COMBINED SUPERVISED AND UNSUPERVISED APPROACHES FOR AUTOMATIC SEGMENTATION OF RADIOPHONIC AUDIO STREAMS

Gaël Richard, Mathieu Ramona and Slim Essid

GET-ENST, 37 rue Dareau, 75014 Paris, France

ABSTRACT

Speech/music discrimination is one of the most studied topics in the domain of audio data segmentation. In this paper, we propose and evaluate a novel method that includes feature selection and a combined supervised and unsupervised strategy for audio streams segmentation. A number of alternatives solutions for each component are assessed and the optimized system is compared to the approaches proposed in the framework of the ESTER campaign.

Index Terms- Speech/Music discrimination, audio segmentation, novelty detection.

1. INTRODUCTION

The segmentation of audio data is of interest for a broad class of applications including audio scene analysis, surveillance applications or indexing of broadcast audio or audiovisual streams. Speech/music discrimination is one of the most studied topics in this domain, probably because it is essential for broadcast news automatic transcription. In fact, numerous approaches have been proposed in the context of automatic monitoring of FM radio channels [1], in the context of broadcast news transcription [2] or coding applications [3]. Traditional approaches are based on a Front-end module that extracts features from the input signal which are then processed by a previously trained classifier. A wide variety of features were proposed ranging from the widely spread zero crossing rates (ZCR) [1] and Mel-Frequency Cepstral Coefficients (MFCC) [4] to more specific features [5]. Several classification strategies were also proposed in the past including frame-based supervised approaches (based on Gaussian Mixture Models or Hidden Markov Models) and combined approaches exploiting a time-domain segmentation [5].

In this paper, we propose and evaluate a novel method that exploits feature selection, supervised and unsupervised classification approaches. A number of alternative solutions for each component are assessed and the optimized system is compared to the methods proposed in the framework of the ESTER campaign [6, 7].

The paper is organized as follows. First, the overall system architecture is briefly described. Then, in section 3, the feature extraction and selection are discussed. We respectively outline the supervised and unsupervised approaches in section 4 and 5. Section 6 is dedicated to the experiments and results obtained, and some conclusions are suggested in the last section.

2. SYSTEM ARCHITECTURE

The overall architecture of the radiophonic audio stream segmenter is depicted on figure 1.

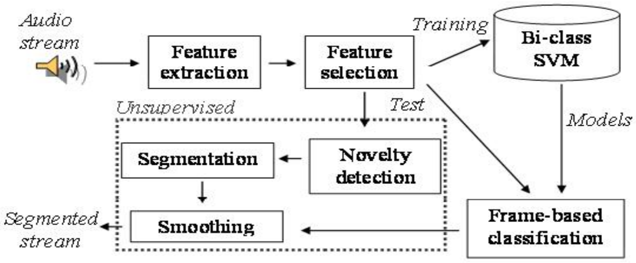

Fig. 1. System architecture

It associates a supervised approach with an unsupervised smoothing procedure. First, an extensive set of features are extracted (*feature extraction*). Second, the most efficient features are retained using a simple feature selection algorithm (*feature selection*) and are then used to train the three binary Support Vector Machines (SVM) classifiers (one per bi-class classification). During the test phase, the feature previously selected are computed and a decision (Speech vs Music vs Mixed) can be taken for each time-window. The unsupervised approach aims at obtaining longer segments for which a single label will be attached to (namely the most frequent label given by the supervised approach on this audio segment).

3. FEATURE EXTRACTION AND SELECTION

The initial set considered encompasses over 500 feature values that can be grouped into: *Temporal features* (ZCR, temporal statistical moments, modulation coefficients,...), *Spectral features* (spectral statistical moments, spectral slope, spectral flux,...), *Cepstral features* (MFCC, Constant Q transform cepstral coefficients) and *Perceptual features* (Relative loudness, perceptive sharpness,...). Note that for most of these descriptors, first and second derivatives are also considered. A detailed presentation of these features can be found in [8]. At this stage, different types of features exist: those computed on short time windows of 20ms (called short-term features) and those computed on longer windows of 2s (called long-term features). Feature integration is achieved by substituting short-term features by their means and variances computed on long windows. Before further processing, each feature is centered and normalized by its respective variance computed on the training

This research was supported by the European Commission under contract *FP6-027026-K-SPACE* and by the National Project *Cap Digital - Infom@gic*.

database.

Two rather simple feature selection algorithms are used in this work. The first approach is based on the Fisher Discriminant ratio which is defined as the ratio of the Between-class inertia to the "average radius" of the scatter of all classes. The features corresponding to the highest ratios are selected (the number of features selected is predefined). However, using such a criterion may result in redundant feature subsets. Hence, as described in [9] an alternative algorithm, called IRMFSP, was proposed to take into account the non-redundancy constraint by introducing an orthogonalization step at each feature selection iteration.

4. SUPERVISED CLASSIFICATION

The supervised classification approach used in this study is known as Support Vector Machines (SVM). SVM non-linearly map (using a Kernel function) their N-dimensional input space into a higher dimensional feature space where the two classes are linearly separable with an optimal margin. SVM solutions can be expressed using dot products in the high dimension space which can be computed through kernel functions. For this study, a radial basis kernel was chosen. Although such classifiers are designed for binary classification and regression estimation tasks, they can also be adapted to perform multi-class classification In this work, three different classes are considered, namely *Speech*, *Music* and *Mixed segments* and a one vs one scheme (or class pairwise strategy) is adopted.

5. UNSUPERVISED SMOOTHING

The unsupervised step aims at segmenting the incoming audio stream in homogeneous continuous segments. Since it is then assumed that such segments are assigned to a single label, this unsupervised segmentation can be used to smooth the frame-based results obtained by the supervised approach. It is based on a specific processing of a novelty detection function.

5.1. Novelty detection

We evaluate, in this work, several novelty detection approaches that use the the same framework. A sliding window $W(k_0)$ of length $2L+1$ centered at frame k_0 is observed. k_0 is considered as a good candidate for being a segment boundary if the content of the "future" data set $S_2(k_0) = \{X_n(k), k \in [k_0, k_0+L]\}$ is novel relatively to the content of the "past" data set $S_1(k_0) = \{X_n(k), k \in [k_0-L, k_0]\}$. To simplify notations, the past and future windows, for a given value of k_0 will be simply referred to as S_1 and S_2, the underlying densities as P_1 and P_2, the entire window as W. The five novelty detection approaches considered are briefly described below (more details can be found in [10]):

Bayesian Information Criterion Being a classical model or order selection criterion, the Bayesian Information Criterion (BIC), has been widely used in speech/music or speakers segmentation problems. It will be considered here as a baseline algorithm.

One-class SVM One-class SVM aim at identifying a region of the feature space in which most of the data points lie. This is obtained by finding the hyperplane that separates the data from the origin with maximum margin. Two different novelty detection approaches based on single-class SVM are envisaged in this work. The first method is based on a likelihood ratio test :

$$R = \frac{\prod_{x\in S_1} P_1(x) \prod_{x\in S_2} P_2(x)}{\prod_{x\in W} P_1(x)} = \frac{\prod_{x\in S_2} P_2(x)}{\prod_{x\in S_2} P_1(x)} > t$$

The estimates of P_1 and P_2 can be deduced from the SVM algorithm solution:

$$P_i(x) = \exp(\sum_k \alpha_k^i K(x, X_k^i) - \rho_i)$$

where ρ_i is a threshold, α_k^i are Lagrange multipliers, $(X_k^i)_k = S_i$ the vectors of the training set, and K is the kernel used.

The second method is the so-called Kernel Change Detection (KCD) approach [11]. It is based on a dissimilarity measure that can be seen as a ratio of "inter-class scatter" to "intra-class scatter" in the transformed space induced by the kernel.

Probabilistic distances Another way of detecting segment boundaries is by using a relevant distance between the data points in S_1 and S_2. We expect these boundaries to be characterized by a higher distance. For the sake of robustness we consider probabilistic distances between the estimates of the distributions P_1 and P_2. Many such distances can be considered among which we chose the Bhattacharryya and Kullback-Leibler divergence (mainly due to the resulting simplification in the following computations). However, to avoid the assumption of Gaussian distribution of the original class observations, the data is mapped from the original space to a Reproducing Kernel Hilbert Space (RKHS) and distances are expressed in terms of kernel evaluations.

5.2. Segmentation

The novelty detection functions $d(n)$, output of the BIC, 1-class SVM or probabilistic distance algorithms usually exhibit large dynamics and therefore need further processing to facilitate the segmentation. The following steps are then performed:

Detrending using a non-linear median filter:

$$d_d(n) = d(n) - \text{median}[d(n-W_a), \ldots, d(n), \ldots, d(n+W_a)]$$

where W_a is the window size,

Normalization to compensate for local variations of peaks amplitude using a standard deviation filter:

$$d_c(n) = \frac{d_d(n)}{\text{std}[d_d(n-W_a), \ldots, d_d(n), \ldots, d_d(n+W_a)]}$$

Peak detection under the constraints that the peaks should be separated by a minimal number of frames W_b; and should be above a given threshold τ. A section boundary is detected whenever a peak is reached. Typical values for W_a and W_b respectively are 360 frames (45 seconds) and 40 frames (5 seconds).

6. EXPERIMENTS AND RESULTS

The experiments, for a large part, are based on data collected in the ESTER campaign that aimed at evaluating French radiophonic audio streams transcription systems (see [6, 7]).

6.1. Databases and evaluation protocol

6.1.1. `SEQ_ESTER` *and* `SEQ_ESTER+` *databases*

The ESTER campaign was conducted in two phases. A total of one hundred hours was recorded and annotated. In this work, the entire database recorded for phase I (40 hours) and the entire training database of phase II (i.e 50 hours) constitute our corpus (a total of 90 hours) from which a development corpus, `ESTER_DEV` (10 hours), is extracted. The test corpus of ESTER phase II is left aside for the final single run test evaluation. The primary goal of this database was the development of speech transcription systems, and therefore its content has a strong bias towards speech (77h30 of speech, 11h45 hours of speech on musical background and only 40 minutes of music alone). Since this disproportion has an impact on the performance, it motivates us to build alternative training corpora. The first training corpus, SEQ_ESTER, is a subset of the initial training corpus but where an equal amount of data for each class (i.e. 40 minutes) is randomly selected. The second training corpus, SEQ_ESTER+, was built by adding 40 minutes of music coming from the RWC database [12] to the previous SEQ_ESTER database to increase the variety of musical signals in the training database.

6.1.2. *Experimental protocol*

The results of the experiments below are given on the development database ESTER_DEV except for the final test results. Note that for this final test our system is evaluated in a single run on the test database similarly to the SES task conducted in the ESTER campaign (that is on the tasks "speech/non speech" and "music/non music" segmentation). The confusion matrix on the three different classes *Speech*, *Music* and *Mixed* are also given.

The performances are evaluated by means of three measures, namely the *F-measure*, the *Recall* (R) and *precision* (P) rates which are computed on the segment boundaries (called events). These events are given in seconds with a tolerance of 0.25s. Let c_i be such an event then : $t(c_i; c_i)$ equals 1 if the event c_i is correctly detected (and 0 otherwise); $t(\bar{c}_i; c_i)$ equals 1 if the event c_i is missed (and 0 otherwise) and $t(c_i; \bar{c}_i)$ equals 1 in case of false alarm (and 0 otherwise).

Recall (R), *precision* (P) and *F-measure* (F) are then given by:

$$R = \frac{\sum_i t(c_i; c_i)}{\sum_i t(c_i; c_i) + t(\bar{c}_i; c_i)} \quad ; \quad P = \frac{\sum_i t(c_i; c_i)}{\sum_i t(c_i; c_i) + t(c_i; \bar{c}_i)}$$

$$F = \frac{2RP}{R + P}$$

6.2. Experiments

6.2.1. *Feature vector dimension*

In this section, we discuss the results obtained by the two feature selection approaches studied. The F-measures obtained with our supervised binary-SVM classifiers (see section 2) as a function of the target feature vector dimensions are given for both algorithms in Table 1. Note that the SVM classifiers are retrained for each new feature combination. On the overall, the IRMFSP algorithm slightly outperforms the more straightforward algorithm based on the Fisher Discriminant ratio, and for all feature vector dimensions between 30 and 70. The best results are obtained with the 70 features selected by the IRMFSP algorithm. All subsequent results provided in this paper are therefore given with this 70 selected features.

Algorithm	**30**	**40**	**50**	**60**	**70**
Fisher	89.4	89.8	90.7	90.7	90.3
IRMFSP	91.4	91.5	91.4	91.3	***91.7***

Table 1. F-measure as a function of the number of selected features for both algorithms on the database SEQ_ESTER

6.2.2. *Impact of the unsupervised smoothing*

In this section, we evaluate the impact of the unsupervised smoothing step on the overall segmentation performances. The five novelty detection function briefly described in section 5 are compared. The global F-measures obtained on the development corpus ESTER_DEV for all combined approaches are compared to the supervised approach alone (see figure 2). Using an unsupervised smoothing approach brings a clear gain in performance for all novelty detection methods. It can be also observed that BIC is the least efficient smoothing approach which is certainly the consequence of the Gaussianity assumption made by this algorithm. However, it still represents a satisfactory trade-off between efficiency and complexity.

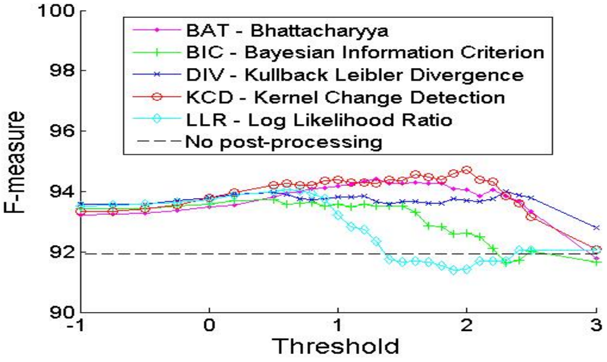

Fig. 2. F-measure for the five unsupervised smoothing approaches compared to the supervised only method (dashed line).

6.2.3. *Training on SEQ_ESTER+*

To evaluate the potential gain achieved by using a richer training corpus in terms of musical content, the same experimental protocol is followed (training on either SEQ_ESTER+ or SEQ_ESTER, testing on the development corpus ESTER_DEV) using the best system configuration (i.e combined supervised approach with unsupervised smoothing based on KCD). Although the system trained on SEQ_ESTER+ obtained better overall results, the difference is not very significant (see Table 3). Complementary experiments using more music material for the training phase have confirmed this observation. This may be explained by the fact that the most frequent music event of the development corpus are jingles which are often very specific musical signals. In fact, it is believed that a significant gain in performances would be obtained by either considering a separate class for the jingles or by developing specific methods dedicated to jingle detection.

6.3. Test results on ESTER segmentation task

Confusion matrix			
Class	Speech	Music	Mixed
Speech	97.6%	0.1%	2.3%
Music	16.1%	47.7%	36.1%
Mixed	25.0%	4.8%	70.2%

Table 2. Results on the ESTER test set

The configuration chosen for this test is the best system obtained that was trained on the SEQ_ESTER database. The results are obtained from a *single run* of this configuration on the

Segments	Mixed			Speech			Music		
Training database	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**
SEQ_ESTER	94.7	94.4	95.7	99.6	99.7	99.4	66.2	74.5	71.1
SEQ_ESTER+	95.3	95.7	95.3	99.4	99.5	99.3	67.1	78.9	69.2

Table 3. F-measure, Recall (R) and Precision (P) comparison for both training databases SEQ_ESTER and SEQ_ESTER+

	All segments			**Speech**			**Music**		
Lab.	F	%fa	%fr	F	%fa	%fr	F	%fa	%fr
ENST	96.5	4.8	4.1	98.9	43.5	2.1	79.3	5.0	8.8
IRIT	94.2	2.1	9.5	98.8	30.1	1.5	52.7	1.2	61.7
IRISA	93.1	1.3	12.1	98.9	9.7	1.9	33.7	1.0	78.5
LIA	92.7	11.7	5.7	99.2	36.6	0.7	54.8	10.9	38.7
LIUM	90.7	1.3	16.2	97.4	8.0	4.9	17.8	1.1	89.6
SIS	83.7	11.5	20.9	93.4	82.2	10.4	12.7	10.4	89.2
UOB	88.2	3.9	18.6	95.1	20.1	8.9	26.2	3.4	82.0
FT R&D	—	—	—	99.1	25.5	1.1	—	—	—
LORIA	—	—	—	97.5	34.2	4.0	—	—	—

Table 4. Comparison with the results of the ESTER campaign for the segmentation task "SES"

test corpus as it was done in the course of the ESTER evaluation campaign. The overall F-measure for the combined supervised/unsupervised approach is 96.5% (98.9% for speech and 79.3% for music) and 95.9% without the unsupervised smoothing step. To provide an alternative analysis of the results, the confusion matrix for the three classes is given in Table 2 and it can be observed that most confusions occur with the mixed segments class.

Finally, the results obtained are compared to those obtained in the ESTER campaign [7]. The best results are obtained with our system although the performance on speech/non speech discrimination is slightly below the best systems (see Table 4). These good results may be explained by a very significant improvement on the music/non music task.

7. CONCLUSION AND FUTURE WORK

A novel and global method for radiophonic audio data segmentation was presented and evaluated using the same protocol and database than those of the ESTER campaign. The proposed approach outperformed the algorithms developed within ESTER but some directions can be suggested for further improvement. In fact, more effort could be dedicated to a finer adaptation of the unsupervised approaches to the specificity of the data. It also appeared that most of the confusions occur with the mixed segment class that, indeed contains segments of both types (i.e. speech and audio). A tempting approach would be to combine the current method with an estimator of the number of active audio sources in the audio stream.

8. REFERENCES

[1] J. Saunders, “Real-time discrimination of broadcast speech/music,” in *Proc. ICASSP ’96*, 1996, pp. 993–996.

[2] G. Williams and D. Ellis, “Speech/music discrimination based on posterior probability features,” in *Proc. Eurospeech ’99, Budapest, Hungary*, Sept 5-9 1999.

[3] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal, “Speech/music discrimination for multimedia applications,” in *Proc. ICASSP ’00, Istanbul, Turkey*, 2000, pp. 2445–2448.

[4] P. C. Woodland, T. Hain, S. E. Johnson, T. R. Niesler, A. Tuerk, E. W. D. Whittaker, and S. J. Young, “The 1997 HTK broadcast news transcription system,” in *Proc. DARPA Broad. News Trans. and Und. Workshop*.

[5] J. Pinquier, J-L. Rouas, and R. André-Obrecht, “Robust speech / music classification in audio documents,” in *Proc. ICSLP’02*, 2002.

[6] G. Gravier, J-F. Bonastre, E. Geoffrois, S. Galliano, K. Mc Tait, and K. Choukri, “The ESTER evaluation campaign of rich transcription of french broadcast news,” in *Proc. Language Evaluation and Resources Conference*, 2004.

[7] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J-F. Bonastre, and G. Gravier, “The ESTER Phase II evaluation campaign for the rich transcription of french broadcast news,” in *Proc. of Interspeech’05*, 2005.

[8] S. Essid, *Classification automatique des signaux audio-fréquences : reconnaissance des instruments de musique (in French)*, Ph.D. thesis, ENST, Dec 2005.

[9] G. Peeters and X. Rodet, “Hierarchical gaussian tree with inertia ratio maximization for the classification of large musical instrument database,” in *Proc. DAFX ’03*, 2003.

[10] O. Gillet and G. Richard, “On the correlation of automatic audio and visual segmentations of music videos,” *IEEE Trans. Circuits and Syst. for Video Techn.*, Accepted 2007.

[11] F. Desobry, M. Davy, and C. Doncarli, “An online kernel change detection algorithm,” *IEEE Trans. Sig. Proc.*, vol. vol. 53(8), pp. 2961–2974, August 2005.

[12] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “RWC music database: Popular, classical and jazz music databases,” in *Proc. ISMIR ’02*, Oct 2002.