Interactive Segmentation of Electro-Acoustic Music

S. Gulluni[1,2], S. Essid[2], O. Buisson[1], and G. Richard[2]

[1]Institut National de l'Audiovisuel
4, avenue de l'Europe 94366 Bry-sur-marne Cedex, France
[2]Institut Telecom, Telecom ParisTech, CNRS/LTCI
37 rue Dareau, 75014 Paris, France

Abstract. In this paper, we present an interactive approach for the segmentation of Electro-Acoustic music by focusing on the local timbre variations. For this purpose, a preliminary low-level segmentation is achieved with an onset detector and an agglomerative hierarchical clustering algorithm is used to cluster inter-onset segments. Subsequently, two alternative scenarios are compared in which the user feedback is used to update clustering. Evaluation is done on both synthetic and real world data. The results show that a simple interactive method can improve significantly the performances of the resulting segmentation.

Key words: music segmentation, clustering, relevance feedback

1 Introduction

In marked contrast to other more conventional musical forms, the composers of *Electro-Acoustic Music* work directly with the "sound material" using recording techniques. Apart from a very few exceptions, the composers have not created a symbolic representation of their pieces that could be assimilated to a score sheet. This renders the analysis and study of this type of music quite complex and totally user-centered, hence our work towards developing adaptative systems capable of analyzing and structuring *Electro-Acoustic Music* in a semi-automatic fashion using user relevance feedback, which to the best of our knowledge is completely novel.

In the present work, we focus on one of the most important structural aspects of contemporary music, i.e. Timbre [5]. The importance of low level timbral features has already been evidenced in previous works [6,7,8] on music segmentation and structuring where the aim is to decompose a conventional piece (generally pop/rock music) into high level sections such as *intro*, *verse*, *bridge*, ... However our concern is significantly different from such proposals in the sense that our primary objective is to obtain a segmentation of the music into short-term segments of homogeneous timbre in a non-supervised fashion. Here, segments duration may vary from fractions of seconds to tens of seconds. For exemple, a sequence of a piano note, a synthetic sound, a trumpet phrase and again a piano note would be labelled: $Timbre1$, $Timbre2$, $Timbre3$, $Timbre1$. The diversity of

timbres to be considered (natural environmental sounds, synthetic sounds and instrumental sounds) may open new fields of applications such as the analysis of audio documents which are not structured in the same way as conventional music (movie soundtracks, audiovisual documents or ambiances).

We propose a simple yet novel timbre segmentation architecture based on an original interactive clustering method that allows us to achieve high accuracy by exploiting minimal user feedback (see Figure 1). It is shown that very simple, hence tolerable feedback significantly improves the segmentation performance compared to a classic totally automatic approach.

The paper is organized as follows: Section 2 describes the first stages of the system including Segmentation and Feature Extraction. Then, Section 3 is dedicated to Interactive Clustering. Section 4 describes the evaluation of the system and the last section is the conclusion of this work.

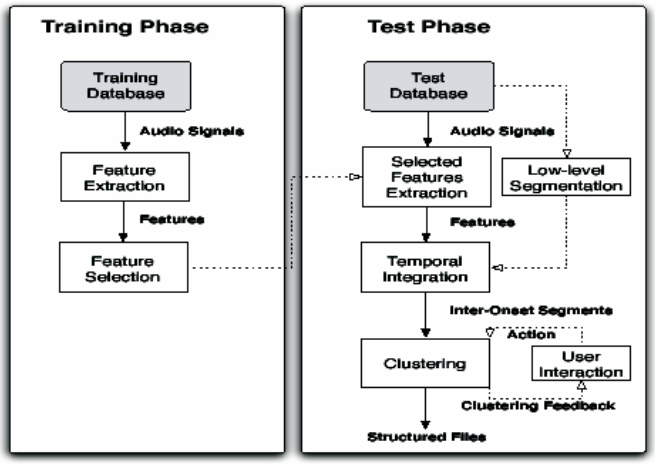

Fig. 1. Architecture of the Timbre Segmentation System.

2 Segmental feature extraction

The goal is to build an efficient timbre descriptor for our task from a set of low-level features. Prior to feature extraction, the signal is downsampled to 22 kHz and split in 20ms frames with 50% overlap.

A total number of 280 feature coefficients were extracted from a wide set of classic audio features (see `http://www.tsi.enst.fr/~gulluni/iseam/` for more details on the extracted features).

The classic Fisher algorithm was then used for automatic feature selection as in [2]. As a result, a reduced vector of 30 coefficients is obtained[1].

[1] The number of selected features was varied from 10 to 40 in a preliminary phase which showed that keeping the best 30 first features was a good choice.

Following [2], we perform a temporal segmentation in order to obtain perceptually relevant sonic units. For this task, the onset detector described in [1] was used. This method is based on the spectral energy flux (temporal derivative of the spectrum). The resulting detection function was used to obtain inter-onset segments. Each inter-onset segment S_i is defined as a collection of d-dimensional vectors: $S_i = (x_{i1}, x_{i2}...x_{in})$. We adopted a simple but efficient temporal integration to sum up the segments. A recent study [2] showed that the reduction of a segment to mean and standard deviation of the vectors gives quite good results compared to more complex methods. As a result, we reduced the inter-onset segment to vectors $X_i = (\mu_i, \sigma_i)$ where μ_i and σ_i are the mean and standard deviation vectors of S_i.

3 Interactive Clustering

3.1 Agglomerative Hierarchical Clustering

The clustering approach follows a classical bottom-up strategy: each observation starts with its own cluster and then the closest pairs of clusters are merged as a bigger one. Consequently, the extreme top of the hierarchy (root) is related to the whole set of observations and similarly the bottom (leaves) to a unique vector (see [4] for more details). To compare clusters, the hierarchical clustering uses a linkage criterion which computes the distance between two sets of vectors A and B as a function of the pairwise distances between observations. For this, the average criterion was used and is defined as follows: $\frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} L_2(a,b)$, where L_2 is the Euclidian distance in our case. The result of this clustering can be viewed as a binary tree where each node is associated to a set of observations. Dendrograms are also used for representation with the benefit of indicating the distances between the objects represented by the vertical axis (Figure 2).

3.2 Improving the clustering by user interaction

To allow the user to efficiently interact with the automatically generated segmentation, a simple, yet intuitive, representation was designed where different colors are used to represent a cluster on a spectrogram (Figure 2).

For the initial clustering, it is assumed that the user approximatively knows how many different "timbres" are in the musical piece. Traditionally, a constant threshold is used to extract the subtree (e.g. *global cut*) yielding the desired clustering. We propose to use a variable threshold which can be defined from user feedback, therefore allowing *local cuts* (Figure 2).

Two alternative scenarios have been considered. In the first scenario, we allow the user either to break or merge timbre segments resulting from the initial clustering. The user can choose the segment he/she wants to rectify (either because two or more different timbre classes have been affected to the same segment or a contiguous timbre segment has been erroneously fragmented). Given that each segment is related to a cluster, the user feedback is taken into account by:

- breaking a cluster into its two children subclusters when the user wants to break a segment (Figure 2),
- merging the appropriate clusters when the user decides to merge two segments.

The second scenario is simpler. In the latter, we consider only one action: at each iteration the user marks the most erratic segment and the system breaks the related cluster (as done in the first scenario). This action is more closely related to a classic relevance feedback since the user marks negatively the most irrelevant segment.

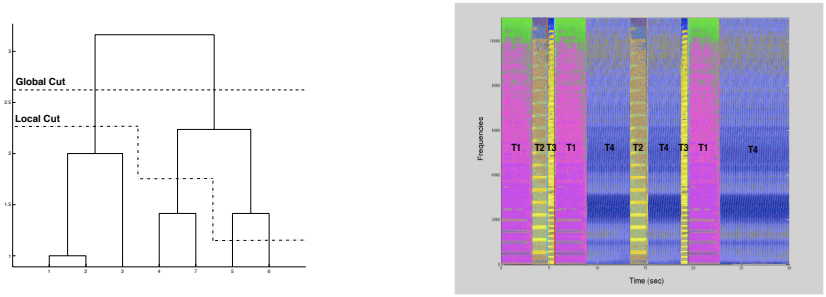

Fig. 2. A dendrogram with global and local cuts (left), Basic representation of timbre clustering (right)

4 Evaluation

4.1 Corpus and Evaluation Criteria

Only a few annotations of *Electro-Acoustic Music* exist and most of them mix interpretation information and description of the different timbres of the piece. The extreme tediousness of *Electro-Acoustic Music* annotation led us to opt for a more pragmatic solution based on a synthetically generated corpus.

This corpus was generated by concatenating sound files extracted from two sound banks mimicking the composition process of some *Musique concrete* (a form of electro-acoustic music [5]) pieces such as *Timbre-durees* (TD) from *Olivier Messiaen*. The sounds are randomly extracted from a collection of environmental and ambiance sounds (INA) and from the RWC instrumental database [3] which contains most instruments of the orchestra.

The resulting corpus gathers 1000 files of 30s. The first 200 were used for the feature selection and the other 800 for testing the algorithm.

Additionally, the musical piece *Timbre-durees* which only contains juxtaposed timbres (only one timbre at a time) was manually annotated to provide preliminary validation on a real composition.

The evaluation of the system is done by comparing the ground truth or classes to the resulting clusters. In the synthetic part of the corpus, the ground truth is created at the same time as the signal generation. The comparison of the ground truth classes and the clustering is not direct, it involves two main

problems: the number of clusters may differ from the number of classes (it is often the case), there is no correspondences between the classes and the labels of the clusters. Consequently, it is necessary to associate each class to a relevant cluster. For this, each class C_i is associated to the cluster W_j which contains the greatest number of frames belonging to this class. Accordingly, the classic recall (R_i) and precision (P_i) measures used in information retrieval are defined as follows: $R_i = \frac{\max_j |C_i \cap W_j|}{|C_i|}$ and $P_i = \frac{\max_j |C_i \cap W_j|}{|W_J|}$, with $J = \arg\max_j |C_i \cap W_j|$. Then, we can compute the overall F-measure as $Fm = \frac{2RP}{R+P}$ where R and P are respectively the average of R_i and P_i on all classes.

To evaluate the impact of user interaction, we take advantage of the fact that the desired segmentation leaves no room for subjective interpretation, i.e. there is only one possible correct segmentation for each file, that is supposed to be formed straightforward by any user. Thus, given the ground truth, it is possible to simulate the behavior of an ideal user. We consider that the user begins to rectify the most erratic segments (the segments with the maximum number of wrong labeled frames in comparison to the ground truth). The clustering is then updated and a new version presented to the user who will repeat the same process with the new proposal until he/she reaches the ground-truth segmentation.

4.2 Experiments

We first evaluate the impact of the chosen number of clusters. It was observed that an improvement is obtained by increasing the number of clusters beyond the correct number of classes (N_C), up to a certain limit. We observed that the best results were obtained for $N_C + 4$ with a reference f-measure score of 0.82. A careful review of the output segmentation, showed that some clusters were formed from low energy frames. This motivates future work on temporal smoothing techniques to address this issue.

The results of the interactive experiments are first given under the form of F-measure averages as a function of the number of user feedback iterations (see Figure 3). These results show that the second method gives the best results

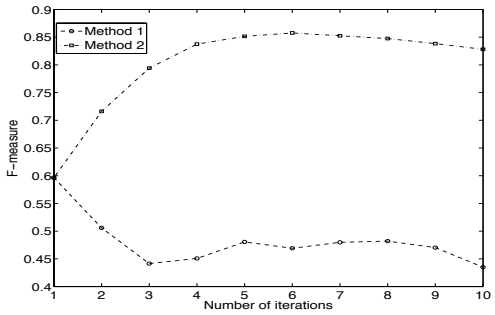

Fig. 3. Evolution of the F-measure as a function of the number of iterations

and that the interaction allows us to improve the initial segmentation. We experimentally observed that the fusion of clusters added instability to the system which explains the performance degradation with the first interaction mode.

These results are confirmed by the evolution of the average of the F-measures maxima regardless of the number of iterations. Starting from a reference score of 0.82, the first and second methods obtain respectively 0.71 and 0.9 and improve 34,2% and 92,5% of the reference scores.

The second method was also applied to *Timbre-Duree* and we observed an improvement of 4% of the reference score (from 0.67 to 0.71).

5 Conclusion

In this paper, we have proposed a system for interactively segmenting a musical piece into homogenous timbre sections for application to *Electro-Acoustic Music*. We have considered two strategies of interaction for improving clustering by user relevance feedback. The results have showed that a simple method can significantly improve the performance by relying on limited user actions. However, despite of its efficiency, the proposed clustering method appears to be limited because of the rigid tree structure on which it is based. As a matter of fact, this structure limits the clustering to the possible partitions (defined by the tree). In future work, we will consider more advanced relevance feedback strategies using active learning techniques which could take advantage of the outputs of the system proposed in this paper. We will also include other descriptive aspects of music in the system like dynamic, harmonicity or rythm.

References

1. Alonso, M., Richard, G., David, B.: Extracting note onsets from musical recordings. In: Proceedings of IEEE International Conference on Multimedia and Expo. Amsterdam (2005)
2. Joder, C. , Essid, S., Richard, G.: Temporal Integration for Audio Classification with Application to Musical Instrument Classification. IEEE Transactions on Audio, Speech and Language Processing.Vol. 17, No. 1., pp. 174-186 (2009)
3. Goto, M., Hashiguchi, H., Nishimura, T., Oka, R.: RWC Music Database: Popular, Classical, and Jazz Music Databases. Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002), pp.287–288 (2002)
4. Jain, A. K., Murty, M. N., Flynn, P. J.: Data Clustering: A Review. ACM Comput. Surv., Vol. 31, No. 3., pp. 264–323 (1999).
5. De Leeuw, T., Groot, R.:Music of the twentieth century: a study of its elements and structure. Amsterdam University Press (2006)
6. Jensen, K.: Multiple scale music segmentation using rhythm, timbre, and harmony. EURASIP J. Appl. Signal Process, New York (2007)
7. Levy, M., Noland, K., Sandler, M.: A Comparison of Timbral and Harmonic Music Segmentation Algorithms. ICASSP 2007. IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. pp.IV-1433-IV-1436 (2007)
8. Cheng, H.-T. , Yang, Y.-H. , Lin, Y.-C., Chen, H.-H.: Multimodal structure segmentation and analysis of music using audio and textual information. IEEE Int. Symp. Circuits and Systems, Taipei (2009)