same stroke can have different labels depending on the context in which it is played) while with drum signals time dependencies can be observed only at the signal-level (the same instrument can sound differently depending on the context in which it is played). Also, one of the main differences of the two studies is that for tabla all performances were representative of a unique style (which is again not the case for the drum loops dataset). This suggests that sequence modelling may become much more efficient if the drum signals are gathered according to a given style.

Another reason for the better performances of the SVM could be that much more training data are used for each class (instrument combination) with SVM since the events are here considered regardless of their left context. Clearly, more variability is attached to the data of a given class, but this is well supported by SVM.

5. CONCLUSION AND FUTURE WORK

This paper proposed novel approaches for drum transcription and evaluated these methods on complex drum loops signals. If promising results were obtained (83.9% using a simplified taxonomy), they suggest that the acoustic model part could be improved and several directions can be envisaged. For exemple, data transformation such as Principal Component Analysis (PCA) which leads to feature vectors with independent components and acoustic models with higher number of gaussian mixtures will be tested. Also, despite the rather large size of our corpus, it clearly appears that better modelling could be achieved with a larger dataset and this especially for HMM approaches. Finally, it is planned to build a combined system that would take into account the respective advantages of SVM and HMM sequence modelling.

6. REFERENCES

[1] S. McDonald and C.P. Tsang, "Percussive sound identification using spectral centre trajectories," in *Proc. of 1997 Postgraduate Research Conference*, 1997.

[2] J. Sillanpää, A. Klapuri, J. Seppänen, and T. Virtanen, "Recognition of acoustic noise mixtures by combined bottom-up and top-down approach," in *Proc. of EUSIPCO-2000*, sept.

[3] F. Gouyon, P. Herrera, and A. Dehamel., "Automatic labelling of unpitched percussion sounds.," *In Proc. of the 114th AES convention*, March 2003.

[4] O. Delerue, F. Gouyon, A. Zils, and F. Pachet, "Automatic extraction of drum tracks from polyphonic music signals.," *In Proc. of WEDELMUSIC2002*, December 2002.

[5] D. FitzGerald, E. Coyle, and B. Lawlor, "Sub-band independent subspace analysis for drum transcription.," *In Proc. of 5th Int. Conf. on Digital Audio Effects (DAFX'02),*, 2002.

[6] J.K. Paulus and A. Klapuri, "Conventional and periodic n-grams in the transcription of drum sequences.," *In Proc. of 5th Int. Conf. on Digital Audio Effects (DAFX'02),*, 2002.

[7] O. K. Gillet and G. Richard., "Automatic labelling of tabla signals.," *In Proc. of the 4th ISMIR Conf.*, 2003.

[8] http://www.enst.fr/~grichard/Publications/Icassp04_1.htm

[9] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. of IEEE-ICASSP*, Phoenix, 1999.

[10] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *JASA*, vol. 103, no. 1, pp. 588–601, 1998.

[11] M. Alonso, B. David, and G. Richard, "A study of tempo tracking algorithms from polyphonic music signals," in *Proc. of 4th COST276 Workshop*, Bordeaux, France, March 2003.

[12] J. Weston and C. Watkins, "Multiclass support vector machines.," In tech. rep. csd- tr-98-04,, Royal Holloway Univ. of London,, 1998.

[13] U. Kressel, *Pairwise classification and support vector machines.*, In Advances in Kernel Methods : Support Vector Learning,, 1999.

[14] Chih-Chung Chang and Chih-Jen Lin., "Libsvm : a library for support vector machines,," Software available at www.csie.ntu.edu.tw/~cjlin/libsvm., 2001.

[15] P. Herrera, A. Dehamel, and F. Gouyon, "Automatic labeling of unpitched percussion sounds," in *114th AES Convention*, Amsterdam, The Netherlands, March 2003.