JINGLE DETECTION AND IDENTIFICATION IN AUDIO DOCUMENTS

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
{pinquier, obrecht}@irit.fr

ABSTRACT

This work addresses the soundtrack indexing of multimedia documents. Our purpose is to detect and locate one or many jingles to structure the audio dataflow in program broadcasts (reports). Each jingle is commonly represented by a sequence of spectral vectors, considered as its "signature". Potential candidates are extracted from the data flow by computing an Euclidean distance. They are validated with heuristic rules. The system evaluation is performed on TV and radio corpora (more than 10 hours, 3 TV channels and 3 radio channels). First results show that the system is efficient: among 132 jingles to recognize, we have detected 130 with our reference jingle table of 32 different key sounds.

1. INTRODUCTION

More and more audiovisual information is available from many sources around the world. The information may be represented in various forms like pictures, videos, audios. To process this information in a smart and rapid way, it is necessary to have robust tools to describe the content of this important volume of media sources. To index an audio document, key words or melodies are semi-automatically extracted, speakers are detected... More recently, the problem of topics retrieval has been studied [1]. Nevertheless all these detection systems presuppose the extraction of elementary and homogeneous acoustic components.

In most studies, the first partionning in audio indexing consists on speech/music discrimination. Different tendencies are observed:

- the musician community gives greater importance to features which increase a binary discrimination [2]: for example, the zero crossing rate and the spectral centroid are used to separate voiced speech from other sounds [3] as the variation of the spectrum magnitude attempts to detect harmonic continuity [4].

- the automatic speech processing community prefers cepstral parameters [5]. Two concurrent classification frameworks are usually investigated : the Gaussian Mixture Model (GMM) framework and the k-nearest-neighbors one [6].

- in a previous paper, we have studied a fusion of these methods and robust results are provided [7].

An alternative to this approach on a complementary partionning consists in detecting pertinent key sounds, so called jingles, which reveal the beginning or the end of a broadcast or announce it. There is no intention to do a topic segmentation task [8], but the purpose is to propose an audio macro-segmentation by finding the temporal structure of broadcast program. In [9], jingle detection appears as an interesting way to audiovisual classification.

In MPEG7 audio specifications (sound recognition section) [10], there is a non-normative list of sound effect categories (with many examples) which could be used to describe audio documents. The idea is to create a dynamic table of reference sounds and to structure audio documents by detecting and locating their occurrences.

This study lies in this scientific framework. We are interested in "jingle" which designs key sounds of few seconds (about three seconds in our collection). Such a jingle has the particularity to contain speech, music or noise and many occurrences of it may be observed. To detect and locate it during the audio data flow, only an example of this jingle is necessary. The so called reference jingle is described by low-level audio descriptors based on a spectral analysis and dissimilarity is measured with Euclidean distances.

This paper is divided into two classical parts. First, we describe our classification system that permits to detect and identify any reference jingle along an audio flow. Then, we present experiments performed on radio and TV documents.

2. CLASSIFICATION SYSTEM

Our classification system is divided in three main parts, frequently used in a pattern recognition problem:

- an acoustic preprocessing module to characterize the signal by parameter vectors,

- a detection module to propose some jingle candidates,
- an identification module to confirm (or cancel) the precedent candidates.

2.1. Acoustic preprocessing

The acoustic preprocessing consists of a spectral analysis. The signal is windowed into frames of 32 ms length, with a 16 ms overlap. The frames are preemphasized and we apply a Hamming window before computing a Fast Fourier Transform (FFT). To reduce the number of spectral coefficients, we use a spectral filtering which we tested in a previous study on speech/music classification[11]. 29 channels are derived from a piece linear scale. Each acoustic vector is composed of 29 spectral coefficients covering the frequency range [100 Hz - 8 kHz]. Spectra are normalized by their mean energy to be independent of the recording level.

2.2. Detection

A reference 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 large sequence of spectral vectors. The signature and the data flow are compared using an Euclidean distance.

The sequence (of N adjacent vectors extracted from the data flow) is moving with a step of S vectors at each comparison (Figure 1).

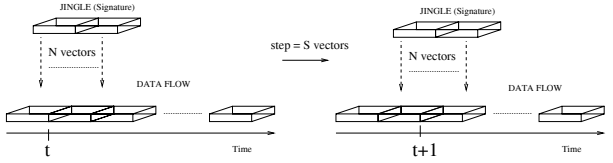

Fig. 1. Comparaison jingle/corpus.

We select the potential candidates by defining minimum values. We calculate the mean value of the distance. If the current distance value is lower than the half of this mean, named M, we decide that it is a minimum value (Figure 2).

We only keep as jingle occurence candidates, the local minima extracted from these minimum values.

2.3. Identification

The figure 2 illustrates an example of the results obtained by computing the Euclidean distance between the reference jingle and the data flow. Five principal minima are first selected. The first two ones correspond to a "good" jingle (the researched jingle). The three other ones indicate the presence of jingles but these jingles do not belong to the reference table. So to select the correct jingle, we propose the following processing.

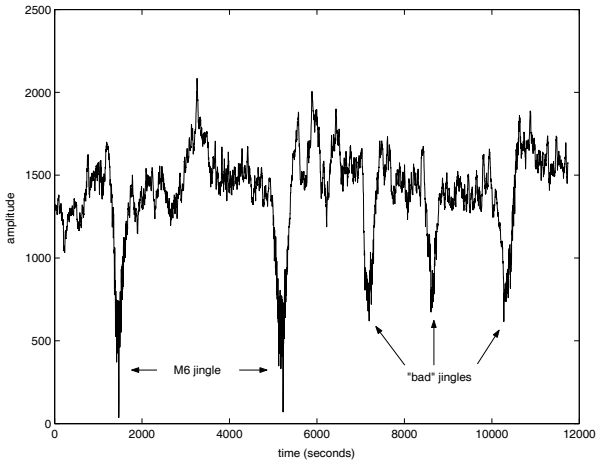

Fig. 2. Euclidean distance during the detection of the "M6 jingle": 3 minutes of signal (corpus M6).

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

- 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 (cf. 3.2).

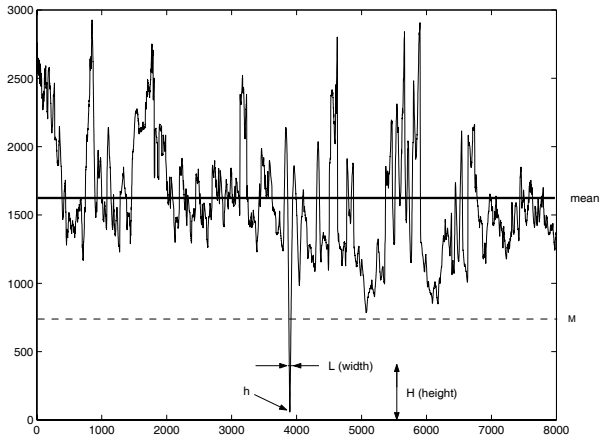

Fig. 3. Distance between the jingle and the data flow.

We introduce a threshold λ. If $L < \lambda$, the peak width is fine and the local minimum is a "good" jingle, else the candidate is rejected ("bad" jingle). Obviously, the threshold λ must be proportional to N.

3. EXPERIMENTS

3.1. Corpus

Our database is made up of six different corpora (Table 1). The total duration is about 10 hours. This database is sampled at 16 kHz.

Table 1. *Description of the database.*

Corpus	Duration	Key sounds	Occurrences
France 3	15 mn	1	4
M6	15 mn	1	16
Canal+	30 mn	1	6
France Info	60 mn	1	12
RFI	360 mn	3	60
Commercials	90 mn	25	34
Total	**570 mn**	**32**	**132**

- France Info corpus is a selection of radio programs: majority of news (current events, weather, sports and reports) but also commercials and musical extracts.
- France 3 corpus is composed of TV soundtracks: many commercials and two songs.
- Canal+ and M6 corpora are French TV newscasts.
- RFI corpus gathers multilingual broadcasts (interviews, reports and news) of Radio France Internationale.
- The last one is a corpus of radio commercials.

More than 50 different jingles appear on the whole of the database. Our goal is to only detect, locate and identify reference jingles. These 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 the 200. The reference jingle table is composed of 32 different key sounds extracted on the database. The jingle duration is between one and five seconds. These jingles are a selection of the jingles present in the corpus.

3.2. Training

To implement the identification method, it is necessary to fix several parameters S, H and λ. For that purpose, we examine the behavior of the distance while processing France 3 and M6 corpora. It appears that S must take large values without degrading the processing. This important delay (more than 1 second for a 16 kHz sampling rate) permits to process the data flow "on-line" (in real-time).

We fix $H = \alpha.h$ where α is a constant value ($\alpha = log2$). The reject threshold λ depends on the ratio between the length of the reference jingle N and the analysis delay S. Experimentaly we find $\lambda = 5 * N/S$.

3.3. Results

We have tested each reference jingle in **all** corpora (Table 2). 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 and their peaks are too large (Figure 4).

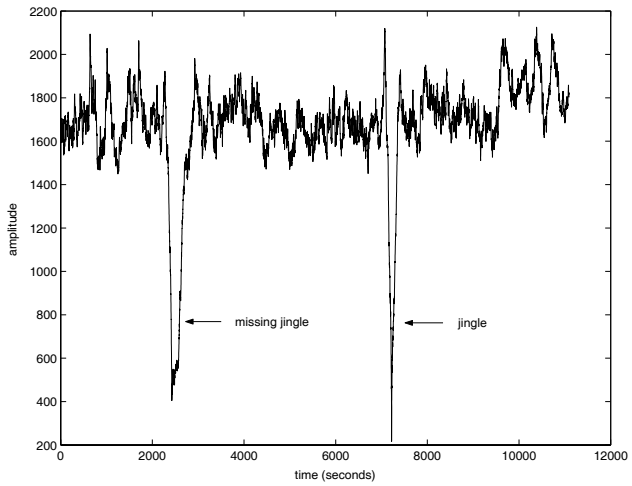

Fig. 4. Detection error (missing) of the "France Info jingle" on 3 minutes of signal (corpus France Info).

The detection is very good: we have no false alarm and only two omissions whereas there were other jingles in the database. We have tested **each** reference jingle in **all** corpora (Table 2). The training corpora for S, H and λ are colored in the table.

Table 2. *Manual and automatic jingle detection for each corpus.*

Corpus	Automatic detection	Manual detection
France 3	4	4
M6	16	16
Canal+	6	6
France Info	11	12
RFI	60	60
Commercials	33	34
Total	**130**	**132**

Considering the variety of the database, TV and radio programs having different conditions of recording, the system has a very correct behavior. Furthermore, the experiments prove the robustness of our system.

During the evaluation phase, we have studied the precision of the detection. The jingle localization gives satisfactory results. Differences between manual and automatic localizations are no more than a half second for each jingle ($S/2$). In an audio indexing task, many decisions are generally taken on every second of the signal. This localization is amply sufficient.

We can note that our system works real-time: for a sound file of 30 minutes, less than 30 minutes are necessary for calculations of all 32 jingles with a processor which is running at 1.4 GHz (AMD processor).

4. DISCUSSION

We present a jingle detection and identification system for audio indexing. Based on an Euclidean distance in a spectral domain, this method is very simple. Nevertheless the results are very satisfactory because we observe no false alarm and only two omissions in extreme conditions (speech superimposed during **all** the jingle time). The localization is very good: we can determinate the beginning of a jingle with a margin of half second (sufficient in an indexing task).

Our system is simple (only based on a spectral analysis), real-time ("on-line"), robust (task-independent) and has good results, so it is efficient. It can be used for high-level description of audio documents, which is essential to structure (or classify) broadcasts programs. This work will be extended by adding video macro-segmentation [12]. We hope to define an audio-video signature of broadcasts programs and appropriate audio-video measure of dissimilarity.

5. ACKNOWLEDGEMENTS

We would like to thank the CNRS (French national center for scientific research) for its support to this work under the RAIVES project.

6. REFERENCES

[1] M. Franz, J. Scott McCarley, T. Ward, and W. Zhu, "Topics styles in ir and tdt: Effect on system behavior," in *European Conference on Speech Communication and Technology*, Aalborg, Denmark, Sept. 2001, pp. 287–290.

[2] S. Rossignol, X. Rodet, J. Soumagne, J. L. Collette, and P. Depalle, "Automatic characterization of musical signals: feature extraction and temporal segmentation," *Journal of New Music Research*, vol. 28, no. 4, pp. 281–295, Dec. 1999.

[3] 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.

[4] 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.

[5] J. L. Gauvain, L. Lamel, and G. Adda, "Systmes de processus lgers : concepts et exemples," in *International Workshop on Content-Based Multimedia Indexing*, Toulouse, France, Oct. 1999, pp. 67–73, GDR-PRC ISIS.

[6] 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.

[7] J. Pinquier, Jean-Luc Rouas, and R. André-Obrecht, "A fusion study in speech / music classification," in *International Conference on Audio, Speech and Signal Processing*, Hong-Kong, China, Apr. 2003.

[8] 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.

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

[10] ANSI, "Iso/iec 15938-4 information technology - multimedia content description interface - audio," Tech. Rep., MPEG, 2001.

[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] P. Aigrain, P. Joly, and V. Longueville, "Medium knowledge-based macro-segmentation of video into sequences," *In Intelligent multimedia information retrieval*, pp. 159–173, 1997.