Audio classification by search of primary components

Julien PINQUIER, José ARIAS and Régine ANDRÉ-OBRECHT
Équipe SAMOVA, IRIT, UMR 5505 CNRS INP UPS
118, route de Narbonne, 31062 Toulouse cedex 04, FRANCE
{pinquier, arias, obrecht}@irit.fr

Abstract

This work addresses the soundtrack indexing of multimedia documents. Our purpose is to detect and locate sound unity to structure the audio dataflow in program broadcasts. We present three different audio classification tools that we have developed. The first one, a speech/music classification tool, is based on three original features: entropy modulation, stationary segment duration and number of stationary segments. It provides about 90% of accuracy for speech and music detection. Another system, a jingle identification tool, uses an euclidean distance in the spectral domain to index the audio data flow. Results show its efficiency: among 132 jingles to recognize, we have detected 130. The last one, a key sound classification tool, permits to extract applause and laughter. Results are perfect for applause (only few shift) and quite good for laugther (some missing). Systems are tested on TV and radio corpora (more than 12 hours).

1 Introduction

To process the quantity of audiovisual information available 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 studied. Nevertheless all these detection systems presuppose the extraction of elementary and homogeneous acoustic components.

In most studies, the first partionning in audio indexing task consists on speech/music discrimination. We observe two tendencies. On one hand, the musician community that gives greater importance to features which increase a binary discrimination: for example, the zero crossing rate and the spectral centroid are used to separate voiced speech from other sounds [1], the variation of the spectrum magnitude attempts to detect harmonic continuity [2]. On the other hand, the automatic speech processing community that prefers cepstral analysis [3]. Two concurrent classification frameworks are usually investigated : the Gaussian Mixture Model (GMM) framework and the k-nearest-neighbors one [4].

In this paper, we present a system able to detect these two basic components (speech and music) with an equal performance and we explore other prior partionning. It consists in detecting pertinent key sounds (like applause, laughter or jingles). There is no intention to do a topic segmentation task [5], but the purpose is to propose an audio macro-segmentation by finding the temporal structure of broadcast program. When we say “jingle”, we mean a redundant audio part of few seconds (about three seconds in our collection). In audio documents like TV or radio broadcasting, they are used to announce the beginning and the end of a segment: weather report, news and adverts. In [6], jingle detection appears as an interesting way to audiovisual classification.

This paper is divided into four parts. First, we describe our speech/music classification system and in particular three original parameters: entropy modulation, stationary segment duration and number of segments. Then, we present our jingle classification system that permits to detect and identify any reference jingle on an audio source. After, two key sounds (applause and laughter) are extracted with use of spectral coefficients. The modeling is based on a GMM. Finally, we perform, for each system, test experiments on TV and radio documents (more than 12 hours).

2 Speech/Music classification system

This system results of the fusion of two detection subsystems: speech detection and music detection (figure 1). For the speech detection, we have used entropy modulation and 4 Hz modulation energy and for music, number of segments and segment duration. For each classifier, we propose a statistical model. The decision is made regarding to the maximum likelihood criterion (scores). Finally, we have two classifications for each second of input signal: the speech/non-speech one and the music/non-music one.

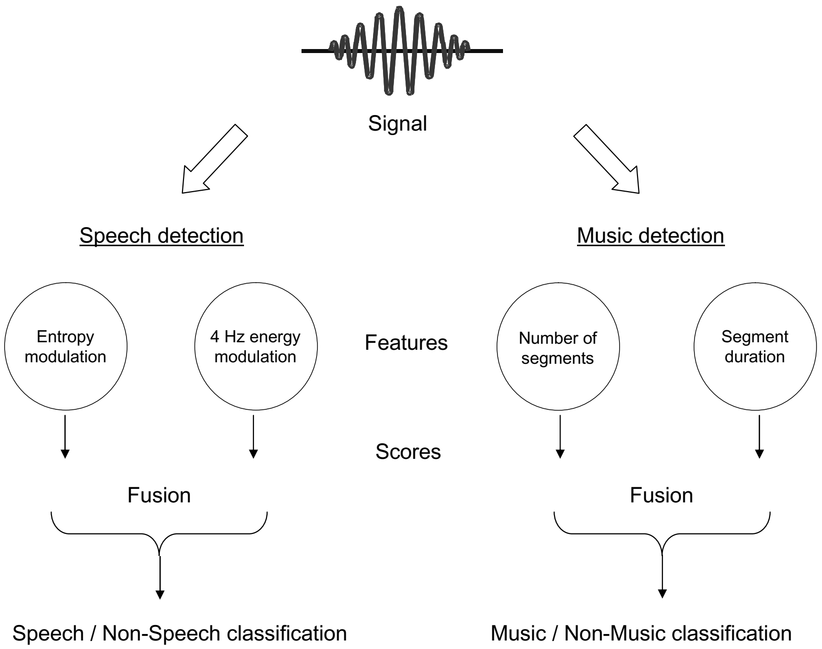

Figure 1: Speech/Music classification system.

2.1 Original features

These features are more detailed in [7].

- 4 Hz modulation energy

Speech signal has a characteristic energy modulation peak around the 4 Hz syllabic rate [8]. Speech carries more modulation energy than music.

- Entropy modulation

Music appears to be more "ordered" than speech considering observations of both signals and spectrograms. To measure this "disorder", we evaluate a feature based on signal entropy ($H = \sum_{i=1}^{k} -p_i log_2 p_i$, with p_i=probability of event i). Entropy modulation is higher for speech than for music.

- Segmentation features

The segmentation is provided by the ”Forward-Backward Divergence algorithm” [9] which is based on a statistical study of the acoustic signal. Assuming that the speech signal is described by a string of quasi-stationary units, each one is characterized by an auto regressive gaussian model. The method consists in performing a detection of changes in the auto regressive parameters.

Number of segments: speech signal is composed of alternate periods of transient and steady parts. Meanwhile, music is more constant. The number of changes will be greater for speech than for music. To estimate this feature, we compute the number of segments on one second of signal, and we model it by gaussian laws.

Segment duration: segments are generally longer for music than for speech. We use segment duration as a feature. We decided that is appropiate to model sound duration by a gaussian inverse law [10]. The probability density function (pdf) is given by:

$$p(g) = \sqrt{\frac{\lambda}{2\pi g^3}} * e^{\frac{-\lambda(g-\mu)^2}{2\mu^2 g}}, g \geq 0$$

with μ = mean value of g and $\frac{\mu^3}{\lambda}$ variance of g.

3 Jingle classification system

Our jingle classification system is divided in three principal parts, frequently used in pattern recognition problem: an acoustic preprocessing module, a detection module and an identification module (figure 2).

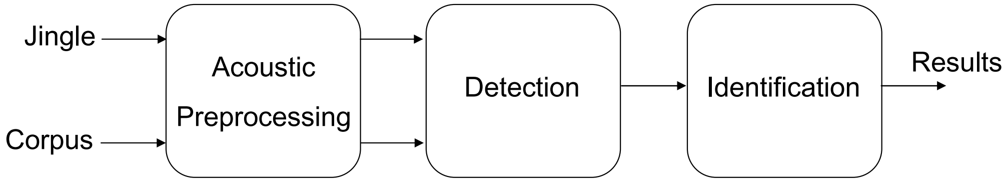

Figure 2: Jingle classification system.

3.1 Acoustic preprocessing

The acoustic preprocessing consists of a spectral analysis. The signal is windowed into frames of 32 ms in length, with adjacent frames overlapping by 16 ms. For each frame, we have an acoustic vector of 29 spectral coefficients: 29 output filters covering the frequency range 100 Hz - 8 kHz [11].

3.2 Detection

Jingle is characterized by a sequence of N spectral vectors which is called the "signature" of the jingle. The size N is the number of analysis frames. The detection consists in finding this sequence in the data flow. So the data flow is transformed into a sequence of spectral vectors. The jingle signature and the data flow (N adjacent vectors extracted) are compared using an euclidean distance.

We select the potential candidates by defining minimum values. We calculate the mean value of the distance. If the current value (jingle/flow distance) is lower than the half of this mean, we decide that it is a minimum value. We only keep as jingle occurence candidates, the local minima extracted on these minimum values. With this stage, we detect candidates who have the same spectral information as the reference jingle.

3.3 Identification

We have noticed that all minima, corresponding of the reference jingle, have a common particularity: they have, without exception, a fine width (figure 3). So, we analyse the peak width (L) of each detected local minimum.

- h the current value of the local minimum,
- L the peak width at the height H, where H is the height where we estimate the width peak. Naturally H and h must be tied.

If $L < \lambda$ (threshold), the peak width is fine and the local minimum is a "good" jingle, else the candidate is rejected ("bad" jingle). This system is more detailed in [12].

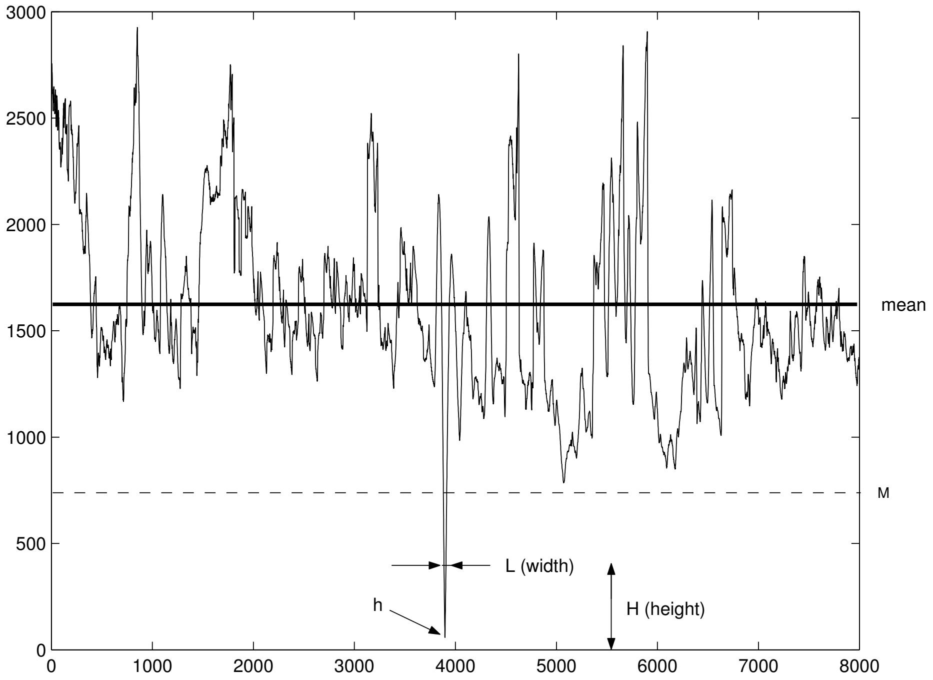

Figure 3: Jingle identification.

4 Key sound classification system

Extraction of training vectors to estimate pdfs for applause and laugther models has been made in a separate way. The system is divided in three modules: acoustic preprocessing (cf. 3.1), training and classification.

4.1 Classification

For each key sound, we have chosen to model the Class (Applause, Laugther) and the Non-class (Non-Applause, Non-Laugther) by a GMM. The classification by GMM is made by computing the log-likelihood among test vectors and each model of Class and Non-Class, assigning to vectors the label of the model with highest score. Following this classification phase, a phase of merging allows to concatenate neighboring frames having obtained the same label during the classification. A smoothing function is necessary to delete insignificant size segments and to keep relevant zones of sounds. This smoothing is 1000 ms.

4.2 Training

The training of GMM consists in an initial estimation of their parameters followed by an optimization step. The initialization step is performed using Vector Quantization (VQ) based on the algorithm of Lloyd [13]. The optimization of the parameters is made by the classic Expectation-Maximization (EM) algorithm [14]. After experiments, the number of gaussian laws in the mixture has been fixed to 64 for Applause model and 128 for Laughter model (figure 4).

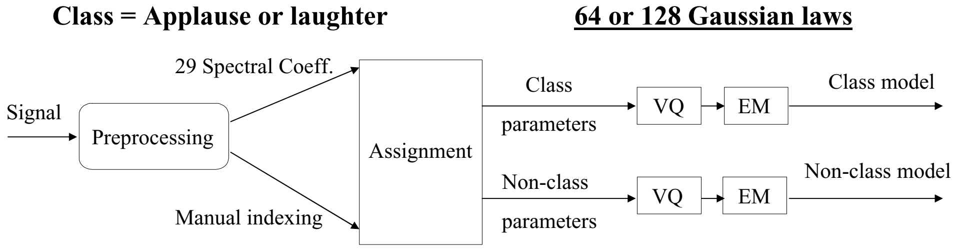

Figure 4: Training of GMM.

5 Experimental results

5.1 Corpus

Our database is very diversified. The speech/music system was trained with read speech (30 mn) and different kind of music excerpts (30 mn). In the RFI (Radio France Internationale) corpus we have TV and radio programs (news, songs, commercials, reports...) in 3 different languages: english, french and spanish. The total duration is about 12 hours from which we had about 7 hours of speech and 3 hours of music to test the speech/music and the jingle detection systems.

More than 50 different jingles appear on the whole database. Our goal is to detect, locate and identify only reference jingles. The reference jingle table is composed of 32 different key sounds extracted on the database. The detections may be identical to the reference one or superimposed to speech if the speaker speaks at the same time. Therefore we have to recognize 132 jingles among 200.

For the key sound system, we used a 6 hours corpus from the TV show "le grand echiquier" (french variety, interviews, gags). We have used 3 hours for training from which we got 4 and 1 minutes of significant (long and clean signals) applause and laughter respectively. We have trained four models: Applause, Non-Applause, Laughter and Non-Laughter.

5.2 Evaluation

- Speech / Music classification system

We have tested separately all the parameters. The experiments (Table 1) provide similar accuracy (about 87 %) for entropy modulation and 4 Hz modulation energy. The number of segments gives about the same accuracy for music detection. Only the Bayesian approach with segment duration and Gaussian Inverse law gives a lower accuracy rate (78 %). The final performance of our system is 90.5 % of accuracy for speech detection and 89 % of accuracy for music detection.

Table 1: *Speech / Music classification.*

Features	Accuracy
(1) 4 Hz Modulation energy	87.3 %
(2) Entropy modulation	87.5 %
(3) Number of segments	86.4 %
(4) Segments duration	78.1 %
(1) + (2) Speech detection	**90.5 %**
(3) + (4) Music detection	**89 %**

- Jingle classification system

The detection is very good: we have no false alarm and only two omissions whereas there were other jingles in the database. Among 132 jingles which must be detected and identified, we have detected 130 (98.5% of accuracy). The two omitted jingles are completely recovered by speech. Considering this variety, the system has a correct behavior. That proves the robustness of our system.

During the evaluation phase, we have studied the precision of the detection. Differences between manual and automatic boundaries are no more than a half second. For an indexing task, the decision is generally taken on every second of the signal. This localization is amply sufficient.

- Key sounds classification system

Applause events are stable signals that are easy to detect even if their boundaries are not always precise, often mixed with music or shouts. For laughter, the main problem is to find a learning corpus that includes all possible ways of laughing. In our 3 hours test corpus we have 10 and 6 minutes of significant applause and laughter respectively to identify (Table 2). Our test results are explained as follows: we detect the most important events and we miss several applause and laughter signals that are not well defined or "polluted" with other informations.

Table 2: *Key sound classification.*

Features	Applause	Laughter
Manual: significant segment	72	175
Manual: total segment	144	359
Automatic	97	102
Accuracy (NIST evaluation)	**98.58 %**	**97.26 %**

- Qualitative examples

Figure 5 gives an example of speech / music classification and jingle detection in the RFI corpus. We can note that the majority of jingles is classified as music. We have also noticed that applause often arrived after music (song). Figure 6 gives an example in a TV corpus.

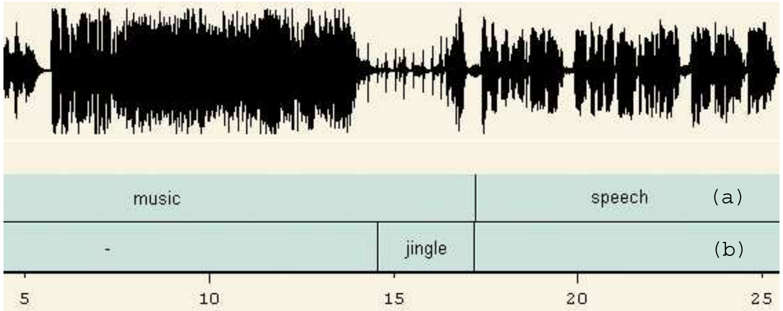

Figure 5: Example of first partionning: (a) speech / music classification and (b) jingle detection.

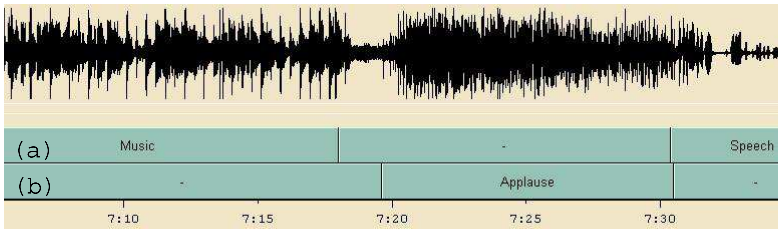

Figure 6: Example of first partionning: (a) speech / music classification and (b) applause classification.

6 Discussion

The first presented system is a speech/music classifier. We have processed four features: entropy modulation, number of segments, segment duration and 4 Hz modulation energy. Considered separately, all those features are relevant. The combination of those approaches allows to raise the accuracy rate up about 90%. Four features and four pdfs are sufficient. Note that training of these models was performed on personal database (different of the RFI database): This system is robust and task-independent.

Then, we describe a jingle classification system. This study is based on an euclidean distance in the spectral domain. The results are very satisfactory because we have no false alarm and only two missings (in extreme conditions). Our jingle system is real-time, robust, and have good results, so it is efficient.

The key sounds detection system based in differentiated modeling using gaussian mixture models gives encouraging results because we detect main trained events and we can easily extend it to identify other environment sounds.

Each primary component can be used for high-level description of audio documents, which is essential to index (or structure) program broadcasts (reports). This work could be extended by the adding of video track to perform sequences detection or by defining an audio-video jingle model.

References

[1] J. Saunders, “Real-time discrimination of broadcast speech/music,” in *International Conference on Audio, Speech and Signal Processing*, Atlanta, USA, May 1996, pp. 993–996, IEEE.

[2] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *International Conference on Audio, Speech and Signal Processing*, Munich, Germany, Apr. 1997, pp. 1331–1334, IEEE.

[3] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *IEEE International Conference on Multimedia and Expo*, New-York, USA, 2000, pp. 452–455, IEEE.

[4] M. J. Carey, E. J. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," in *International Conference on Audio, Speech and Signal Processing*, Phoenix, USA, Mar. 1999, pp. 149–152, IEEE.

[5] R. Amaral, T. Langlois, H. Meinedo, J. Neto, N. Souto, and I. Trancoso, "The development of a portuguese version of a media watch system," in *European Conference on Speech Communication and Technology*, Aalborg, Denmark, Sept. 2001.

[6] J. Carrive, F. Pachet, and R. Ronfard, "Clavis - a temporal reasoning system for classification of audiovisual sequences," in *Proceedings of Content-Based Multimedia Information Access (RIAO) Conference*, Paris, France, Apr. 2000.

[7] J. Pinquier, Jean-Luc Rouas, and R. André-Obrecht, "Robust speech / music classification in audio documents," in *International Conference on Spoken Language Processing*, Denver, USA, Sept. 2002, vol. 3, pp. 2005–2008.

[8] T. Houtgast and J. M. Steeneken, "A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria," *Journal of the Acoustical Society of America*, vol. 77, no. 3, pp. 1069–1077, 1985.

[9] R. André-Obrecht, "A new statistical approach for automatic speech segmentation," *IEEE Transactions on Audio, Speech, and Signal Processing*, vol. 36, no. 1, Jan. 1988.

[10] N. Suaudeau and R. André-Obrecht, "An efficient combination of acoustic and supra-segmental informations in a speech recognition system," in *International Conference on Acoustics, Speech and Signal Processing*, Adélaide, Australie, April 1994, IEEE.

[11] J. Pinquier, C. Sénac, and R. André-Obrecht, “Indexation de la bande sonore : recherche des composantes parole et musique,” in *Congrès de Reconnaissance des Formes et Intelligence Artificielle*, Angers, France, Jan. 2002, pp. 163–170.

[12] J. Pinquier and R. André-Obrecht, “Jingle detection and identification in audio documents,” in *International Conference on Audio, Speech and Signal Processing*, Montréal, Canada, May 2004.

[13] J. Rissanen, “An universal prior for integers and estimation by minimum description length,” *The Annals of Statistics*, vol. 11, pp. 416–431, Nov. 1982.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society*, vol. 39 (Series B), pp. 1–38, 1977.