Audio Events Detection For An Audio-based Surveillance System

Chloe Clavel
Thales Research and Technology France
DAS-HIT
Orsay, FRANCE zipcode

Thibaut Ehrette
Thales Research and Technology France
DAS/HIT
Orsay, FRANCE zipcode

Abstract

The present research focuses on the development of an audio-based surveillance system. We are interested in detecting events occurring in abnormal situations and particularly in detecting "shots". This study presents a two-class classification system to detect shots in an audio stream. We first use for the training step of the class "shot" a "clean" database. The classification system provides insufficient performances in situations where the surrounding noise is very present compared to the shots. To handle this problem, we elaborated a test protocol to evaluate the performances of the system in various noise conditions. We use more appropriate training databases for the class "shot" in which surrounding noise is mixed with the shots with varying noise signal ratio. The performance increased considerably.

1. Introduction

Audio events detection represents an interesting topic for a wide set of potential applications, such as indexation system, and survey application systems (homeland security). Classical security systems are only based on visual cues detection without any recording of audio signal. Our global aim is to use acoustic cues to detect and analyse abnormal situations. We define an abnormal situation as an unplanned event, consequent upon a human action, present or imminent or natural disaster, which implies human life threatening and requires prompt action to protect life or to limit its damages. Among the abnormal situations we can mention natural damages such as fires, earthquakes, flood etc, physical or psychological threatening and aggression against human beings (kidnapping, hostages etc). We assume that in these situations a lot of charistical events, such as shots, cries occur. The present research focuses here on the automatic detection of these charistical events, which will be a module of the aimed audio based surveillance system. The potential applications are dealing with security of public places, for example, bank or subway surveillance.

In the domain of event detection most of works deal with audio events classification with an indexation problematic (citer papiers) and only a few work are considering event detection in an audio stream (citer Cai). In this paper we handle event detection in an audio stream for an audio-based video surveillance system. ??In an indexation problematic the shot detection in a movie for example don’t imply a large context training database.?? This new problematic implies to consider two major quality criterions: precision and noise robustness. Indeed an audiovideo surveillance system should be suitable for the kind of place it has to survey, from usually quiet places, such as bank to more noisy places, such as stadium. We consequently focus here on the evaluation of our system in different noise to signal ratio situations, choosing more appropriate and context independant database for the training. ??dper la plupart utilise des bases de donnes tres proches des donnes tests mais notre systme doit pouvoir detecter tous les types de coups de feu en diffrents environnements?? ?? hypothese suffisament de micros pour les coups de feu soient toujours assez proches?? ??Methods used for classification such as HMM, SVM, dper but here we just focus on the influence of the training database on the performance. Influence des bases de donnes propres ou non, influence de l’endure de la base de donnes d’apprentissage de sa représentativité et de son adquation avec l’application faire des classes spécifiques chaque type d’armes ou faire une classe pour les armes en gneral. In the following sections we describe the detection system and the databases used to test it (section 1) and the test protocol (section 2) and its performance for different noisy situations (section 3). Finally, in section 4 we present conclusions and further work projects.

1.1. Previous Work

There are various bibliographic and citation schemes available in LaTeX, but we choose to use the simplest one in this example. Throughout I may cite references of the form [1] or [2], and LaTeX will keep track of numbering. The numbers are based on the order you place them in the bibliography, not the order they appear in the text. They should (I believe) be in alphabetical order. LaTeX will put square brackets about the number within the text of your paper. For
those of you new to \LaTeX, you may have to run the latex process twice to allow all references to be resolved. You will get a warning about a missing .aux file. Just rerun latex and it will be ok.

2. System framework

The proposed final audio-based automatic surveillance system illustrated in fig? consists of different modules providing informations from different natures that will be merged by an information fusion system for situation analysis. Audio module hence used vocal and non vocal manifestations of abnormal situations and focus both on the emotional content and on charistical events, such as cries or shots. The detection system described here deals with shots detection in an audio stream. CITER ARTICLE ICSLP

2.1. Resources and materials

Natural corpora with charistical events in abnormal situations for surveillance applications are not available because of the confidential character of the data. Therefore we build our database from sound library CD (CD of Radio France “ Sound Library ”) containing public places recordings in normal situations and shots recordings. The training and test corpus are both constructed to study the system robustness to noise. CITER SONOTHEQUE The training database

For each of both classes (“shot” and “others”), a training database is built. An ideal database should be constituted for the “shot” class by a large amount of different shots recorded in the same place, the place to survey and for the “other” class by recordings of the public place in normal condition. In order to be as close as possible from our application, the training database is built as follow :

- The shots are extracted from CD of library of sounds and 134 shots fired by various weapons and recorded in different conditions (indoor and outdoor), without any environment noise. DECREIRE LES ARMES ET LES LIEUX For the class ”shot”, a database of shots mixed in the noise is generated from 134 initial shots. The used noises come from recordings in public places and shots are inserted for Noise (environment) Signal (shot) Ratio (NSR) going from 20 to 0 dB. distinguer clean database et noise database (faire deux items)

- Sequences used for the ”other” class are extracted from the same CD and contain recordings of public places such as airports, stations, stadiums, theaters. DECREIRE NOMBRE DE LIEUX DIFFERENTS

The test database

Authentic shots are randomly inserted in sequences extracted from the CD of Radio France and containing sound recordings in various public places, as airports, stations, stadiuims, theaters. We will call these sequences “surrounding sequences”. Various noise signal ratio are presented to test the robustness of the system.

For each noise signal ratio (from 0 to 20 dB) a total of 134 sequences are generated for the test corresponding to the 134 shots.

2.2. Shot detection in an audio stream

We develop here a system of classification to detect the presence of shots for each frame in audio sequences. The classification stage is based on two classes: the class ”shot” and the class ”others” which is composed by surrounding noises. The audio stream is segmented in windows of half a second with a 50% overlap (that is a quarter of second). Sound features. For each window audio signal is represented by a 12-dimensional feature vector consituted by 12 order MFCCs, and delta MFCCs (to model the signal time variation).

voir pour d’autres caractristiques??

The training step The Gaussian Mixture Models (GMM) are used to learn models for each class. Each parameter is modeled by the characteristics of each Gaussian distribution of the class: the likelihood $\pi_m$. traduire ici l’annexe GMMs et voir article medical ... phase d’initialisation, iteration par l’algorithme EM nombre de gaussiennes (celui qui donne le meilleur score) The detection step

In the test step a likelihood (ou MAP???) score associated with each window is calculated for each class. The average likelihood is then calculated on a set of 12 windows with a 50% overlap between the frame. We classify the frame into the class that has the maximum likelihood score. Results are supplied in the form of a signal of detection corresponding at each frame and equal to 0 or 1 if a shot was detected.

?? What vocal manifestations concerns the emotional content is about to be taken into account with the emotion detection module and the event detection module for cries detection.?? ??First step, the heart of audio processing algorithms consists of finding the audio signal representation-space to make the distinction of sounds we want to detect more accurate. This choice of the representation space is highly problem dependant and existing works used always the same set of features for a general audio classification. We focus here on the selection of relevant features for a 2 class shot /others discrimination.

3. Evaluation

We want her to evaluate the performance of our detection system in different noise conditions. As we control the
building of the database the labels are automatically generated: the label is 1 if the frame contains a portion of shot and 0 otherwise. Labels can then be compared to the detection signal and a false rejection ratio and a false acceptance ratio is computed for each test.

3.1. First experiment: training with clean data

At first, the used training data are the "clean" data that means that the shot are not noisy. The tests are made first on the clean data (protocol Leave one out) then for different NSR going of 20dB to 0 dB with a 5 dB step. The results obtained are described in following figure.

Table 1: FR (false rejection) and FA (false acceptance) ratio for different noisy conditions with clean training database.

<table>
<thead>
<tr>
<th>NSR test data</th>
<th>clean</th>
<th>20dB</th>
<th>15dB</th>
<th>10dB</th>
<th>5dB</th>
<th>0dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR in %</td>
<td>3.75</td>
<td>4.51</td>
<td>5.26</td>
<td>12.03</td>
<td>17.29</td>
<td>27.82</td>
</tr>
<tr>
<td>FA in %</td>
<td>0.17</td>
<td>0.7</td>
<td>0.86</td>
<td>1.02</td>
<td>1.11</td>
<td>1.16</td>
</tr>
</tbody>
</table>

Calculer le taux d’erreur (ERR) For weak RSB performances are, as expected, really unsufficient : the "clean" database is not suitable for noisy case. We are going to test the performances of the system if we adapt the training database to the noise signal ratio of the test database.

3.2. Second experiment: training with noisy database

During this test the training data are those with the same NSR of the test database. In other words, during the test of the sequences where the inserted shot is 5 dB superior to the surrounding noise we use the models extracted from the training database containing shots mixed in the surrounding noise in a 5dB ratio. Results are reporting in the following table:

Table 2: FR (false rejection) and FA (false acceptance) ratio for different noisy conditions with noisy training database.

<table>
<thead>
<tr>
<th>NSR test data</th>
<th>20dB</th>
<th>15dB</th>
<th>10dB</th>
<th>5dB</th>
<th>0dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>false rejection in %</td>
<td>2.26</td>
<td>3.76</td>
<td>5.26</td>
<td>5.26</td>
<td>7.52</td>
</tr>
<tr>
<td>false acceptance in %</td>
<td>2.39</td>
<td>2.54</td>
<td>3.23</td>
<td>3.55</td>
<td>5.29</td>
</tr>
</tbody>
</table>

The false acceptance rate appreciably increased, compared to the previous test, but stay enough low. However the false rejection rate falls of 20 % for a RSB equal to 0, what for a problem of video surveillance is major. Indeed a false rejection has heavier consequences in our surveillance application than a false acceptance. This test thus puts in evidence the interest to adapt the database of learning at the surrounding noise level. For that purpose, a noise level estimation is necessary to select the best adapted training database.

trouver le meilleur compromis concernant la base de donne. choix du RSB le mieux adapt pour la base d’apprentissage.

3.3. A class for each weapon

4. Further Works

We assume that for a surveillance application charistical events of abnormal situation don’t occur frequently. This module is thus divided in two stages: a burst detection stage which allows us to eliminate quiet windows. this one is based on an entropy energy criterion and ZCR? and sigma? Windows under a threshold for these criterions are eliminated and a classification stage which allows us to indentify burst windows as shot. Test 4 : apprentissage sur base RSB, estimation du niveau de bruit et slection de mod-leLors de ce test, la base de donnees de test contient des squences qui cette fois ne呈现ent pas un RSB donn. Le systme devra alors estim le niveau de bruit afin de sletionner le modle issu de la base d’apprentissage ayant le RSB le plus proche de la squence tester.

Test 5 : apprentissage sur base ” propre ” + dbruitageUne autre ide consiste utilise la base de donnees d’apprentissage constitue de la base propre et, afin d’amliorer les performances, d’estimer le niveau de bruit dans le but de dbruter la squence avant de dtecter le coups de feu. Ainsi, le RSB sera plus lev au moment de la dtaction, qui devrait alors tre plus facile.

5. Summary and Conclusions

This template will get you through the minimum article, i.e., with no figures or equations. To include those, please refer to your LaTeX manual and the IEEE publications guidelines. However, for a vision conference you will probably want the following equation somewhere:

\[ g(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-x^2/2\sigma^2} \]

Good Luck!

Acknowledgments

This is how to do an unnumbered subsection. For submission, there should be no acknowledgments as this could lead to identification of the author.

References