Drum Event Detection by Noise Subspace Projection and Classification

Olivier Gillet and Gaël Richard
GET-TELECOM Paris
37, rue Darreau
75014 Paris, France
[olivier.gillet, gael.richard]@enst.fr

The first section describes our algorithm in general terms, independently of the way it was adapted / implemented for the MIREX evaluation. The second section details how our algorithm was modified for the MIREX evaluation. Finally, the third section discusses the results of the evaluation.

1 Summary of the algorithm

Pre-processing of stereo signals A pair of gains is selected for each channel, in order to maximize an impulsiveness criterion on the envelope of the remixed monophonic signal. The impulsiveness measure I is computed as follows: Firstly, an envelope signal $s'(t)$ is obtained by half-wave rectifying, decimating, low-pass filtering, and differentiating $s(t)$. Then a contrast factor on this envelope is computed. This contrast factor is defined as the ratio between the arithmetic and geometric mean of the envelope signal:

$$I(s) = \frac{\sum_{t=1}^{T} s'(t)}{T \sqrt[T]{\prod_{t=1}^{T} s'(t)}}$$

Filter Bank The filter bank used is an octave-band (dyadic) filter bank, with $M = 8$ voices - each frequency band being one octave large. The sampling rate of the input being equal to 44100 Hz, it results in the following eight frequency bands (in Hz): [0,172], [172,345], [345,689], [689,1378], [1378,2756], [2756,5512], [5512,11025] and [11025, 22050]. The filter was implemented using a 100th order FIR filter as a prototype.

Band-wise noise subspace projection The noise subspace projection stage is based on the *Exponentionally Damped Sinusoidal* (EDS) model. The tracking of the signal subspace itself is achieved using the classical EVD iterative algorithm, with 46ms long windows, using a 3/4 overlap. Two exponentially damped sinusoids are used for $x_1(t)$ (lowest frequency band), five for $x_2(t)$, ten for $x_3(t)$ and $x_4(t)$; and eight for the other bands.

The output of the noise subspace projection is thus 8 sub-band noise signals $e_k(t)$. Because of the multirate implementation of the filter bank, these signals need to be resynchronized in time, by upsamling them and by applying a synthesis filter.

Onset detection and classification The sub-band noise signals are directly used to detect onsets. Each of these sub-band noise signals is half-wave rectified and low-pass filtered with a half-Hanning window, the resulting signal being noted $b_k(t)$. The onsets are found by peak-picking $\frac{d}{dt} b_k(t)$. Onset are subsequently grouped.

Features extraction For each onset located at time t, the following features are computed over a 100ms long window starting at t:

- The energy in the first 6 sub-bands. These features can directly be computed from the decomposition.
- The average of the 12 first MFCC (without c_0) across successive frames. The MFCC are computed on the full-band noise signal $\sum_k \hat{e}_k(t)$

SVM classification Three classifiers are trained, each of them detecting the presence of a different class of drum instrument (one classifier for bass drums, one for sanre drum, and the other for hi-hats and cymbals). The classifiers used are Support Vector Machines, with a general-purpose kernel (radial basis function). The training set consists of residual drum signals extracted from polyphonic music using the noise subspace projection.

The output of each SVM $f(x)$ is mapped to the interval $]0, 1[$ with a sigmoid function: $p(x) = \frac{1}{1+e^{Af(x)+B}}$. The parameters A, B are fitted using maximum likelihood estimation on a subset of the training data.

An event is detected whenever the output probability is higher than a given threshold. The thresholds are manually adjusted.

Computation time (2.4 Ghz Athlon) The following times are estimated for 30s excerpts.

Task	time (seconds)
Stereo pre-processing	56
Decomposition	250
Onset detection	15
Features extraction	6
SVM classification	2

Further details Each of these components is presented in details in (Gillet and Richard, 2005).

2 MIREX 2005 submission - Implementation details

In order to ease the deployment of our original algorithm, which was developed with a combination of Matlab and C tools called from a Python script, we decided to rewrite the entire algorithm in Matlab. *The Spider*, a machine learning toolbox, was used to perform the Support Vector Machine (SVM) learning and classification. Unfortunately, we experimented a few problems with the *platt* function of this toolbox, which is implementing the computation of probabilistic outputs for SVM. As a result, we simply used as a decision rule the sign of the SVM output, rather than a calibrated probability.

The variants 1 and 2 of our algorithm correspond to two settings of the C parameter for SVM ($C = 10$ for algorithm 1 ; $C = 1$ for algorithm 2). The choice of these parameters is discussed in the following section.

In the onset detection module, we used a FIR differentiator (derivative of a local polynomial interpolator) rather than a simple differentiation. This resulted in smoother detection functions.

Finally, the first step of the algorithm (stereo preprocessing) was not implemented, as monophonic files were used for the contest.

3 Discussion of MIREX 2005 results

3.1 Hi-hat detection.

For each music collection, our algorithm obtained very poor scores for the hi-hat detection task (0.343 for the average F-measure).

Actually, our algorithm, as it was presented in (Gillet and Richard, 2005), does not transcribe the hi-hat tracks. We adapted our submission to this task by simply:

- Computing the noise subspace projection in the 8 bands of the decomposition, rather than in the first 6 bands of the decomposition as we did in our original work. This also results in an increased complexity.
- Including in the features vector the energy in these 2 extra bands.
- Adding a third SVM classifier, performing the hi-hat/non hi-hat discrimination.

However, this approach, adapted from the bass drum/snare drum detection modules, is not optimal for the hi-hat detection task. Firstly, our motivation to use the noise subspace projection is that in the residual noise signal, the bass drum and snare drum are louder than the note attacks from pitched instruments. This observation is not true for hi-hats, and thus, a better discrimination model, or better features for classification are needed.

Secondly, in the case where only two categories are used (bass drum and snare drum), there is a rather small number of occurences where both events occur simultaneously. However, when the hi-hat category is added, the number of simultaneous events such as bass drum + hi-hat or snare drum + hi-hat increases. The transcription of such simultaneous events is easier with a complete source separation approach, that considers each instrument as a distinct source, rather than with our approach which only aims at extracting a single source containing all the percussive instruments.

3.2 Fine-tuning the precision / recall trade-off.

Our algorithm achieved, for all music collections, abnormally high onset precision, and abnormally low onset recall scores. The average onset detection precision is 77.09% for our algorithm, 65.68% for the best of the others. The average recall score is 40.63% for our algorithm, and 63.38% for the worst of the others. Three hypothesis can account for these results:

Incorrect onset detection threshold. A first possible explanation is that the threshold used for onset detection was set too high. Consequently, only the most salient onsets were detected.

Temporal imprecision of onset detection. During the evaluation of our algorithm in (Gillet and Richard, 2005), we allowed an error of up to 100ms between the original (ground-truth) event and the detected event. In the scope of this evaluation, the highest temporal difference tolerated was 40ms. It is likely that the less accentuated drum events, or the events occuring slightly before or after more accentued notes have been detected with a temporal error greater than 40ms.

Rejection model Even after the noise subspace projection, the residual noise signals still contain attacks or transients from pitched instruments. Thus, the onset detection stage will not only detect onsets associated to drum events, but also to the loudest of the attacks from other instruments. The discrimination between these two cases is performed by the SVM classifiers: in case each classifier returns a negative result (non-bass drum, non-snare drum, non-hi hat), the onset is rejected. The low recall rate of our algorithm can be explained by a too selective rejection model. In our original work, we achieved a good precision/recall trade-off by detecting a bass drum or a snare drum when the output of the corresponding SVM classifier (expressed as a probability) was greater than 0.4. The algorithm we submitted to the evaluation does not use probabilistic outputs, and thus corresponds to the case where this threshold is set to a higher value (> 0.5).

It is not clear which of these hypothesis account for the high precision rate. The availability of the evaluation data, or at least the ground truth transcriptions and the output of our algorithms, would help us to understand our mistakes.

3.3 Importance of training data.

Contrary to the other submissions, our approach is based on statistical machine learning. The performance of such approaches are mostly determined by two factors : the

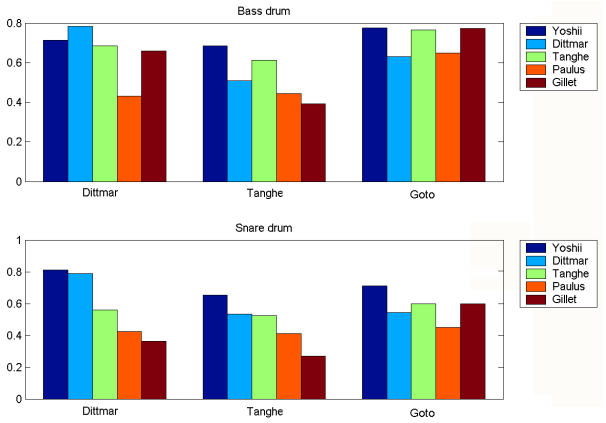

Figure 1: F-measure score for the bass drum and snare drum detection tasks, on the three music collections. Only the best algorithm of each contestant is shown.

learning (generalization) capacity of the machine learning algorithm used ; and the size of the training database.

The two versions of the algorithm we submitted correpond to two settings of the parameter C of the Support Vector Machine classifiers. This parameter expresses a trade-off between the complexity of the model (the number of support vectors), and its training error. Higher values of C will result in a perfect fitting of the training dataset, while smaller values will yield better generalization properties. The algorithm 1 used $C = 10$, while the algorithm 2 used $C = 1$. It can be seen here that improving the generalization capacity of the classification algorithm (while increasing its error on the training set) leads to better results. Smaller values of the parameter C could have possibly given better results.

Since the database we used in (Gillet and Richard, 2005) did not contain annotations for the hi-hat tracks, we decided to use only the sample data provided by the contest organisators to train our algorithm. This dataset consisted of 4 30s excerpts from the Christian Dittmar collection ; 9 30s excerpts from the Koen Tanghe ; and 10 full-length excerpts (ranging from 3m11s to 6m07s) from the Masataka Goto collection. The Goto collection, which represents 86% of this training set, is the collection on which our algorithm achieved the best results (2nd rank for bass drum and snare drum detection, see figure 1). Our classifiers clearly over-fitted the Goto dataset, or failed to learn the properties of the other collections from the limited number of examples.

3.4 Refinements and improvements.

Several refinements and modifications are thus possible to improve the results of our algorithm.

First of all, a better classification scheme than the 3 parallel classifiers has to be found. This scheme showed its limitations in the addition of an extra class (the hi-hat), and we expect it to give even worse performances when other classes (such as toms or cymbals) will be added. Another possible approach could be to consider each combination of instruments (hi-hat + bass drum for example) as a distinct class.

Secondly, the performances of the onset detection module have to be improved. For this purpose, the availability of the ground truth data (transcription as text files) of the music collection used for the evaluation, as well as the output of our algorithm would be very helpful. This would especially allow us to validate or invalidate the three hypothesis we formulated regarding the abnormally high precision rate.

Then, an efficient decision rule that will not result in a high rejection rate will have to be selected.

Finally, training the algorithm on a larger and more balanaced database will probably improve its performances. We plan to finish the annotation of the hi-hats tracks of the database presented in (Gillet and Richard, 2005), and release it to the contest organisators, for its inclusion in future instances of the contest. Training the algorithm with lower values of the C parameter of support vector machines will also result in better generalization properties - which prove to be an essential factor in the overall performances of the algorithm.

References

O. Gillet and G. Richard. Drum track transcription of polyphonic music using noise subspace projection. In *Submitted to the 6th ISMIR conference 2005*, 2005.