EVENTS DETECTION FOR AN AUDIO-BASED SURVEILLANCE SYSTEM

C. Clavel [1,2], *T. Ehrette* [1], *G. Richard* [2]

[1]Thales Research and Technology France, Domaine de Corbeville, 91404 Orsay Cedex, France
[2] GET-ENST, 46 rue Barrault, 75634 Paris Cedex 13, France

ABSTRACT

The present research deals with audio events detection in noisy environments for a multimedia surveillance application. In surveillance or homeland security most of the systems aiming to automatically detect abnormal situations are only based on visual clues while, in some situations, it may be easier to detect a given event using the audio information. This is in particular the case for the class of sounds considered in this paper, sounds produced by gun shots. The automatic shot detection system presented is based on a novelty detection approach which offers a solution to detect abnormality (abnormal audio events) in continuous audio recordings of public places. We specifically focus on the robustness of the detection against variable and adverse conditions and the reduction of the false rejection rate which is particularly important in surveillance applications. In particular, we take advantage of potential similarity between the acoustic signatures of the different types of weapons by building a hierarchical classification system.

1. INTRODUCTION

Audio events classification/detection is receiving a growing interest by the scientific community. It is especially the case in the context of audio retrieval and indexing applications but also in the context of multimedia event detection applications where audio can be used as a complementary source of information [1], [4], [8]. In surveillance or homeland security (security of public places such as bank, subway, airport,...) most of the systems are only based on visual clues to detect abnormal situations. Typical abnormal situations include natural damages such as fires, earthquakes, flood etc, physical or psychological threatening and aggression against human beings (kidnapping, hostages etc). In some of these situations audio conveys a more significant information than video. Our goal is then to use acoustic clues as complementary information to automatically detect and analyse abnormal situations.

A complete multimedia automatic surveillance system would then consist of different modules providing information from different modalities that will be merged by an information fusion system for situation analysis (see figure 1). In this targeted system, the audio module will use vocal and non vocal manifestations of abnormal situations and will deal with both emotional content [2] and typical events, such as cries, shots or explosions. In this paper we propose an approach to develop an audio key-event detection system. Although our event detection system is currently limited to shot detection, the methodology and the approach followed for this system could be extended to other classes of characteristic sounds of abnormal situations in a given environment.

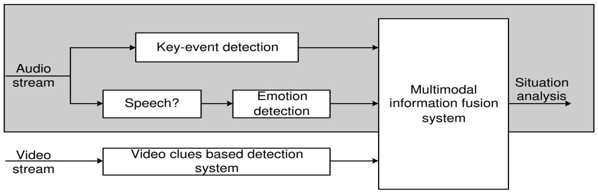

Fig. 1. *Multimedia event detection system architecture*

One of the major difficulties of an audio detection system is linked to the environmental noise that is often non-stationary and that may be loud compared to the audio event to detect. The shot detection system presented in this paper is based on a novelty detection approach [6]. Indeed novelty detection offers a solution to detect abnormality (abnormal audio events) when a given distance to a model of the *normal* situation (built from acoustic data of a given environment) exceeds a predefined threshold. The focus of this paper is on two of the main problems of an automatic audio event detection system, namely the robustness of the detection against variable and adverse conditions and the reduction of the false rejection rate which is particularly important in surveillance applications. In particular, we take advantage of potential similarity between the acoustic signatures of the different types of weapons by building a hierarchical classification system.

The paper is organized as follows. First, our shot detection system is described in section 2. Then, the database and test protocol used to evaluate the system are given in section 3. The different experiments and results obtained are presented in section 4. Finally, we suggest some con-

0-7803-9332-5/05/$20.00 ©2005 IEEE

clusions and future work in section 5.

2. THE SHOT DETECTION SYSTEM

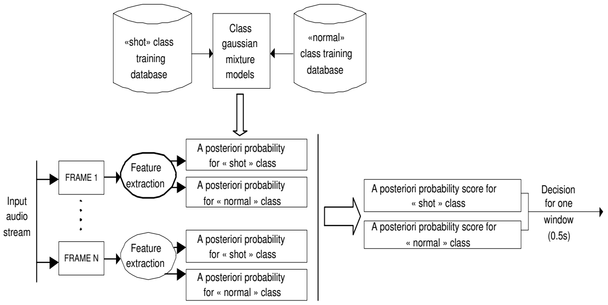
«shot» class training database
Class gaussian mixture models
«normal» class training database
Input audio stream
FRAME 1
FRAME N
Feature extraction
Feature extraction
A posteriori probability for « shot » class
A posteriori probability for « normal » class
A posteriori probability for « shot » class
A posteriori probability for « normal » class
A posteriori probability score for « shot » class
A posteriori probability score for « normal » class
Decision for one window (0.5s)

Fig. 2. *Gaussian mixture models detection system.*

The goal of our shot detection system is to segment the input audio stream into successive segments and to label these segments according to the two main classes (the *shot* class and the *normal* class that represents the environment acoustic characteristics). The architecture of our audio event detection system includes a feature extraction module, a training module that is used to build the model of the two classes (using Gaussian Mixture Models or GMM) and a classification module that, based on the previous models, labels the successive audio segments. As depicted on figure 2, the input audio stream is first segmented in short frames (20 ms) but a label is only given for segments of 0.5 second (with 50% overlap).

2.1. Audio features extraction

Feature extraction is made on 20 ms analysis frames with a 50% overlap. Computed features are chosen among the most popular in audio processing algorithms and the more likely to fit with our classification problem.
-*Short time energy* describes a signal energy at a given time and is alternatively referred to as loudness or volume.
-The first eight *Mel-Frequency Cepstral Coefficients*.
-The first two *spectral statistical moments*, namely the spectral centroid which is the mean of the power spectrum for a given time and the spectral spread.
-The *first and second derivatives* of each of the above features.
Feature vector dimension is then reduced by principal component analysis procedure. We keep the 13 first components as significant. Each analysis frame of the input audio signal is thus represented by a 13-dimensional vector.

2.2. The training step

For each class a Gaussian Mixture Model (GMM) is built. The appropriate number of Gaussian for each class is estimated thanks to the Bayesian Information Criterion [3]. The parameters of the models are estimated using the traditional Expectation-Maximization algorithm [7] initialized by a basic binary splitting vector quantization algorithm.

2.3. The detection step

Detection is made using the Maximum A Posteriori (MAP) decision rule : the mean a posteriori log-probability on a 0.5-second decision window is computed for each class model (by multiplying the probability obtained for each short time analysis frame). The decision window is then classified according to the class that has the maximum a posteriori score. Silence windows are not considered and are automatically removed.

3. DATABASE AND PROTOCOL

3.1. Database

Corpora of typical audio events in ecological conditions, such as surveillance applications, are not available mainly because of the confidential nature of the data but also because abnormal situations are rarely recorded. To be as close as possible to real conditions for our application, we have built artificial data from a set of multiple public places and gun shots recordings extracted from a CD of sounds for the national French public radio [5].

- The shot-event database: a total of 134 shots (296 seconds) composed by pistol (P), rifle (R), submachine gun (S), grenade (G) and cannon fire (C) are extracted. Description of the weapon repartition in *shot* class data is presented in table 1.

weapon	P	R	S	G	C
files number	5	15	79	8	27
duration	5s	24s	134s	28s	105s

Table 1. The shot-event database

- The environmental database: the CD provides various public places recordings (mainline station, airport, stock exchange, exhibition hall, stadium, market, ...) that are called *surrounding sequences*. The most represented type of place (market) totals 797 seconds of recordings of four various types of market. For every four recording the last 75 seconds, are kept for the training of the *normal* class. The rest of the environmental database is used for building of test database (see 3.2)

3.2. The experimental protocol

The test database results from a mix between the shots and *surrounding sequences*. A shot occurs for each sequence at a random moment with various local Signal to Noise Ratio (SNR). The SNR is computed for the part of the surrounding sequence where the shot is inserted and data are previously normalized before mixing. Each test sequence is 30 seconds long and is randomly chosen among test part of market surrounding sequences. For each SNR (from 20 to 5 dB) 134 sequences totalling about 67 minutes are generated for the test corresponding to the available 134 shots. Such mixed test sequences provide a simulation of abnormal situation in a public place as close as possible to the reality (in the case of gun shot occurrence). Despite their artificial nature these sequences allow us to control the SNR and therefore to test the system noise robustness but also to have a ground truth annotation of the test files (i.e. an exact localisation of all shot events in the *surrounding sequences*).

For different SNR conditions, the labels given by our automatic event detection system are compared to the ground truth annotation. The overall results are given by computing a false rejection (FR) ratio and a false detection (FD) ratio which are defined as follows :

$$FR = \frac{\textit{number of failed detections}}{\textit{number of events to detect}}$$

$$FD = \frac{\textit{number of false detected windows}}{\textit{number of windows}}$$

We use *leave one shot out* cross validation method for the training of the *shot* class : every shot to be detected in the test database is removed from the training database during the training step for each test sequence.

4. EXPERIMENTS

4.1. First experiment : noisy training database versus clean training database

This first experiment aims to better understand the effects of the noise level of the shot training database. For the *shot* class, a database of shots mixed with surrounding sequences segments is generated from 134 initial shots. Shots are inserted for SNR going from 20 to 5 dB with 5dB step. Figure 3 provides results of shot detections for each SNR level of a group of test sequences with varying SNR levels of the training databases.

As expected results rapidly degrade when the SNR condition of the test sequences decreases, or in other words when the shot energy decreases compared to the *surrounding sequences* energy. In particular clean shot (training) databases provide insufficient results in terms of false rejection for the noisiest test sequences. However, it can also be seen that the use of too noisy shot training databases triggers off a considerable increase of false detection rate which, in the worst case, is reaching 43% (5 dB SNR training database and 5 dB SNR test sequences condition).

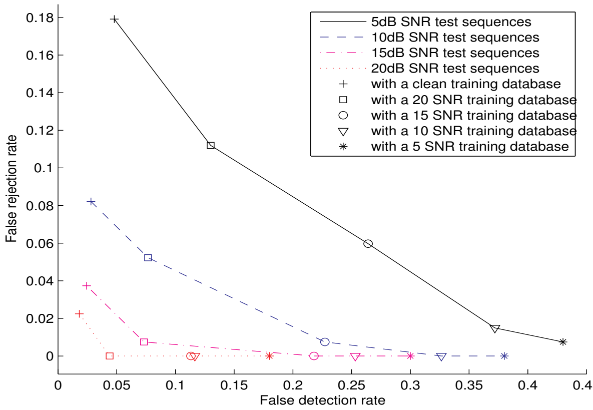

Fig. 3. *False rejection rate as a function of false detection rate for various SNR training database and test sequences. Each line corresponds to the performance of the system for a given SNR of the test sequences.*

This experiment illustrates the necessary trade-off between false rejection and false detection when choosing the appropriate SNR level between shot and *surrounding sequences* for the training database. For surveillance application, it is particularly important to keep the false rejection rate as low as possible and it appears that acceptable results are obtained in all test conditions with a 20 dB SNR training database (false rejection less than 11 % and false acceptation less than 15 %).

4.2. Second experiment : 4-classes hierarchical classification versus two-broad classes classification

In the previous experiments, only two classes have been considered: *shot* and *normal*. But the *shot* class is defined by different types of weapon reported in table 1. They all have a specific acoustic signature implying that the more different the signatures are the worst the *shot* model should be due to acoustic confusion. In order to reduce confusion and finally improve our detection system performance, it is reasonable to think that more specific models could be built. The basic idea would be to split the shot data into sub-classes gathering a sufficiently high number of training items that are acoustically close. Due to the limited size of our database, it is not possible to build a specific model for each weapon class, and we therefore aim merging the classes that are close with respect to a given distance. A convenient way to represent distance between every weapon subclasses is the hierarchical classification ([9]). To match with our problem and bring correlated subclasses closer, we choose anti-correlation factor (1-r Pearson) as aggregation distance. Acoustic measures of each shot were represented by their mean and standard deviation on each analysis windows. Results of the classification are presented in figure 4. Topologic distance between two subclasses is equivalent to the anti-correlation value (1-r). We can observe that pistol (P) and rifle (R) are very closed, just as grenade (G) and cannon (C). That means that pistol and rifle acoustic values are more correlated than pistol and grenade ones for example. One single subclass seems to be more isolated: the submachine gun (S), eventhough it is closer to the set (P+R). The three subclasses (P+R, S and G+C) represent the best trade-off between independence (i.e.: distance to the other subclasses) and future model quality (i.e.: sufficient number of members for training each subclass and maximum number of subclasses).

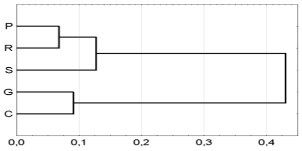

Fig. 4. *Five weapon subclasses hierarchical classification tree. The X-axis represents the aggregation distance between subclasses marked out on the Y-axis.*

Our second experiment then consists on using a 4 classes (3 weapon classes and the *normal* class) classification system and on assessing the performance of this hierarchical approach compared to the previous binary classifier. For each decision window the a posteriori probability score corresponding to the three weapon classes is computed and compared to the a posteriori probability score of *normal* class. Classification shot/normal is then performed using the following decision rule applied over each classification pair P+R/normal, S/normal, G+C/normal: shot classification is decided from the moment that the decision window is not classed *normal* by one of the three classification pairs. Other decision rules such as the majority vote of the three classification pairs could also be chosen but lead to higher false rejection rate.

Figure 5 shows the great enhancement provided by the hierarchical approach. The false rejection rate falls from 18% to about 10% when using subclasses. In the same time false detection rate relatively increases but stays sufficiently low, less than 5% even for noisy test conditions.

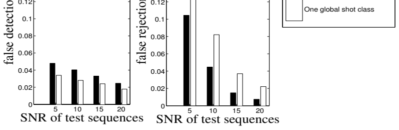

Fig. 5. *False detection rate and false rejection rate with a global shot class and with three shot subclasses.*

5. CONCLUSION AND FURTHER WORKS

In this paper, a robust audio-based shot detection system was introduced. This system represents an essential building block of a complete multimedia surveillance system. It is based on a binary classifier (shot/normal classification) and several experiments were conducted in order to reduce the false rejection and false detection rates. We show that the noise level of the training database has a significant impact on the performance of the system which allows to select the most appropriate noise level of the training database for a targeted false rejection rate. The performance of the system was also significantly improved by considering a hierarchical approach. Future work will be dedicated to the extension of the current system to different types of acoustic events that occur in abnormal situations such as shouts, cries or manifestation of fear.

6. REFERENCES

[1] R. Cai, L. Lu, H.-J. Zhang and L.-H. Cai, ”Highlight Sound Effects Detection in Audio Stream”. *Proceedings of IEEE International Conference on Multimedia Expo.*, 2003.

[2] C. Clavel, I. Vasilescu, L. Devillers and T. Ehrette, ”Fiction Database for Emotion detection in Abnormal situation,” *International Conference on Speech and Language Processing*,2004.

[3] C. Fraley, A.E Raftery ”How many clusters? Which clustering method? Answers via model based cluster analysis,” Technical Report 329, University of Washington, Department of statistics, 1998.

[4] O. Gillet and G. Richard, ”Automatic transcription of Drum sequences using audiovisual features,” *Proceddings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2005.

[5] Denis Mercier, ”Sound Library,” *Audivis Distribution, 1989.*

[6] M. Markou and S. Singh, ”Novelty detection: a review,” Signal Process., Vol.83, number 12, 2003.

[7] T. K. Moon, ”The Expectation-Maximization algorithm,” *IEEE Signal Processing Magazine*, 1996.

[8] S. Pfeiffer, S. Fischer and W. Effelsberg, ”Automatic Audio Content Analysis,” *Proceddings of 4th ACM international conference on Multimedia*, 1996.

[9] Tryon, R. C., ”Cluster Analysis,” *Ann Arbor, MI: Edwards Brother*, 1939.