ALTERNATE LEVEL CLUSTERING FOR DRUM TRANSCRIPTION

Mathias Rossignol[*], *Mathieu Lagrange*[*] [†], *Grégoire Lafay*[†], *Emmanouil Benetos*[‡]

[*] IRCAM, CNRS, UPMC
[†] IRCCYN, CNRS, Ecole Centrale de Nantes
[‡] Centre for Digital Music, Queen Mary University of London

ABSTRACT

This paper introduces a clustering-based unsupervised approach to the problem of drum transcription. The proposed method is based on a stack of multiple clustering and segmentation stages that progressively build up meaningful audio events, in a bottom-up fashion. At each level, the inherent redundancy of the repeating events guides the clustering of objects into more complex structures. Comparison with state-of-the-art approaches demonstrate the potential of the proposed approach, both in terms of efficiency and of ability to generalize.

Index Terms— Audio segmentation, automatic music transcription, drum transcription, unsupervised learning.

1. INTRODUCTION

Automatic music transcription (AMT) refers to the process of converting an acoustic musical signal into some form of musical notation, and is considered to be a key problem in the field of music signal processing. A major part of the general AMT problem is that of drum transcription (also called percussion transcription), which refers to the process of locating and identifying acoustic events issued for a drum kit from an acoustic musical mixture. While several algorithmic approaches have been proposed in the literature, the problem is still considered to be open.

Roebel *et al.* [8] propose to divide drum transcription methods into three groups. The first one, usually called *match and adapt*, matches templates of audio events of interests to the signal in order to trigger detection [11]. Similarly, *separate and detect* approaches attempt to isolate templates of audio events from the input signal, using extended versions of Non-negative Matrix Factorisation (NMF) [7]. *Separate and detect* methods include Non-negative Matrix Deconvolution [8] and Probabilistic Latent Component Analysis [1, 9]. Even though significant progress has been made towards the latter approaches [6], they remain strongly dependent upon the prior knowledge used to acquire templates and, to the best of our knowledge, little is known with respect to the generalization capabilities of such approaches. Consequently, we believe that there is a need both for more in-depth experimental validation of such approaches and the development of purely unsupervized approaches that rely on fewer priors. The approach introduced here belongs to the third scheme, *segment and cluster*, and it only relies on the use of templates at a later stage, once all the audio signal has been segmented and events of interest organized into clusters, maximizing generalization potential. Thus, the main contributions of this paper are: 1) introducing a semi-supervised approach to the drum transcription problem, 2) studying the generalization capabilities of both a state-of-the-art supervised method and the proposed approach.

Partially funded by ANR Houle under reference ANR-11-JS03-005-01. EB is supported by a Royal Academy of Engineering Research Fellowship.

2. PROPOSED APPROACH

We propose to represent an audio scene in a hierarchical manner, as a tree whose nodes correspond to audio fragments; the nodes of each level are created by concatenating on the temporal axis nodes of the level immediately below it, with the first level being composed of elementary overlapping frames of the digital recording.

This is in keeping with Bregman's consideration of audio scene perception, taking the form of an iterative object-forming process occurring at different time-scales [2]. Objects formed at small time scales are themselves embedded in larger ones based on their spectral similarity and time proximity. The proposed clustering-based approach follows this framework.

2.1. Alternate Clustering

Having chosen to adopt a bottom-up approach of progressive agglomeration of objects into bigger and bigger ones, the practical question to tackle is: at level n of the analysis, given a sequence of objects (be they elementary frames or complex objects), which ones should we merge to produce objects at level $n + 1$? To take this decision, we consider several cues, coming from classical Gestalt theory:

- *continuity* means that we preferably identify as objects sound fragments without breaks,

- *similarity* means that in cases when there are breaks in the object, all its components should still share common spectral properties, since produced by a same source,
- the *structural* cue, finally, reflects the way that listeners can recognize complex objects by identifying repeated patterns: for example, complex mechanical sounds or animal calls can be made up of several non-continuous fragments.

Although the cues of continuity and similarity, purely based on spectral content, can directly be considered in a frame based manner by computing similarities between object features, taking into account structural cues requires a more indirect approach. In order to identify repeated patterns in the sequence of audio fragments, we first cluster them into k classes, thus allowing a representation of that sequence using an alphabet of k symbols. Such quantization schemes are quite powerful and have been considered in many other fields, from time series matching [5] to Asian language processing [10]. We then make use of the sequential mutual information between symbols to identify which ones have a strong bond and are likely to belong to a same object.

The Alternate Level Clustering (ALC) algorithm reads:

1. **Initialize** by a trivial clustering: group successive audio frames that have a very high spectral similarity to form "level 0" objects
2. **Repeat** for the specified number of levels:
   - Compute spectral similarities between objects (see below for details)
   - Cluster objects into classes C_i following those similarities, using the kernel k-means algorithm
   - Compute the mutual information (MI) between C_i, based on how often an instance of C_j follows an instance of C_i:

$$MI\left(C_i, C_j\right) = \log\left(\frac{p\left(C_i C_j\right)}{p\left(C_i\right) p\left(C_j\right)}\right)$$

     where $p(C)$ is the probabilities that a given object belong to a given class, and $p(C_iC_j)$ is the probability that a sequence of two consecutive objects is an instance of C_i followed by an instance of C_j
   - Generate a decision curve along the time axis, whose value between two consecutive objects is a weighted average of:
     - the sequential MI between the two classes those objects belong to,
     - the spectral continuity, computed as the spectral similarity between the last frame of one object and the first frame of the next.
   - Segment the sequence of objects based on this decision curve, by merging objects appearing between local minima.

Spectral similarity is computed from the features of the objects using the Dynamic Time Warping (DTW) distance in order to account for timing discrepancy. To produce usable results, a final clustering into the number of desired output classes is performed on the objects produced at the last level, using the same similarity and algorithm as for intermediary level clusterings.

Figure 1 gives an example of the kind of analysis ALC can perform: at the lowest level (bottom of the figure) very small fragments start being grouped into more coherent units. At the intermediate level, the fragments are still very small, but already large enough to be clustered in a somewhat meaningful way. At the top level, event onsets can be found, and fragments are grouped according to the type of object they belong to. In this example, the configuration of the algorithm leads to an over-segmentation of the signal, and in the last level the clustering correctly identifies hi-hat hits (blue class), distinguishes between two types of snare drum hits (orange and purple classes, actually identifiable by ear, and visually on the signal image) but confuses the bass drum (in yellow on the signal) with the background signal (light pink class).

It is important to note that our approach is not simply one of standard hierarchical classification, where objects are gathered into bigger and bigger classes of similar objects: here, the objects are not simply put together in a set, they are merged to create larger objects. The hierarchy is therefore one of increasing size and complexity of objects, that also implies a difference in nature: objects at one level are essentially different from objects at another.

For the drum transcription task, we experimentally set the number of levels to 3. The second level considers 50 classes to account for the diversity of the musical background. The last level considers 4 classes, one for each drum event and one for the musical background.

2.2. Object Matching

The proposed method is completely unsupervised, that is, objects are built in a bottom-up fashion. In order to assign those objects to the target classes, one has to resort to the use of prior knowledge. Instead of following the classification approach [4] that would require an actual modeling of the background music, we adopt a semi-constrained clustering approach: for each desired class, some provided samples are considered as reference objects. The resulting objects built by the ALC algorithm (scene objects) and those reference objects are then clustered together using the kernel k-means algorithm under the DTW distance. The clustering is constrained by providing the following allocation as an initialization: each reference object is assigned to its drum class ("hi hat", "bass drum" or "snare drum") and the scene objects are assigned to the background class. By doing so, the clustering algorithm iteratively assigns some scene objects to the drum classes they correspond to before reaching convergence.

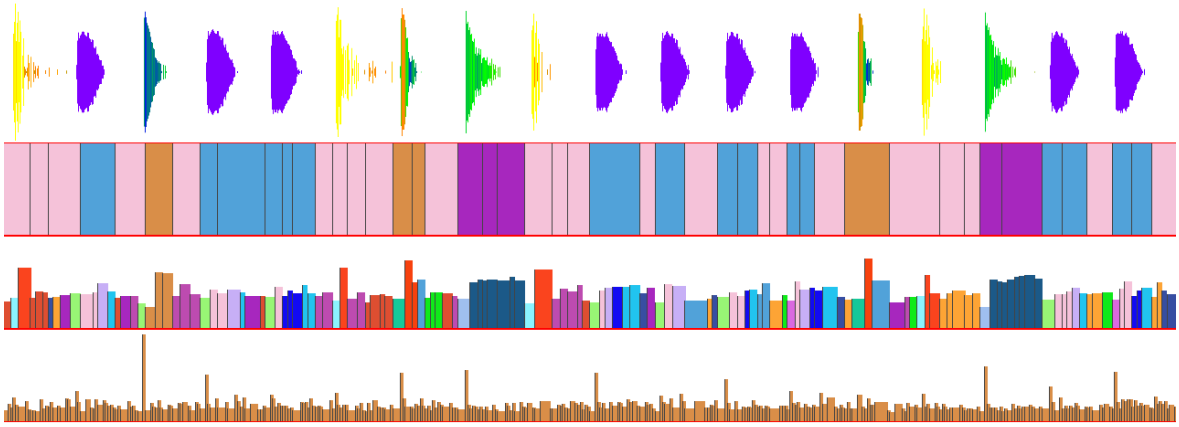

Fig. 1. Sample of ALC analysis. The waveform is given for reference and colored according to zero-crossing rate. At each level, the bar diagram indicates, in width the extent of objects produced at this level, in height the values of the generated decision curve at that level, and with colors the classes the objects belong to.

3. EXPERIMENTAL PROTOCOL

Simulated acoustic scenes are considered to evaluate the algorithm performance. A simulated acoustic scene may be seen as an audio file built from isolated recordings of " events " and " backgrounds ". In the case of the drum transcription task, single sounds of drum components such as hi-hat, snare drum or bass drum are considered as audio events, whereas pitched sounds of other instruments are considered as background. The use of simulated scenes allows us to refine the algorithm performance analysis, by controlling simulation parameters such as Event to Background Ratio (EBR)[1].

3.1. Datasets

Two distinct corpora are used to simulate the scenes, respectively the drum kit components 2 and 3 of the ENST-Drums database [2]. Each corpus is made of single drum sounds of closed hi-hat (hh), snare drum (sd) and bass drum (bd). The use of those three classes follows common practice in drum transcription algorithms evaluation [8]. Additionally, we use two musical backgrounds respectively named bg_{easy} and bg_{hard}, the latter having been found more complex than the former by experimenters. All single drum sounds have been cut to 250 ms and faded-out using a 10 ms window. Backgrounds were long enough to fit the evaluation scene duration, which has been set to 30 seconds.

Drum event selection and inter-onsets spacing between events are managed class-per-class. Drum events are randomly chosen using pseudo-random integers drawn from a discrete uniform distribution, while inter-onsets values are drawn from a normal distribution with mean $\mu = 1$ seconds, and standard deviation $\sigma = 0.25$ seconds. To better simulate a natural scene, separate occurrences of a same event category are synthesized using randomly chosen distinct sound samples. As the drum events have the same duration, all classes are equally represented in each scene.

Two datasets are built: one with drum kit 2 and bg_{easy} and another one with drum kit 3 and bg_{hard}. In the experiments dedicated to supervised approaches, the former is used as a training set and thus termed ' Train ", and the latter as a testing set and termed " Test ". For each of those datasets, 7 different EBRs are considered: -24 dB, -12 dB, -6 dB, 0 dB, 6 dB, 12 dB and 24 dB[3]. The simulation process has been replicated ten times for each drum kit/background and EBR conditions, yielding 28 datasets (4×7) composed of 10 scenes each. It should be noted that we seeded the pseudo-random generator before changing the corpus or the EBRs so that :

- each related scene of each drum kit/background and EBR conditions has the same temporal structure of events;
- each related scene of each EBR conditions for one drum kit/background condition has the same temporal structure of events, and is composed of the same sound samples.

3.2. Metrics

The evaluation is performed using onset-based Precision, Recall and F-measure (P-R-F), as defined in [3], with a 100 ms tolerance window. Multiple onsets detected within the same window are considered as false alarms. P-R-F are computed both for all classes (F), and in a class-wise way (Fsd, Fbd and Fhh). Additionally we use a class-blind F-measure (F_{onset}) as defined in the MIREX audio onset detec-

[1] The simulator is available at: `https://bitbucket.org/mlagrange/simscene`

[2] `http://www.tsi.telecom-paristech.fr/aao/2009/11/25/enst-drums-pistes-de-batterie-annotees`

[3] Those datasets are available at: `http://archive.org/details/simSceneDrum2015`

tion evaluation, which only evaluates the simpler task of onset presence /absence.

3.3. Methods

Scenes are represented by spectral features, namely the log power spectrogram, the Mel Frequency Cepstral Coefficients (MFCCs), and the Constant-Q transform (CQT), computed using standard libraries and parameters, for fair comparison.

Three methods with different flavors are considered. First, the proposed approach in its direct unsupervised form detailed in Section 2.1, and its semi-supervised form presented in Section 2.2. For this algorithm, MFCC features give the best overall performance.

Second, in order to compare with another unsupervised approach, we consider a standard Non-negative Matrix Factorization (NMF) scheme with 4 basis vectors. Clustering is performed by selecting at each frame the basis with the highest activation. Euclidean and Kullback-Leibler divergences were evaluated over the four types of features, and the best results were obtained with Euclidean distance over the log power spectrogram.

Third, a recent supervised Probabilistic Latent Component Analysis (PLCA) algorithm [1], designed for transcribing both drum and pitched sounds, is considered. The method takes as input a CQT spectrogram of an audio signal and decomposes it into pitched and unpitched part, by relying on a dictionary of harmonic and drum templates. The algorithm returns a " piano-roll " representation for the pitched part and a " drum-roll " representation for the drum kit components. For this paper, we used 2 variants of the method in [1]: one with a dictionary of drum sounds for bass drum, snare drum, and hi-hat extracted from isolated drum samples of the RWC database [4], and another system for which the dictionary was trained using samples from the Train dataset. The latter system thus gives an indication of the upper performance limit of a supervised approach. As the method's activation curves have to be thresholded to trigger the detection of drum events, the thresholds are determined by optimizing the F-measure over a specified development set.

4. RESULTS

Several experiments are conducted to validate the proposed approach. First, the proposed approach in its unsupervised form is compared to an unsupervised NMF scheme. Then, several variants of the supervised PLCA scheme of [1] are studied. With those results as context, we then present the results achieved by the semi-supervised proposed approach with object reallocation. Unless otherwise stated, the results are given for a 6 dB EBR, which corresponds to standard music production ratio between drums and other instruments.

[4] https://staff.aist.go.jp/m.goto/RWC-MDB

	dataset	F (%)	Fbd (%)	Fhh (%)	Fsd (%)	F_{onset} (%)
NMF	train	10.4±1.4	10.6±2.0	10.6±2.1	10.2±2.2	14.8±2.1
	test	9.8±0.9	10.0±2.1	10.6±1.6	9.5±1.4	12.1±1.3
ALC	train	23.3±1.8	24.7±3.9	25.5±4.8	26.3±7.0	**56.4±1.6**
	test	21.2±2.3	24.7±4.5	25.8±4.4	23.0±4.1	**55.9±1.6**

Table 1. Performance of the NMF and the proposed unsupervised scheme.

Te	Th	d.set	F (%)	Fbd (%)	Fhh (%)	Fsd (%)
tr	tr	train	96.6±0.6	93.0±1.9	97.6±1.7	98.8±0.8
tr	tr	test	70.6±2.5	57.8±7.6	60.2±1.4	99.7±0.7
o	tr	train	61.0±4.8	72.9±6.6	42.8±3.7	75.0±11.0
o	tr	test	37.0±2.8	0.8±1.7	38.1±2.9	62.3±10.7
o	o	train	48.1±5.7	73.7±7.5	2.6±2.8	52.0±10.9
o	o	test	5.2±1.8	0.0±0.0	10.4±5.0	4.5±2.2

Table 2. Performance of the PLCA method. *Te*: template, *Th*: threshold, *o*: original values, *tr*: optimized on dataset

4.1. Unsupervised approaches

As those methods are purely unsupervised, the predicted classes have to be aligned to the ground truth classes. This matching is done by selecting the optimal permutation leading to the best classification accuracy.

Even with optimal alignment, the NMF scheme does not perform satisfactorily. The proposed approach with its hierarchical structure behaves better, as shown on Table 1. Most notably, the detection of the onsets (F_{onset}) is much better, which leads us to the conclusion that the actual assignment of objects to drum or background classes could benefit from a semi-supervised scheme. Also, it appears the test dataset is somewhat harder than the train one as both methods behave slightly worse on the latter.

4.2. Supervised approaches

The PLCA algorithm requires prior knowledge of what shall be detected (Templates) and which level of activation triggers detection (Thresholds). Each can be taken directly from [1] or optimized on the training dataset.

As can be seen on Table 2, the closer the prior knowledge is to the evaluation data, the better the performance. Closer inspection shows that drum kit 2 is quite comparable with the RWC drum kit, whereas drum kit 3 is quite different, which may explain the drop of performance between the train and test datasets for the original templates, especially pronounced for the o-o approach, where templates and threshold have been optimized on the RWC drum kit. The major drop in performance on the last line of the table shows the often neglected importance of threshold tuning for such approaches.

dataset	F (%)	Fbd (%)	Fhh (%)	Fsd (%)
train	34.2±4.6	5.7±9.7	59.7±18.3	29.5±15.6
test	26.3±9.0	9.8±12.8	44.2±31.3	6.8±14.8

Table 3. Performance of the proposed system (ALC) with object reallocation.

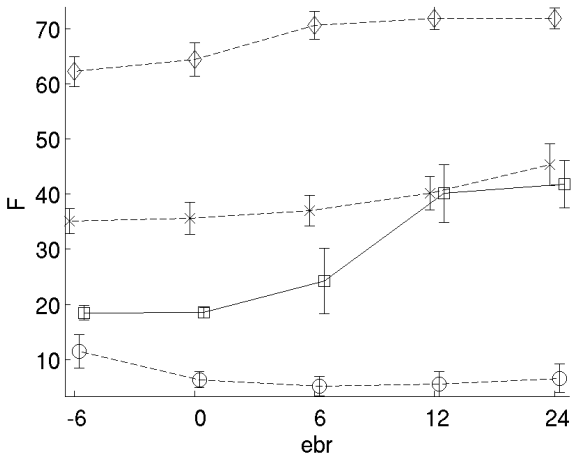

Fig. 2. Performance on the test set of the proposed system (solid line) and the PLCA system (dashed line) with o-o (circle), o-tr (cross) and tr-tr (diamond) conditions.

4.3. Semi-supervised approaches

Drum kit 2 is used as reference for the object reallocation system described in Section 2.2. As can be seen on Tables 1 and 3, hi-hat hits are best matched, bass drum the worst; this is probably an effect of the greater contrast to background of the former. Figure 2 compares the proposed approach and the PLCA approach for several EBRs in terms of global F-measure on the test dataset. Notably, the object reallocation system efficiency decreases with lower EBRs, most probably due to the difference between clean reference objects and extracted noisy objects. Adaptation of the reference objects with addition of musical background during matching shall be investigated in the future to reduce this discrepancy.

5. CONCLUSION

We introduced a hierarchical alternate clustering scheme suitable for detecting drum events in musical acoustic scenes. Based on Gestalt-like principle, the proposed algorithm is quite generic and flexible in terms of design.

Experiments on controlled simulated data shows that state-of-the-art supervised approaches based on matrix factorization require a significant degree of compatibility between training and testing data to perform well. The proposed approach, by relying less on trained priors, is more stable while facing new types of data, even though its raw performance still needs to be further improved.

Future work will include more extensive validation on real data, improved fusion of segmentation cues, and better modelling of prior knowledge in the object reallocation.

REFERENCES

[1] E. Benetos, S. Ewert, and T. Weyde. Automatic transcription of pitched and unpitched sounds from polyphonic music. In *IEEE ICASSP*, pages 3107–3111, Florence, Italy, May 2014.

[2] A. S. Bregman. *Auditory scene analysis: The perceptual organization of sound*, page 644. MIT press, 1994.

[3] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, Mathieu Lagrange, and Mark D Plumbley. Detection and classification of acoustic scenes and events: An ieee aasp challenge. In *IEEE WASPAA*, 2013.

[4] O. Gillet and G. Richard. Automatic transcription of drum loops. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, volume 4, pages iv–269–iv–272 vol.4, May 2004.

[5] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. Experiencing sax: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15(2):107–144, 2007.

[6] H. Lindsay-Smith, S. McDonald, and M. Sandler. Drumkit transcription via convolutive NMF. In *International Conference on Digital Audio Effects (DAFx)*, York, UK, September 2012.

[7] Jouni P. and Tuomas V. Drum transcription with non-negative spectrogram factorisation. In *European Signal Processing Conference (EUSIPCO)*, pages 65–69, Antalya, Turkey, september 2005.

[8] A. Roebel, Pons J., M. Liuni, and M. Lagrange. On automatic drum transcription using non-negative matrix deconvolution and itakura saito divergence. In *ICASSP*, Brisbane, Australia, April 2015.

[9] Madhusudana Shashanka, Bhiksha Raj, and Paris Smaragdis. Probabilistic latent variable models as non-negative factorizations. *Computational intelligence and neuroscience*, 2008. Article ID 947438.

[10] H. Wang, J. Zhu, S. Tang, and X. Fan. A new unsupervised approach to word segmentation. *Computational Linguistics*, 37(3):421–454, 2011.

[11] K. Yoshii, M. Goto, and H. G. Okuno. Adamast: A drum sound recognizer based on adaptation and matching of spectrogram templates. *1st Annual Music Information Retrieval Evaluation eXchange (MIREX)*, 2005.