Interactive Classification of Sound Objects for Polyphonic Electro-Acoustic Music Annotation

Sebastien Gulluni[1 2], Slim Essid[2], Olivier Buisson[1], and Gaël 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*

Correspondence should be addressed to Sebastien Gulluni (gulluni@telecom-paristech.fr)

ABSTRACT

In this paper, we present an interactive approach for the classification of sound objects in electro-acoustic music. For this purpose, we use relevance feedback combined with 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. Three active learning criteria are compared in the evaluation of a system classifying polyphonic pieces (with a varying degree of polyphony). The results show that the interactive approach achieves satisfying performance in a reasonable number of iterations.

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 [1]. 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 adaptive classification systems capable of analyzing and structuring electro-acoustic music in a semi-automatic fashion using user relevance feedback [2], which to the best of our knowledge remains an original approach.

Previous works on polyphonic timbre classification have focused on "standard" instruments and percussion used in the majority of conventional music [3, 4, 5, 6]. In these approaches, as individual timbres are known, it is possible to build supervised systems by using large audio databases which involve the corresponding standard instruments. In the electro-acoustic case, composers exploit various sound sources and we do not have a-priori knowledge about these sources which are most of the time polyphonic and heterogeneous. The reader can refer to [7] (a multimedia presentation on the works of important composers of the genre) for examples of electro-acoustic compositions.

Relevance feedback has been widely used in content-based image retrieval tasks (see [8] for an overview). Many works use classifiers to learn high level semantic concept from low level features for the image retrieval task. The user gives feedback to the system by qualifying the images returned as "relevant" or "irrelevant". The present work uses this approach to identify complex sounds.

Opposing to image retrieval, relevance feedback and active learning have only been used in a few studies [9, 10] in the field of audio retrieval. In [9] the study is focused on the task of pop music retrieval based on user preferences and [10] is about mood and style classification.

In this work, we propose an interactive approach with relevance feedback adapted to the analysis of electro-acoustic compositions which are traditionally organized in sound objects. Here, we define "sound object" as any sound event perceived as a whole [1]. Most of the time a music piece does not expose separated sound objects, *i.e.* simultaneous sounds are masking each others due to polyphony. As in [6], we use sound mixtures which contain the target object as positive samples, and sound mixtures which do not contain the target object as negative samples for learning. The interactive classification of sound objects uses relevance feedback and active learning segment selection (see Figure 1). From a user's point of view, the search for a target sound object begins with

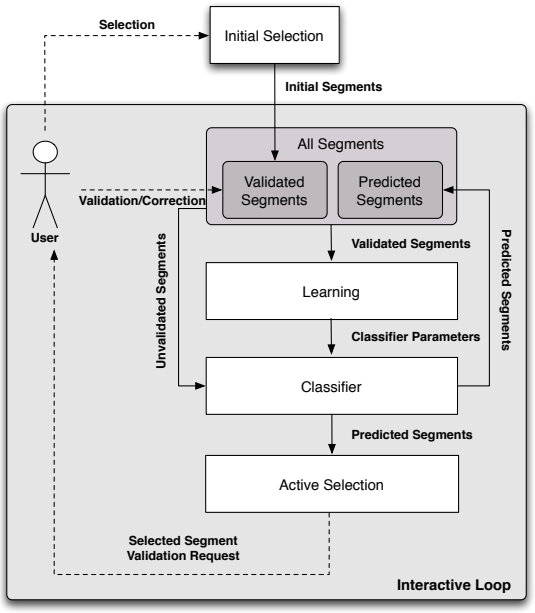

Fig. 1: Overview of the interactive system

the selection of 2 segments: the first contains the target sound (positive samples) and the second does not (negative samples). Then, the system enters in an interaction loop and suggests, at each iteration, segments to be annotated by the user 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 annotation. We compare different active learning criteria and show that we can obtain satisfying results in a reasonable number of iterations for different degrees of polyphonic complexity.

The paper is organized as follows: Section 2 describes the interactive classification approach including the user scenario and active learning segment selection. Section 3 is dedicated to the evaluation of the method and the last section suggests some conclusions.

2. INTERACTIVE CLASSIFICATION SYSTEM

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

2.1. System architecture

Figure 2 is a representation of a polyphonic piece which involves potential sounds masking: the distinct sound layers are arranged in parallel timelines (one for each sound). The goal of the annotation is to mark the presence of the different sounds in the whole piece. The classification operates on *segments*, *i.e.* temporal fragments of homogeneous timbre (as shown with vertical orange lines in Figure 2). In this work, the segment boundaries are supposed to be known to allow us to focus on the classification problem. In future work, the segmentation could be obtained interactively as in [11].

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

1. The user selects a segment S_{i+}. This segment should be the most characteristic instance of a class C_i (see Figure 2). Hence, the chosen segment should be the one in which the target sound class is perceived by the user to be the least masked by other signals.

2. The user selects a segment S_{i-} which does not contain the target sound class C_i.

3. The system learns from the validated segments and enters in the classification process to automatically annotate the remaining parts of the signal.

4. In order to improve the previous classification, the system selects a segment, based on one of the active learning strategies described in Section 2.4, and asks the user to validate or correct its current label.

5. If the user is not satisfied with the current overall annotation, the system goes back to step 3. Otherwise, the system goes back to step 1 to annotate the next class C_{i+1} until all the target classes are annotated.

2.2. Feature extraction

The features are calculated on 20ms windows with 50% overlap. The sampling rate of the sound files is 44.1kHz. To cope with the complexity of the sounds to be classified, a large set of audio features is considered and feature selection is used and updated at each relevance feedback iteration. The reader can refer to [12, 13] for a complete description of the features. All the features used and the corresponding number of attributes are listed in Table 1. Feature extraction was performed using the YAAFE software [14]. A total of 217 attributes were extracted from 25 descriptors.

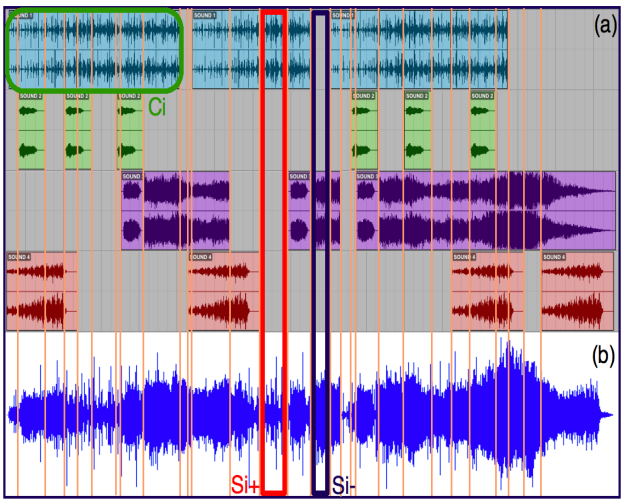

Fig. 2: Time-line representation of a polyphonic piece with C_i (target class), S_{i+} (initial positive segment) and S_{i-} (initial negative segment). Though the distinct sound layers are here displayed in parallel time lines (a), in real situations the user can actually only see the final mix made by the composer that appears as a single track (b). The initial user selection and subsequent validations are done by listening.

Feature Name	Attributes
Auto Correlation	49
Root Mean Square Energy	1
Amplitude Envelope	6
Envelope Shape Statistics	4
Linear Predictive Coding	2
Line Spectral Frequencies	10
Loudness	24
MFCC (and derivates order 1,2)	39
Octave Band Signal Intensities	10
Octave Band Signal Intensities Ratio	9
Perceptual Sharpness	1
Perceptual Spread	1
Spectral Crest Factor Per Band	23
Spectral Decrease	1
Spectral Flatness	1
Spectral Flatness Per Band	23
Spectral Flux	1
Spectral Rolloff	1
Spectral Shape Statistics	4
Spectral Slope	1
Spectral Variation	1
Temporal Shape Statistics	4
Zero Crossing Rate	1
Total number of attributes	217

Table 1: List of the extracted features

2.3. Classification

In this system, the classification task consists in detecting the presence of a given class for all the segments of the music piece. Classifications are performed independently for all the classes of the music piece. A characteristic of this system is that it uses polyphonic segments as in [6] in a "one vs all" fashion for the learning phase. In other words, 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 sound mixtures which contain other sounds.

The classification phase begins with a feature selection based on the Fisher discriminant [15]. 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 current 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.

After the selection process, the feature vectors of the current validated segments (Figure 1) are used to train a Support Vector Machine (SVM) classifier [16]. In the same way as we do in the feature selection phase, a separate database was used to find the optimal parameter settings for the SVM. We use probabilised output SVMs[1] to obtain a frame-level posterior probability $p(C_i|X_j)$ of the class C_i on each frame feature vector X_j [18]. Then, a segment-level probability $P(C_i|X_{j_\tau},...,X_{j_\tau+L_\tau-1})$ is computed for each segment. 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: $P(C_i|X_{j_\tau},...,X_{j_\tau+L_\tau-1}) = \sum_{j=j_\tau}^{j_\tau+L_\tau-1} \log p(C_i|X_j)$. Finally, the label of a texture segment is given by the maximum probability criterion.

[1] we use the libSVM implementation [17].

2.4. Active learning for segment selection

Relevance feedback has been widely used in multimedia Information Retrieval and the reader can refer to [2] 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 choice of an adapted sampling strategy criterion is crucial to obtain correct labeling quickly (See section 3.3).

In this work, we compared the following sampling strategies which were used successfully with SVM classifiers in other relevance feedback studies [citer papier RF].

- **Most Positive:** this strategy chooses in priority the samples which have the highest probability to contain the target class;
- **Most Negative:** in contrast to the previous strategy, this one selects first the samples which have the lowest probability to contain the target class;
- **Most Ambiguous:** this strategy chooses first the uncertain samples (probability near 0.5). In the SVM classifier point of view, most ambiguous samples are the closest to the hyperplane in the feature space.

For each probability given by the classifier, we compute a score in accordance with the used sampling strategy (see Figure 3). Given this score for each frame of audio, we obtain a score for each segment by temporal integration, where the segment score is the mean of the underlying frame scores. The temporal integration allows us to obtain a unique sampling strategy score for each segment and to rank them. The segment which maximizes the chosen sampling strategy is selected and the segment validation request is sent to the user.

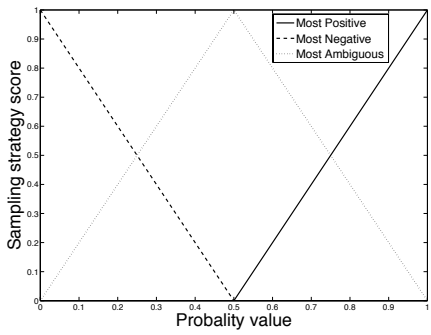

Fig. 3: Sampling strategies score calculation

3. 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.

3.1. Synthetic pieces generation

The goal of the synthetic pieces generation is to create a polyphony of complex sounds. As a consequence, the sounds used for the generation are initially complex and have a temporal evolution.

Three composers of electro-acoustic music from the *Groupe de Recherches Musicales* (INA-GRM) were involved to provide sounds. These sounds, for the most of them, come from personal sound recordings and were chosen independently by the composers without any compositional intents. However, the only constraint was to opt for acoustically homogeneous sounds in the sense that the main timbral characteristics of each sound selected had to remain stable over its duration in order to consider it as an individual class instance. The three composers selected a total of 24 sounds (hence 24 classes) which were used for the generation of the synthetic pieces. The most important characteristics of the selected sounds are the length and complexity:

- Lengths vary from one second to a minute;
- Some sounds are built from an aggregate of smaller elementary sound events;
- Some sounds are composed from the superposition of many elementary sound events.

In order to make a more accurate study of the polyphonic evolution, 5 versions of the same basic piece were generated with a different degree of polyphony. The first version of each piece is monophonic and the fifth has a polyphonic degree of 5 sounds. As a result, for the i^{th} version of the piece, we have a maximum of i sounds playing at the same time. A total of 100 pieces were generated with 5 polyphonic versions for each. All pieces are 2-minute long. The reader can refer to the website of this paper[2] for examples of individual sounds and synthetic pieces. The generation process to make sequences of sound events was to take 5 arbitrary sounds from the 24 available and then to extract randomly segments in the selected sounds to make different instances of the same class. By alternating sound events and silence, we obtained sound layers that we juxtaposed accordingly with the polyphony of the generated piece. In these synthetic files, the different instances of the sound classes are considered as the target sound objects.

3.2. User simulation

In this work, we focus on the classification of segments longer than 0.5 s since shorter segments could be misjudged by the user when asked for validation, due to human perception limitations. The successive interaction steps of the user with the system exposed in Section 2.1 were simulated for the 500 sound files of the whole corpus. For the initial selection of the segment $Si+$, the segment with the smallest masking degree is selected: the simulation algorithm first filters the segments which do not contain the sound class C_i and the segment with the smallest polyphonic degree, *i.e.* the one involving the smallest number of sound classes, is selected.

3.3. Results

For the validation of the interactive method, we monitored the behaviour of the F-measure scores for 500 pieces over the sequence of relevance feedback iterations: the user simulation algorithm loops until the maximum score is reached. The goal is to minimize the number of iterations. We compute the F-measure score F_i for the class C_i using the segment-level predictions: $F_i = \frac{2R_iP_i}{R_i+P_i}$ where R_i is the recall and P_i is the precision for the i^{th} class.

[2] http://www.tsi.enst.fr/~gulluni/aes2k11/

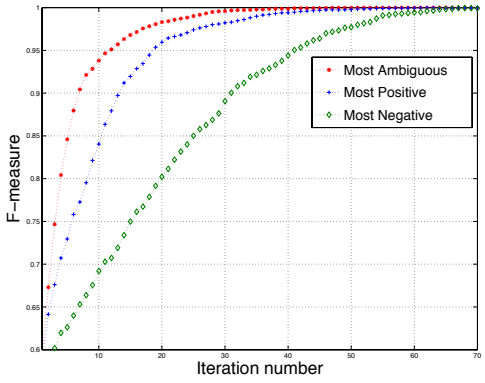

Fig. 4: Average F-measure versus number of iterations for the three active learning criteria (polyphony = 3).

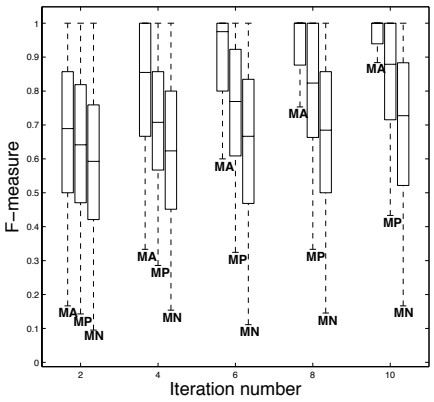

Fig. 5: Detailed performance for the first iterations by sampling strategy. The central mark is the median, the edges of the box are the 25th and 75th percentiles and the whiskers extend to the minimum and maximum data points. MA = Most Ambiguous, MP = Most Positive and MN is for Most Negative (polyphony = 3).

Figure 4 is a global view of the average F-measure evolution for all iterations of the experiment. Figure 5 shows the detailed performances of the first iterations for the three sampling strategies (Most Ambiguous, Most Posi-

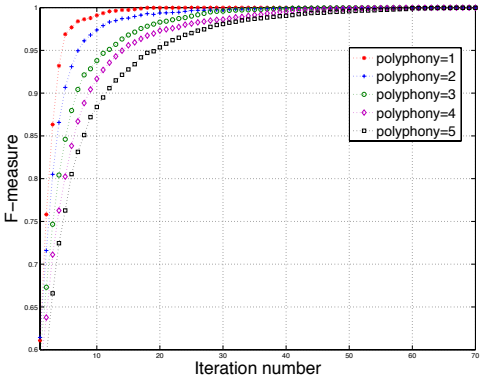

Fig. 6: Average F-measure versus number of iterations for five polyphonic degrees with the MA sampling strategy.

tive, Most Negative). The two figures are the results for the intermediate polyphonic degree (polyphonic degree = 3) and show that the Most Ambiguous strategy performs significantly better than the Most Positive and the Most Negative strategies. An average number of 12 iterations to get a F-measure of 0.95 is shown in Figure 4 for the Most Ambiguous strategy. The Most Positive strategy takes an average number of 19 iterations to get the same score and the Most Negative is the worst with 41 iterations.

Figure 6 shows the average F-measures for the five polyphonic categories in the Most Ambiguous case. As expected, the performance decreases significantly when the polyphony becomes more complex. The monophonic case takes 4 iterations on average to get a performance score of 0.95 and the same score is obtained in 20 iterations for the most polyphonic cases.

4. CONCLUSION

In this study we have proposed an interactive classification system adapted to the annotation of electro-acoustic music. The lack of a-priori knowledge of the sound sources makes the classic techniques for polyphonic music classification difficult to apply [3, 4, 5]. Three sampling strategies have been compared and the Most Ambiguous criterium has been shown to perform best. Sound classes can be successfully annotated in an average of 4 iterations for the monophonic case, 12 iterations for the intermediate case (polyphonic degree = 3) and 20 iterations for the most polyphonic case.

Future work will focus on limiting the number of interactions with the user. More than one segment could be selected by the system and the user could give more feedback before a new learning phase is launched. In parallel, to extend the evaluation to real users and real music pieces, dedicated effort will be devoted to the design of an appropriate user interface.

5. REFERENCES

[1] D. Teruggi, "Technology and Musique Concrete: The Technical Developments of the Groupe de Recherches Musicales and Their Implication in Musical Composition.," *Organised Sound*, vol. 12, no. 3, pp. 213–231, 2007.

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

[3] M. R. Every, "Discriminating Between Pitched Sources in Music Audio," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 267–277, 2008.

[4] 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.*, vol. 2007, pp. 155–155, January 2007.

[5] 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.

[6] D. Little and B. Pardo, "Learning musical instruments from mixtures of audio with weak labels," *in Proc. of ISMIR*, 2008.

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

[8] 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.

[9] 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*, 2003, pp. 110–119.

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

[11] S. Gulluni, S. Essid, O. Buisson, and G. Richard, "Interactive Segmentation of Electro-Acoustic Music," *2nd International Workshop on Machine Learning and Music*, 2009.

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

[13] S. Essid, G. Richard, and B. David, "Musical instrument recognition by pairwise classification strategies," in *IEEE Transactions on Speech, Audio and Language Processing*, 2006, vol. 14, pp. 1401–1412.

[14] 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.

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

[16] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[17] 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`.

[18] T.F. Wu, C.-J. Lin, and R.C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, 2004.