EVALUATION OF CLASSIFICATION TECHNIQUES FOR AUDIO INDEXING

José Anibal Arias, Julien Pinquier and Régine André-Obrecht

Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS INP UPS
118, route de Narbonne, 31062 Toulouse cedex 04, FRANCE
{arias, pinquier, obrecht}@irit.fr

ABSTRACT

This work compares two classification techniques used in audio indexing tasks: Gaussian Mixture Models (GMM) and Support Vector Machines (SVM). GMM is a classical technique taken as reference for comparing the performance of SVM in terms of accuracy and execution time. For testing the methodologies, we perform speech and music discrimination in radio programs and environment sounds (laughter and applause) are identified in TV broadcasts. The objective of the study is to establish references and limits to be considered in practical implementations of audio indexing platforms. Tests show complementary properties between methods and data-driven solutions are suggested as conclusion.

1. INTRODUCTION

To process the quantity of audiovisual information available today in a smart and rapid way, it is necessary to have robust tools. Commonly, to index an audio document, key words or melodies are semi-automatically extracted, speakers are detected or topics are identified. Automatic audio indexing is just the first stage of more ambitious intended functions over multimedia documents. We believe that macro-segmentation models can be proposed to manipulate more than the temporal structure of these data types and break through event search or automatic summarizing.

Metadata generation for multimedia data can be achieved using speech recognition [14], image processing [6] and environment sounds detection as indexing technologies. That implies document segmentation in significant units and classification among predefined categories. Through audio track content understanding of a multimedia document we can determine, for example, whether we are dealing with news, commercials or sports programs. In our study case, we identify primary audio sources: speech, music and environment sounds. We consider laughter and applause identifications because they are key sounds in variety broadcasts and they carry out special semantic information. Two classification frameworks are analyzed: Gaussian Mixture Models and Support Vector Machines.

The organization of the paper is as follows: Section 2 describes the two classification methods (GMM and SVM) implemented. Next, we describe pre-processing stage of audio signals and the design of our prototypical indexing platform. In section 4 and 5 we present results of tests experiments and conclusions.

2. CLASSIFICATION TECHNIQUES

2.1. Gaussian Mixture Models

Based on the Fisher algorithm for pattern recognition [3], we consider that each vector belongs to a class, and each class is modeled by a probability distribution function (pdf) of type Gaussian Mixture Model (Figure 1):

- during the training phase, pdf parameters (or models) for each class are estimated,
- during the classification phase, a decision is taken for each test observation by computing the maximum log-likelihood criterion.

One GMM is a combination of k Gaussian laws. In the mixture, each law is weighted (p_k) and specified by two parameters: the mean m_k and the covariance matrix Σ_k.

$$f(x) = \sum_{k=1}^{K} p_k N(x, m_k, \Sigma_k)$$

$$p_k \geq 0 \quad and \quad \sum_{k=1}^{K} p_k = 1$$

Reviewing speech indexing platforms [4] (respectively music [9]), we see how speech segments are selected while the others segments are rejected. The speech/music detection is not studied with the same framework because speech and music have different structure [11]. We use a simpler approach which consists in detecting the two basic

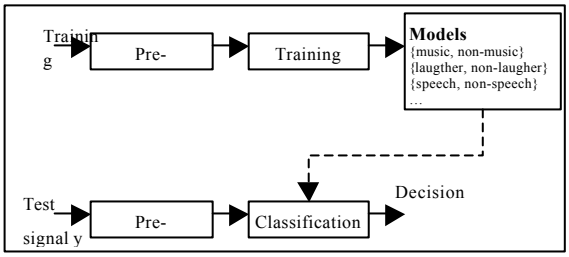

Figure 1. GMM classification system.

components, speech and music, with the same framework. In that purpose, two classification systems are independently defined, a speech detection one and a music detection one. We use the class/non class classification approach. In that way we can fusion the results of both systems and be able to find segments containing speech + music. We define a GMM classification system for each kind of sounds:

- a speech/non-speech classification system,
- a music/non-music classification system,
- an applause/non-applause classification system,
- a laughter/non-laughter classification system.

2.2. Support Vector Machines

This method is a recent alternative for classification [13]. The hypothesis is the existence of a high-dimensional hyperplane (or, in the general case, a non-linear function) that separates two classes. The original vector space is transformed into a Hilbert space by means of a mapping performed with a function called kernel. The hyperplane is calculated using training vectors. The optimal hyperplane is orthogonal to the shortest line connecting the convex hull of the two classes, and intersects it half-way.

To solve a SVM, we have to find the decision function that correctly classifies the data and for every vector x_i satisfies:

$$(1)\quad \begin{cases} w \cdot x_i + b \geq +1 & if \quad y_i = +1 \quad (class\ 1) \\ w \cdot x_i + b \leq -1 & if \quad y_i = -1 \quad (class\ 2) \end{cases}$$

These restrictions are usually expressed as:

$$y_i (w \cdot x_i + b) \geq 1 \quad (2)$$

$1/\|w\|$ corresponds to the distance from the outline of each class to the separating function. The SVM problem is to minimize $\|w\|$ subject to (2), which is a quadratic problem.

A typical approach for solving such a problem is to convert it into its dual expression using the Lagrange multipliers method. It has the form:

$$L_D = \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i,j=1}^{l} a_i a_j y_i y_j K(x_i \cdot x_j) \quad (3)$$

under constraints: $\sum_{i=1}^{l} a_i y_i = 0$

$$0 \leq a_i \leq C \quad i = 1, \ldots, l$$

with x_i =training vectors, y_i =class label of x_i ,

$K(.)$=kernel function, C=tradeoff between misclassification of training data and achieved margin

The SVM decision function:

$$f(u) = sign\left(\sum_{i=1}^{l} a_i y_i K(u, x_i) + b \right)$$

may correspond to a very complex nonlinear decision surface in the original space (Figure 2). We adjust the parameter C to define the decision function that separates learning data.

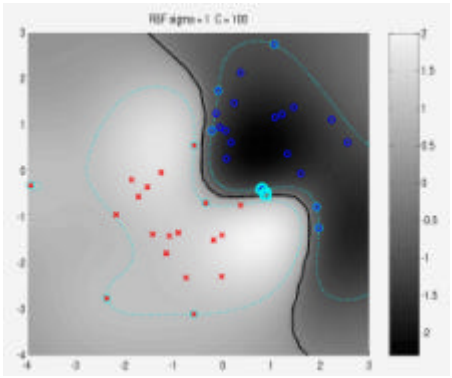

Figure 2. Non-linear SVM classification example.

3. CLASSIFICATION SYSTEMS

3.1. Pre-processing

Audio signals are complex. Two sounds can look very different even if they carry the same information [2]. We perform two different feature extraction processes to keep relevant characteristics from signals. In order to identify the speech contained in the audio track, we make a cepstral analysis of data [5]. The signal is decomposed in 16ms frames and for each frame 18 parameters are used: 8 MFCC plus energy and their associated derivatives. Cepstral features are normalized by cepstral subtractions. For music and environment sounds detection, a spectral analysis is performed on the same frames [7]. 28 filters outputs plus energy are computed for each frame.

3.2. GMM

The GMM training consists of an initialization step followed by an optimization step. The initialization step is performed using the vector quantization algorithm [8]. The

optimization of the parameters is made by the Expectation-Maximization algorithm [12]. After experiments, the number of Gaussian laws in the mixtures has been fixed to 128 for music and 64 for speech and environment sounds. The classification by GMM is made by computing the log-likelihood of tests vectors against the corresponding class and non-class models.

Following the classification, a phase of merging allows concatenating neighboring frames with the same label. A two-steps smoothing function is necessary to delete insignificant segments. The first step deletes segments of size lower than 20 ms (not significant). The second step consists in keeping the important zones (in size) of speech (respectively of music). This smoothing is about 500 ms for the speech/non-speech system, about 2000 ms for the music/non-music system and 1000 ms for the other systems (applause and laughter).

3.2. SVM

The training phase of SVM is the optimization problem described in 2.2. Hence we have to solve the Lagrangian L_D equation for the variables α_i. L_D is a convex function and its solution defines a training model. We call the algorithm SVM due to the fact that only few training vectors have a non-zero α_i, and they are called support vectors. These vectors are situated near the limit of each class. The classification by SVM is discriminative. It computes the decision function for each test vector with class or non-class result. A phase of merging and smoothing is also necessary (cf. 3.1.).

In conclusion, we have a pre-processing stage followed by a training/classification step with two different technologies. Platform can be easily extended to other kind of sounds.

4. EXPERIMENTAL RESULTS

4.1. Corpus

For speech and music discriminations we used 7 hours of radio programs from RFI (Radio France International) database. The program content is varied (news, songs, commercials), in 3 different languages: English, French, and Spanish. For environment sounds, we used a 6-hours corpus from the TV show "Le Grand Echiquier" (French variety). Figures 3 and 4 give examples of delay between manual and automatic detection for applause and laughter. Due to the nature of signals (multiple sources mixed), automatic indexing has slightly difficulties to properly find the beginning and the end of sounds.

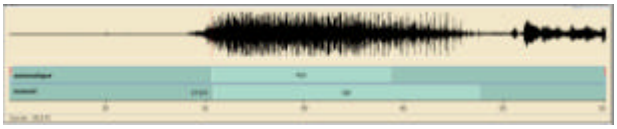

Figure 3. SVM applause/non applause classification. Upper band shows the result of automatic indexing, lower band represents manual labeling.

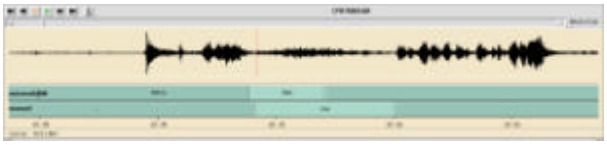

Figure 4. GMM laughter/non laughter classification.

4.2. Speech and music detections

Using the GMM approach, 200000 vectors (~2-hours corpus) are necessary to reach good scores. Maximum average scores are achieved when we take into account the full 6-hours learning corpus, with {769501 class, 769501 non-class} speech vectors and {429826 class, 429826 non-class} music vectors (cf. Table 2). Test corpus is about 1-hour length and the best scores are 97.63 % accuracy for music and 98.93 % for speech (cf. Table 1).

When the number of vectors is high (more than 30000) it is difficult to implement SVM (event though we use the decomposing algorithm [1]). We use two techniques to reduce the learning corpus: one choice is to take randomly 1 vector each 150 samples from the corpus; the other is to invoke VQ algorithm as a means of summarizing the information with minimal distortion. The first technique is better for speech and the other for music.

We implement cross-validation procedures to test several kernels functions and parameters: polynomial, sigmoid and radial basis functions are revisited. The latter achieves the best scores and the fastest processing time. With 2048 vectors from VQ, SVM accuracy is 97.53 % for music and with random set reduction, SVM accuracy for speech attains 96.4 %.

4.3. Laughter and applause detections

We used one 3-hours broadcast as learning corpus from which we got 4 minutes of applause and 1 minute of laughter. The test broadcast last also 3-hours.

Table 1. Classification results.

System	Music	Speech	Applause	Laughter
GMM	97.63 %	98.93 %	98.58 %	97.26 %
SVM	97.53 %	96.4 %	98.35 %	97.12 %

GMM and SVM systems give equivalent results for each kind of sounds (c.f. Table 1) and we can say we detect the most important events present in broadcasts. For SVM input we perform random set reduction to make the classification problem solvable (cf. Table 2).

Table 2. Number of vectors used for training.

System	Music	Speech	Applause	Laughter
GMM				
class	429826	769501	7512	1714
nonclass	429826	769501	337160	342874
SVM				
class	1024	11333	7512	1714
non class	1024	5130	20000	20000

In the last test we built a background non-class model for applause sounds that was tested using the same vectors for both technologies (cf. Table 3). We observe a dependence of GMM scores respect to the number of training vectors available, but a faster performance than SVM. We remark that SVM scores do not vary too much when we reduce the number of vectors for training.

Table 3. Results for applause indexing using the same training vectors.

System	Class vectors	Score	CPU[1] time - training	CPU time - classif.
GMM	7500	98.58%	2' 55''	3' 25''
GMM	2500	95.04%	57''	3' 03''
SVM	7500	98.35%	6' 28''	22' 23''
SVM	2500	97.56%	2' 29''	12' 15''

5. CONCLUSIONS

We have developed a framework to compare two different technologies of classification. GMM is a robust classical approach which obtained very good performance. SVM come from a strong mathematical background. Experiments show advantages and disadvantages of each method. For each kind of sound (speech, music, applause, laughter) we have equivalent accuracy scores for GMM and SVM, but we need less training vectors for SVM models. SVM optimization solution is seriously compromised when we manage important data volumes while GMM needs a big amount of data to train proper learning models. For speech or music detection, where the available volume of data can be very important, a solution based on GMM is powerful. On the other hand, for rarer sounds, like applause or laughter, SVM, requiring less data, is the best alternative.

[1] Processor Pentium IV 2.5GHz, 1.5 Gb of RAM

6. ACKNOWLEDGMENTS

This work has received financial support from the Mexican National Council of Science and Technology. We thank INA (Institut National de l'Audiovisuel) for give us authorization of using TV shows for research purposes.

7. REFERENCES

[1] C-C Chang, C-J Lin, « LIBSVM: a Library for Support Vector Machines, National Taiwan University, 2004.

[2] M. Davy, S. J. Godsill, « Audio information retrieval: A bibliographical study », CUED/F-INFENG/TR429, Cambridge University Engineering Department, February 2002.

[3] R. O. Duda, P. E. Hart, D. G. Stork, « Pattern Classification», Ed. John Wiley & Sons, 2001.

[4] J.L. Gauvain, L. Lamel and G. Adda, "Audio partitioning and transcription for broadcast data indexation", *CBMI'99*, Toulouse.

[5] L. Lu, H-J Zhang, H. Jiang, « Content Analysis for Audio Classification and Segmentation », IEEE Trans. on Speech and Audio Processing, vol. 10, n°7, October 2002.

[6] M. T. Maybury (Editor), « Intelligent Multimedia Information Retrieval ». The MIT Press, 1997.

[7] J. Pinquier, A. Arias, R. André-Obrecht, « Audio classification by search of primary components », International workshop on Image, Video and Audio Retrieval and Mining, Québec, Canada, October 2004.

[8] J. Rissanen, « An universal prior for integers and estimation by minimum description length ». The Annals of Statistics, vol 11, pp. 416-431, November, 1982.

[9] S. Rossignol, X. Rodet, J. Soumagne, J.L. Collette and P. Depalle, « Automatic characterisation of musical signals: feature extraction and temporal segmentation », *Journal of New Music Research*, 2000, pp. 1-16.

[10] B. Schölkopf, A. Smola, « Learning with Kernels », Ed. The MIT Press, 2002.

[11] E. Scheirer, M. Slaney, « Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator », ICASSP'97, Munich, Vol. II, pp. 1331-1334, 1997.

[12] C. Tomasi, « Estimating Gaussian Mixture Densities with EM – A Tutorial », http://citeseer.nj.nec.com.

[13] V. Vapnik, « The Nature of Statistical Learning Theory », Ed. Springer, 1995.

[14] M. Witbrock, A. Hauptmann, « Speech recognition for a digital video library ». http://www.informedia.cs.cmu.edu.