AN INTERACTIVE SYSTEM FOR ELECTRO-ACOUSTIC MUSIC ANALYSIS

Sébastien Gulluni, Olivier Buisson
Institut National de l'Audiovisuel
Bry-sur-marne, France
{sgulluni,obuisson}@ina.fr

Slim Essid, Gaël Richard
Institut Telecom, Telecom ParisTech,
Paris, France
{slim.essid,gael.richard}@telecom-paristech.fr

ABSTRACT

This paper, presents an interactive approach for the analysis of electro-acoustic music. An original classification scheme is devised using relevance feedback and active-learning segment selection in an interactive loop. Validation and correction information given by the user is injected in the learning process at each iteration to achieve more accurate classification. An experimental study is conducted to evaluate and compare the different classification and relevance feedback approaches that are envisaged, using a database of polyphonic pieces (with a varying degree of polyphony). The results show that the different approaches are adapted to different applications and they achieve satisfying performance in a reasonable number of iterations.

1. INTRODUCTION

Being composed directly with the "sound material" using recording techniques [18], electro-acoustic music differs from other more conventional musical forms. Composers of the genre do not use score sheets to write music and there is no common agreement on a standard notation system to be used to create symbolic representations for such compositions. Electro-acoustic music is traditionally organized in *sound objects*. Here, we define "sound object" as any sound event perceived as a whole [18]. Most of the time a musical piece does not expose separate sound objects as simultaneous sounds are masking each others due to polyphony. Consequently, the analysis of this music is quite complex and totally user-centered as it is essentially concerned with the subjective identification of sound objects of interest to the user. The reader can refer to [1] for examples of electro-acoustic compositions.

This work presents an interactive classification system for electro-acoustic music analysis using relevance feedback. In previous works, relevance feedback has been widely used in content-based image retrieval tasks (see [4] for an overview). By contrast, in the field of music information retrieval, relevance feedback and active learning have only been exploited in a few music information retrieval studies, for pop music retrieval based on user preferences [11] or mood/style classification [15]. More closely related works in this field have focused on "standard" instruments and percussion timbre classification [7, 8, 14] by building supervised systems based on large databases. In the electro-acoustic case, composers exploit various sound sources and one does not have a-priori knowledge about these sources which are most of the time polyphonic and heterogeneous.

In this paper, following our previous works [9, 10], we propose a complete system for electro-acoustic music analysis, and evaluate and compare different relevance feedback approaches to our problem. The initialisation of the system is achieved through an interactive segmentation phase (mostly similar to [9]) to obtain initial texture segments (see Figure 1 and 2). Then, these segments are processed by an interactive classification module using relevance feedback and active learning segment selection. From a user's point of view, the search for a target sound object begins with the selection of a characteristic segment for each sound class. Then, the system enters in an interaction loop and suggests, at each iteration, segments to be annotated by the user so as to make learning progress. On each new proposed segment, the user can correct the system's label prediction. The interaction loop ends when the user is satisfied with the labels. We compare different classification and relevance feedback approaches for different degrees of polyphonic complexity. This study shows that different methods are more adapted to different applications.

The paper is organized as follows: Section 2 presents the musical motivations and the results of musicologists' interviews that were carried out to acquire prior knowledge on their approach to the analysis of electro-acoustic music. Section 3 describes the interactive system including the user interaction scenario and active learning segment selection strategy. Section 4 is dedicated to the evaluation of the method and the last section suggests some conclusions.

2. MUSICAL MOTIVATIONS

Our work attempts to address musicologists's need for new tools for the analysis of non written music. Thus, for a better understanding of their expectations, interviews were held with three musicologists with a special expertise in electro-acoustic music analysis. The questions were about their personal methodology for analysis and the utility of computer-based sound analysis tools to their work. By analyzing their answers, some common habits can be identified in their methodologies. Of note is the fact that they always listen to the whole piece from 4 to 10 times to locate prominent sound objects and build a viewpoint to begin the analysis. Another common habit is to listen to the same piece several times and focus on one sound category at each time. In all the interviews, the musicologists approach the analysis as a *sound object transcription* task . For some of them, the transcription helps forming a viewpoint of the piece being analysed, whereas the others already have one when they begin the transcription. All the subjects mentioned that they do not transcribe all sound objects of the piece but only those which are useful for their personal analysis viewpoint.

For the question about the utility of computer-based tools, they expressed some wishes which are all related to the *sound object transcription*. The first was to locate the main sound objects of the piece and help them verify their transcription. Another important wish was to find all the instances of one sound object by giving a segment of the target sound to the tool. This function could also help them to discover sound instances that they did not notice.

This work takes those musical motivations into account and proposes an interactive system for helping musicologists in the transcription task.

3. INTERACTIVE CLASSIFICATION SYSTEM

In this section, we describe all the aspects of the system including the expression of the user point of view.

3.1 Architecture

Figure 2 (A) is a representation of a polyphonic piece which involves potential sound masking: the distinct sound layers are arranged in parallel timelines (one for each sound class). The goal of the transcription is to mark the presence of all target sounds in the whole piece. The classification operates on *texture segments*, *i.e.* temporal fragments of homogeneous timbre (as shown with vertical red lines in Figure 2). The system architecture is divided in two distinct parts: the *initialisation* and the *interaction loop* which performs the classification of the *texture segments* and asks feedback from the user. We compare two different *interaction loop* approaches in this work. The first approach is *multi-pass*: the interaction loop focuses on one sound class at each pass, following the habits of the musicologists who are used to listen to the same piece several times and focus on one sound category for one listening (see Section 2). The other approach is referred to as *one-pass*: the interaction loop considers all the sound objects simultaneously at each time and consequently the user feedback applies to all the classes of interest.

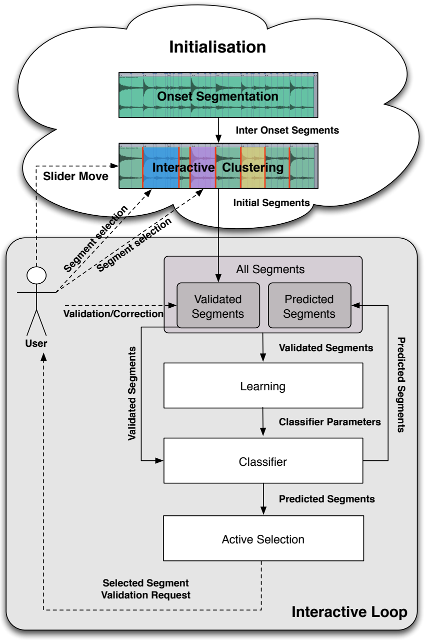

Figure 1. Overview of the interactive system

The interactions of the user with the system can be summarized as follows:

1. Initialisation
   (a) The system starts with an interactive segmentation phase. If the user is transcribing N classes, for each class C_i a characteristic segment $\mathcal{S}_i$ is associated with $i \in \{1...N\}$ (see Figure 2). In order to obtain the initial characteristic segments of all the sound classes corresponding to the user's point of view, in this first interaction phase, the user moves a slider which controls the global segmentation level until the most adapted segmentation is reached. *Texture segments* are created from this segmentation.
   (b) The user selects a characteristic *texture segment* $\mathcal{S}_i$ for each target sound class
2. Interaction loop
   (a) The system learns from the validated segments and enters in the classification process to auto-

matically predict labels for the remaining parts of the signal.

(b) In order to improve the previous classification, the system selects a segment, based on the active learning strategies described in Section 3.3.5, and asks feedback from the user. In the *multi-pass* approach, the system predicts the presence/absence of the current target class and the user validates or corrects the selected segments prediction. In the *one-pass* approach, the user corrects the presence/absence prediction of all the target classes for the selected segment.

(c) In the multi-pass approach, one vs all classification and feedback ((a) and (b)) iterations proceed until the user is satisfied with the result for the current class, before entering a new pass, that is a new interaction loop, for the next classes, until all classes have been covered. In the one-pass approach all classes are considered jointly from the very beginning of the interaction loop and the system iterates multi-class classification and feedback until the user is satisfied with the overall prediction.

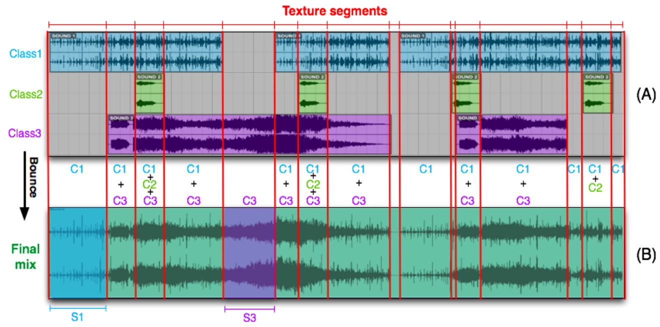

Figure 2. Time-line representation of a polyphonic piece with 3 sound classes and the characteristic segments of the target classes. Though the distinct sound layers are here displayed in parallel time lines (A), in real situations the user can actually only see the final mix made by the composer that appears as a single track (B). The initial user selection and subsequent validations are done by listening.

A total of 217 feature coefficients are extracted from 25 classic audio descriptors on 20 ms windows with 50% overlap, to be used both for the initial segmentation and the subsequent classification. The reader can refer to [6, 17] for a complete description of the features. All the feature vectors used and the corresponding dimensions are listed in the website of the paper [1]. Feature extraction was performed using the YAAFE software [16].

3.2 Interactive Clustering

The goal of the clustering is to obtain a segmentation adapted to the users' point of view as described in the initialisation paragraph of section 3.1. The reader can refer to [9] for a detailed explanation of the initial clustering. First, onset detection is performed and the resulting detection function is used to obtain *inter-onset segments*. Subsequently, a clustering is performed on the *inter-onsets* vectors X_j with an agglomerative hierarchical approach to obtain texture segments. The number of target clusters of the algorithm is controlled by the user in the interface with a slider to obtain an adapted segmentation.

3.3 Classification

In this system, the classification task consists in detecting the presence of given sound classes in every texture segment of the musical piece. Support Vector Machine (SVM) classifiers [2] with probabilistic outputs [2] are used in a "one vs all" fashion. Three different methods are compared to obtain the final prediction: a *multi-pass* approach and two variants of a *one-pass* approach.

3.3.1 Feature Selection

After the initialisation phase, a feature selection based on the Fisher discriminant [5] is performed. The algorithm iteratively selects the attributes which maximize the Fisher discriminant and the d best features are kept to define the feature space for the target class. The parameter d was experimentally determined using a separate database and a value of $d = 10$ has been found to be an appropriate trade-of between performance and complexity. The goal of the selection is to create a relevant descriptor for each sound class. As this selection is part of the *interaction loop*, the sound descriptors may evolve accordingly with the user feedback. This method is adapted to our problem since we do not have prior knowledge on the sound sources.

3.3.2 Multi-pass (MP)

In this approach, the N sound classes are treated sequentially: the user tries to spot all occurrences of the current class C_i before beginning the next class. This enables the user to focus on one sound category at each time following the habits described in Section 2. Therefore, the corresponding *feedback* is quite simple: the user validates/corrects the presence or absence of the current class for the segment selected by *active learning* (see 3.3.5). For the learning phase, positive samples are those which contain the target sound class and negative samples are those which do not. This implies that the positive segments may be complex

[1] http://www.tsi.enst.fr/~gulluni/ismir2k11/

[2] we use the libSVM implementation [3].

sound mixtures which contain other sounds. Using probabilistic SVMs, posterior probabilities $p(C_i|X_k)$ are obtained for each frame observation X_k.

3.3.3 One-pass

In the *one-pass* approach, the classification is carried out as in a "standard" multiclass problem, where all classes are jointly taken into account. Consequently, the user tries to transcribe all the sound classes at the same time and the corresponding feedback requested to the user is to validate/correct the presence or absence of all the sound classes for the selected segment (see 3.3.5). Two classification methods are compared in this approach.

The first one (*one-pass 1*) uses the same classification method as the MP approach: for N sound classes, N classifiers are trained with positive samples being those which contain the target sound class and negative samples being those which do not.

The second method (*one-pass 2*) differs in that it can introduce new classes through the iterations by considering *texture classes* deduced from the user's feedback, *i.e.* for a given feedback iteration, if the user formulates that the corresponding selected segment contains more than one sound class, say classes A and B, a new texture class is created, that is composed of the union of those classes (*i.e.* $A \cup B$), and the corresponding classifier trained. Hence M classifiers are here used in the polyphonic case, with $N \leq M \leq 2^N$.

3.3.4 Segment-level predictions

Given the posterior probability $p(C_i|X_k)$ of class C_i on each frame feature vector X_k, $P(C_i|X_{k_\tau}, ..., X_{k_\tau+L_\tau-1})$ (a segment-level probability) is computed for each texture segment obtained in the clustering phase. For this, the sum of all frame-level log probabilities is used. The probability on the τ^{th} texture segment of length L_τ is given by:

$P(C_i|X_{k_\tau}, ..., X_{k_\tau+L_\tau-1}) = \sum_{k=k_\tau}^{k_\tau+L_\tau-1} \log p(C_i|X_k)$. Then, the label of a texture segment is given by the maximum probability criterion.

3.3.5 Active learning for segment selection

Relevance feedback has been widely used in multimedia Information Retrieval. The reader can refer to [12] for an overview. In the context of this work, our approach consists in gradually adding new segments validated by the user in the learning process. As a consequence, the labels predicted for the other segments may evolve at each iteration of the algorithm. The process begins with a limited number of segments for training the classifier and the training segment dataset grows step by step as user-validated segments are injected. The goal of this approach is to obtain the correct labeling of samples in a reasonable number of iterations. Active learning theory proposes sampling strategies which are used to select the segments to be user-validated first. The two *interaction loop* approaches use different sampling strategies. The *Multi-pass* approach uses the *most ambiguous* strategy: in the SVM classifier, most ambiguous samples are the closest to the hyperplane in the feature space. This strategy is adapted to binary classification problems and was shown to give the best results in a previous study [10]. The *one-pass* approach uses the *best versus second best* strategy which has been successfully used in image classification [13]. This strategy uses the difference between the probabilities of the two classes having the highest estimated probability value which provides an estimation of the confusion about class membership.

For each frame-level probability, we compute a score $s(k)$ in accordance with the sampling strategy used. Given this score, for each frame of audio, we obtain a score for each texture segment by temporal integration, where the segment score is the mean of the underlying frame scores: $S_\tau = 1/L_\tau \sum_{k=k_\tau}^{k_\tau+L_\tau-1} s(k)$ for the τ^{th} texture segment. The temporal integration allows us to obtain a unique sampling strategy score for each segment and to rank them. Therefore, the segment which maximizes the score is selected by the system and a feedback request is sent to the user.

4. EVALUATION

User-based experiments are very time consuming and require the creation of ground-truth annotation of numerous music pieces, which often turns out to be even more tricky, especially as far as electro-acoustic music is concerned. Indeed, there exists only a few annotations in this case which mix the description of sound objects with the annotators' subjective interpretation of the pieces. As a result, to validate our method with a descent number of files and easily compare the different parameters settings, we opted for a user simulation with synthetic music pieces generation. Nevertheless, much care has been taken in order to make this procedure completely realistic as will be further explained hereafter.

4.1 Synthetic pieces generation

The synthetic pieces generation process is similar to our previous work. The reader can refer to [10] for a complete description. 24 homogeneous sounds (hence 24 classes) of different lengths (from a second to a minute) were selected by composers of the *Groupe de Recherches Musicales*[3] (INA-GRM) for the generation process. 100 pieces of 2 minutes containing 5 different sound classes for each were generated. 5 versions of each piece were obtained by varying the polyphonic degree from 1 to 5 (*i.e.* 500 synthetic sound files). Consequently, the n^{th} variation of a given piece will have a maximum number of n sounds playing simultane-

[3] `http://www.inagrm.com/`

ously. In the generation process, 5 distinct sounds are selected randomly for a given piece and different instances of each sounds are concatenated/juxtaposed accordingly with the polyphonic degree of the piece.

4.2 User simulation

In this work, we exploit a *user feedback simulator* to facilitate the evaluation. It is used both in the initialisation phase and in the interaction loop. For the initialisation, the slider position controls the overall segmentation level of the piece and the user has to choose the position which best matches his/her viewpoint assuming that the user will tend to try to maximise the segmentation F-measure score. Hence, the latter is computed for all the slider positions, *i.e.* all possible levels of the hierarchical clustering used for segmentation. The F-measure is computed with a temporal precision window of 0.5 s over the segmentation's boundaries. The slider position which maximizes the F-measure is used as the initial segmentation level before entering the *interaction loop*. For the selection of class initialisation segments $\mathcal{S}_i$, the segments in which the target sound class C_i is the loudest were selected. For each texture segment τ we compute an energy ratio: $R_{C_i,\tau} = E_{C_i,\tau}/\sum_{l\neq i} E_{C_l,\tau}$ where $E_{C_i,\tau}$ is the *root mean square* energy of the τ^{th} texture segment for the class C_i. $E_{C_i,\tau} = \sqrt{1/L_\tau \sum_{k=k_\tau}^{k_\tau+L_\tau-1} x_i^2(k)}$ with x_i the signal of the class C_i. For a given sound class C_i, the texture segment which maximizes the ratio R_{C_i} is selected as the initialisation segment for C_i. In the *interaction loop*, the successive interaction steps of the user with the system, exposed in Section 3.1 were simulated for the 500 sound files of the whole corpus. In this work, for active learning segment selection, we filter segments shorter than 0.5 s since they could be misjudged by the user when asked for validation, due to human perception limitations. A basic version of the function *undo* is also simulated: if an acceptable level of satisfaction (F-measure $\geq$ 0.85) is reached for a given class C_l, the results must not decrease in the next iterations. Therefore, if the results decrease, we suppose that the user will use the *undo* function and lock the class C_l to retain the previous classifier predictions and re-use them (without further updating) for the next iterations.

4.3 Results

We monitored the behaviour of the F-measure scores for 500 pieces over the iterations of the algorithm with the different interactive approaches. In the different methods, we fix a maximum number of iterations of 30 since good results should be obtained in a reasonable number of interactions.

As it was observed in our previous work [10], the results decrease accordingly with the polyphonic complexity of the pieces. Figure 3 shows the F-measure results across the iterations for a particular class with the MP approach (polyphony=4). It is observed that good results can be obtained after 10 iterations with this reasonable polyphonic degree. Given the nature of this approach, which permits the user to focus on one class for the whole process, the obtained number of iterations must be multiplied by the number of

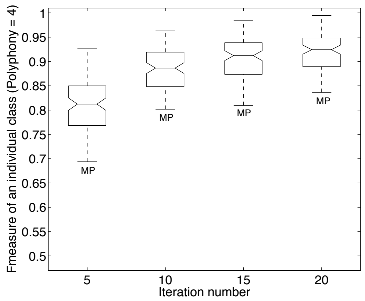

Figure 3. F-measure versus number of iterations for the MP approach (polyphony = 4). The central mark is the median, the edges of the box are the 25th and 75th percentiles and the whiskers extend to the minimum and maximum data points.

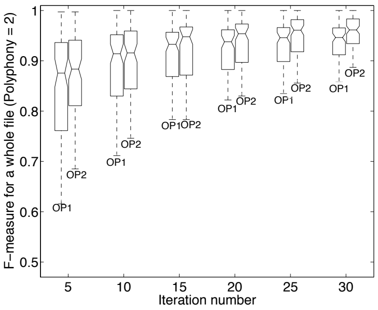

Figure 4. F-measure versus number of iterations for the OP1 and OP2 approaches (polyphony = 2).

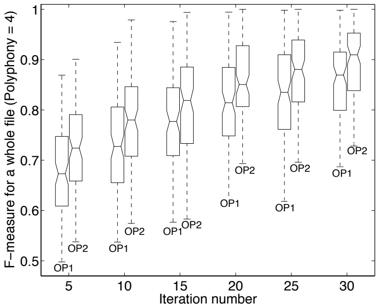

Figure 5. F-measure versus number of iterations for the OP1 and OP2 approaches (polyphony = 4).

classes to adress in the music piece.

Figure 4 compares the *one-pass* approaches (OP1 and OP2) for a polyphonic degree of 2. The Figure 5 compares the same approaches for a polyphonic degree of 4. The results show that the method OP2 which introduces new mixture classes with user feedback gives better results. These approaches are both considering all the classes of interest at the same time and we observe that they can reduce the total number of iterations comparing to the MP approach in which the user must repeat the process to address all the classes. A satisfying median F-measure of 0.85 can be obtained in 20 iterations with OP2 for a whole piece.

5. CONCLUSION

In this paper, we have proposed two different interactive approaches for helping the analysis of electro-acoustic music. In the *multi-pass* approach, the user focuses on one sound class at each time. In the *one-pass* approaches, the user gives a more informative feedback to treat all the classes of the file simultaneously. The results show that the MP approach is more adapted to a small number of classes: if the number of classes to transcribe is important, satisfying results can be obtained in a smaller number of iterations with OP2 (the most effective *one-pass* approach).

Future works will focus on integrating the labeling informations of the initial clustering. To validate the system with real pieces, we will extend the evaluation to real users and work on the design of an appropriate user interface.

6. REFERENCES

[1] `http://www.inagrm.com/sites/default/files/polychromes/problematique/modulePP/index.html`.

[2] C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[3] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] M. Crucianu, M. Ferecatu, and N. Boujemaa. Relevance feedback for image retrieval: a short survey. In *State of the Art in Audiovisual Content-Based Retrieval, Information Universal Access and Interaction including Datamodels and Languages (DELOS2 Report)*, 2004.

[5] R. Duda, P. Hart, and D. E. Stork. Pattern classification, 2001. New York: Wiley-Interscience.

[6] S. Essid, G. Richard, and B. David. Musical instrument recognition by pairwise classification strategies. In *IEEE Transactions on Speech, Audio and Language Processing*, volume 14, pages 1401–1412, 2006.

[7] M. R. Every. Discriminating Between Pitched Sources in Music Audio. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2):267–277, 2008.

[8] F. Fuhrmann, M. Haro, and P. Herrera. Scalability, generality and temporal aspects in automatic recognition of predominant musical instruments in polyphonic music. *in Proc. of ISMIR*, 2009.

[9] S. Gulluni, S. Essid, O. Buisson, and G. Richard. Interactive segmentation of electro-acoustic music. *International Workshop on Machine Learning and Music*, 2009.

[10] S. Gulluni, S. Essid, O. Buisson, and G. Richard. Interactive classification of sound objects for polyphonic electro-acoustic music annotation. *in Proc. AES 42nd Conference on Semantic Audio*, 2011.

[11] K. Hoashi, K. Matsumoto, and N. Inoue. Personalization of user profiles for content-based music retrieval based on relevance feedback. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 110–119, 2003.

[12] X. Jin, J. French, and J. Michel. Toward Consistent Evaluation of Relevance Feedback Approaches in Multimedia Retrieval. *Adaptive Multimedia Retrieval: User, Context, and Feedback*, pages 191–206, 2006.

[13] A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. *IEEE Conference on Computer Vision and Pattern Recognition (2009)*, pages 2372–2379, 2009.

[14] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. Instrument Identification in Polyphonic Music: Feature Weighting to Minimize Influence of Sound Overlaps. *EURASIP J. Appl. Signal Process.*, 2007:155–155, January 2007.

[15] M. Mandel, G. Poliner, and D. Ellis. Support vector machine active learning for music retrieval. In *ACM Multimedia Systems Journal*, 2006.

[16] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard. Yaafe, an easy to use and efficient audio feature extraction software. *in Proc. of ISMIR*, 2010.

[17] G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Tech. rep., IRCAM, 2004.

[18] D. Teruggi. Technology and Musique Concrete: The Technical Developments of the Groupe de Recherches Musicales and Their Implication in Musical Composition. *Organised Sound*, 12(3):213–231, 2007.