ON ACOUSTIC SURVEILLANCE OF HAZARDOUS SITUATIONS

[†]Stavros Ntalampiras, [‡]Ilyas Potamitis, [†]Nikos Fakotakis

[†]Department of Electrical and Computer Engineering, University of Patras, Greece, sntalampiras@upatras.gr,
[‡]Department of Music Technology and Acoustics, Technological Educational Institute of Crete, potamitis@stef.teicrete.gr

ABSTRACT

The present study presents a practical methodology for automatic space monitoring based solely on the perceived acoustic information. We consider the case where atypical situations such as screams, explosions and gunshots take place in a metro station environment. Our approach is based on a two stage recognition schema, each one exploiting HMMs for approximating the density function of the corresponding sound class. The main objective is to detect abnormal events that take place in a noisy environment. A thorough evaluation procedure is carried out under different SNR conditions and we report high detection rates with respect to false alarm and miss probabilities rates.

Index Terms— acoustic surveillance, content based audio recognition, MPEG-7

1. INTRODUCTION

Research in the area of automatic surveillance systems is mainly focused on detecting abnormal events based on the acquired video information [1]. Current implementations typically consist of a large number of cameras distributed in an area and connected to a central control room. While this kind of analysis provides valuable information we concentrate on detecting atypical events by exploiting only the acoustic modality. This approach offers several advantages such as: a) low computational needs, b) the illumination conditions of the space to be monitored and possible occlusion do not have an immediate affect on sound. Previous approaches on the subject of acoustic monitoring include cases such as in [2] where a gunshot detection system is presented based on features derived from the time-frequency domain and GMM classifier. They use different SNRs during the training phase for achieving 10% and 5% false rejection and false detection rate respectively. In [3] they present an emotional recognition scheme for public safety. Their main objective is fear vs. neutral classification and by using different models for voiced and unvoiced speech they reach 30% error rate. In [4] they report on building a parallel classification system based on GMMs for discrimination of ambient noise, scream and gunshot sounds. After a feature selection algorithm they result in 90% precision and 8% false rejection rate. Last but not least, an audio-based surveillance system in a typical office environment is described in [5]. The background noise model is continuously updated for serving *interesting* event detection while both supervised and k-means data clustering are inspected. In [6] audio data recorded using simultaneously 4 microphones are classified with two different approaches - GMM and SVM - for shout detection in a railway environment. The proposed implementation exploits PLP features combined with the SVM classifier.

The main goal of this paper is to efficiently characterize the acoustic environment in terms of threatening/non-threatening conditions while using a single microphone. The outcome of the system is to help/warn authorized personnel to take the appropriate actions for preventing crime and/or property damage. In order for such an implementation to be useful and practical it must offer very low false alarm rate while keeping detection accuracy as high as possible under noisy conditions. Our approach is basically motivated by the fact that sound provides information that is hard or impossible to obtain by any other means. On top of that, such a method comprises a low cost and relatively easy during setup, solution. In this article we concentrate on detecting atypical sound events (scream, gunshot and explosion) in a metro station environment. The current methodology is inspired by the work of Wilpon et al [7] regarding keyword spotting. We extend this idea to the field of keysound spotting, where screams, gunshots and explosions are considered as keysound effects. In our case the non-interesting/garbage model is the metro station soundscene which presents highly non-stationary properties (it includes horns, opening/closing doors, people talking in the background, train movement etc).

We have carried out extensive experimentation regarding the best set of features to be included in the feature extraction process. The final set is consisted of the well known Mel frequency cepstral coefficients augmented by a second group of parameters based on the MPEG-7 audio standard. Subsequently feature sequences are modeled by probability density functions represented by GMMs and HMMs.

The next of this paper is organized as follows: in section 2 a brief overview of the system is given along with the description of MFCC and MPEG-7 sets of parameters. Section 3 explains the experimental protocol that was used while our conclusions are reported in the last section.

2. SYSTEM OVERVIEW

Our system is designed as a two stage topology which was proven to provide better recognition rates than the single stage one. The incoming signal is first classified as normal (metro station environment) or abnormal (scream, gunshot or explosion) and in case it is decided to be abnormal the system proceeds into a second processing stage where the type of abnormality is identified. The proposed architecture comprises a fully probabilistic structure based on ergodic HMMs for describing each sound category.

2.1. Feature Extraction Analysis

In this section we comment on the groups of descriptors that were employed in order to train probabilistic models that represent the a priori knowledge we have about the sound classes. We make use of the Mel-scale filterbank because of its ability to lower the

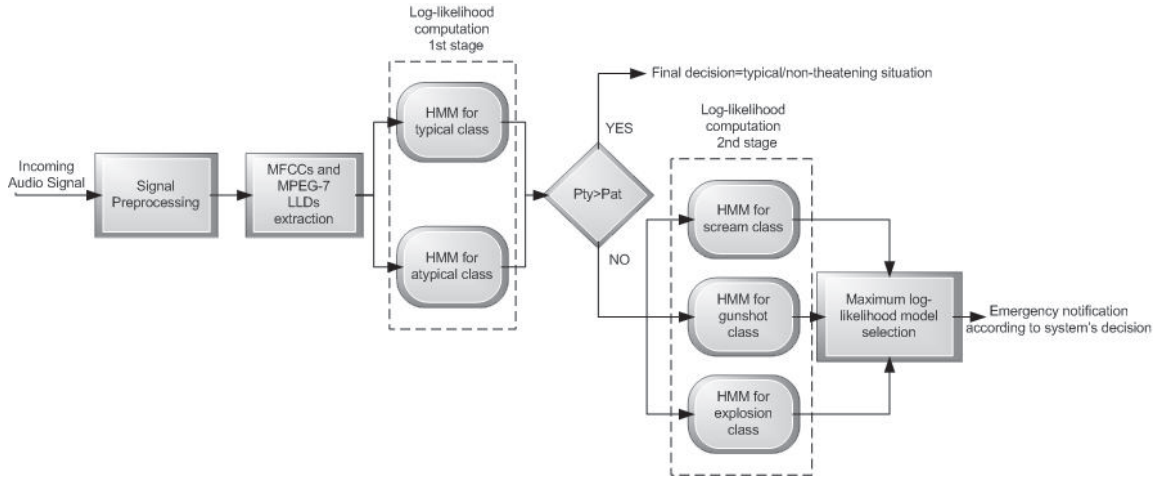

Fig. 1. Block diagram of the probabilistic based acoustic surveillance system.

dimensionality of the Fourier transformed vector. We also include the logarithmic portioning of the data, a process which mimics the natural frequency selectivity of the human middle ear to some extent. Secondly, the MPEG-7 protocol is employed since it currently constitutes the state of the art mechanism for automated content-based generic audio recognition. We adopt the next four low-level descriptors: *Waveform Min, Waveform Max, Audio Fundamental Frequency and Audio Spectrum Flatness* (ASF). The above mentioned feature sets were evaluated both separately and combined using different values of parameters. They were chosen because they capture different aspects of the information provided by the MFCC features. MFCCs are a Mel-scaled projection of the log spectra while ASF constitutes a higher level description of the audio waveform indicating how flat a given signal is. It should be noted that the incorporation of a set of parameters based on Teager energy operator (critical band based TEO autocorrelation envelope area) proposed in [8] was also tested. They are used for stress classification but their combination with the above mentioned features did not provide improved results.

In this work we are dealing with keysound spotting under subway environment· acoustic signals characterized by a long duration need to be processed for the purpose of atypical sound event detection. Thus the instantaneous value of each feature is computed over a larger frame size than the one commonly used in speech recognition (namely 30ms). After several experiments it was decided that all sound samples should be cut into frames of 200ms with 50% overlap. Mean removal and variance scaling are also applied. A short analysis of the feature extraction processes follows.

Mel-Frequency Cepstral Coefficients

For MFCC's derivation we compute the power of the short time Fourier transform for every frame and pass them through a triangular Mel scale filterbank so that signal components which play an important role to human perception are emphasized. Afterwards the data are compressed and decorrelated using the logarithmic scale and the discrete cosine transform respectively. Thirteen coefficients are kept (including the 0-th coefficient which reflects upon the energy of the signal) and in combination with their respective derivatives a twenty six-dimension vector is formed.

MPEG-7 Audio Protocol Descriptors

Provide a general framework for efficient audio management. Furthermore, it includes a group of fundamental descriptors and description schemes for indexing and retrieval of audio data. We employed three audio descriptors namely: Spectrum Flatness (ASF), Waveform (AWF) and Fundamental Frequency (AFF).

2.2. Classification Schemas

We employed Gaussian mixture models and Hidden Markov models with two different topologies (left-right and fully-connected). Subsequently the previously created models are used for computing a degree of resemblance (e.g. log-likelihood) between each model and an unknown input signal. This type of score is compared against the rest and the final decision is made with a simple maximum log-likelihood determination. Torch implementation (provided at http://www.torch.ch) of GMM and HMM, written in C++ was used during the whole process. The maximum number of K-means iterations for initialization was 50 while both the EM and Baum-Welch algorithms had and upper limit of 25 iterations with a threshold of 0.001 between subsequent iterations.

3. EXPERIMENTAL SET-UP

Natural corpora with extreme emotional manifestation and atypical sounds events for surveillance applications are not publicly available because of the private character of the data, their scarcity and unpredictability. Our corpus consists of audio acquired from professional sound effects collections. These kinds of collections comprise an enormous source of high quality recordings used by the movie industry. An important detail, which is not widely known, is that the audio in a movie is not the exact audio recorded at a scene but it is processed and in most cases added separately to the audio stream later. Therefore, there is a vast corpus of real vocal and non-vocal audio available for the construction of trained probabilistic classification models. Sound samples from the following compilations: (i) BBC Sound Effects Library, (ii) Sound Ideas Series 6000, (iii) Sound Ideas: the art of Foley, (iv) Best Service Studio Box Sound Effects and (v) sound effects from internet search constructed the final corpus.

Category	Number of recordings	Duration (sec)
Explosion	131	13.77
Gunshot	187	32.94
Scream	270	4.04
Subway	32	44.88
Total	620	23.9

Table 1. The parts of the final corpus

3.1. Model Construction and Recognition Accuracy

The data belonging to each class were splitted into 75% for training and 25% for testing in a random way. A fully-connected HMM was built for each category to capture this property while testing consists of a simple comparison of log-likelihoods. Due to the system architecture we first constructed two kinds of models: typical (metro station soundscape) and atypical (including explosion, gunshot and scream). After extensive experimentations we used 6 states each one modeled with 19 Gaussian components and 98.87% average recognition rate was achieved. Regarding the second stage three HMMs were built for describing each atypical situation. The same parameters provided the highest average recognition accuracy - 93.05% - and the corresponding confusion matrix is tabulated in Table 2.

Presented \ Responded	Explosion	Gunshot	Scream
Explosion	**86.06**	11.62	2.32
Gunshot	1.72	**93.10**	5.17
Scream	0	0	**100**

Table 2. Confusion Matrix for three Atypical Sound Events (%)

We observe that scream sound events are recognized with the best accuracy. This is due to the different spectral/energy distribution that scream vocal reactions exhibit when compared with the rest of the classes. The lowest accuracy is obtained regarding to explosion sound events, of which 11.62% is misclassified as gunshots. Many of the errors occur because of the great variability among sound samples of the same category.

3.2. Atypical Sound Event Detection in a Metro Station

Emergency situations located in a metro station were created by merging abnormal sound events with subway recordings at different SNRs (from -5dB to 15dB with 5dB step). The proposed architecture was tested using Detection Error Tradeoff (DET). Two series of experiments were conducted dedicated to each stage of our implementation. The DET curves for both stages are depicted in Fig. 2 and Fig. 3 for stage 1 and stage 2 respectively. Figure 2 provides results of atypical event detection regarding to all three different sound events. The log-likelihood values of two statistical models (typical/atypical) were utilized for the DET curves creation. Results follow a rapid degradation when the SNR condition of the test signals decreases. Abnormal sounds are adequately detected even at extremely low SNR conditions. Average equal error rate (EER) regarding all types of events at -5dB SNR is 8.53% while its minimum value (best detection rate) is achieved in the gunshot class. The audio signals that are most vulnerable to background noise corruption are the explosion ones with 12.88% EER at -5dB SNR. For surveillance tasks an energy ratio of 0dB represents the real world conditions appropriately. The proposed framework demonstrates very good performance in the respective ratio, having EER of 4.8% and false alarm probability of 1.83% which is particularly important for this kind of applications.

Figure 3 illustrates system's capabilities regarding the detection of each atypical sound category alone merged with metro station recordings at different SNR levels. The misclassifications that occur at this processing level comprise errors that are of less importance in comparison to the previous ones. Here a threatening situation has been detected and the system tries to identify which type of abnormality is present while an authorized person has already been alerted in order to take the appropriate action. Thus, at this stage of recognition our main interest is to obtain very low miss probability and then try to have as low false alarm rate as possible. The output log-likelihoods obtained by the probabilistic models which describe each atypical sound class, were used during this phase. We can observe that gunshot events are detected with relatively low EERs across all SNR values in contrast with the two other kinds of atypical events which are detected with satisfying accuracy. As expected, miss detection probability falls as the SNR conditions increase from -5dB to 15dB. More precisely explosion sounds, corrupted by metro station environmental noise with -5dB ratio are detected with EER of 13.2%, gunshot sounds with 24.5% and scream sounds with 28.2%. Additionally, our implementation provides very good false alarm probability with a mean value of 6% among the three sound event categories with 0dB SNR conditions. The corresponding EERs achieved by the system as regards explosion, gunshot and scream detection are 8.54%, 24.5% and 21.1%.

4. CONCLUSIONS

In this work we presented and evaluated a two stage probabilistic framework for acoustic monitoring in a metro station environment. Its main aim is to identify on time the sensed situation and deliver the necessary warning messages to an authorized officer. The proposed methodology is practical, can operate in real-time and elaborates on three abnormal sound events corrupted by highly non-stationary metro station. The recognition results under a variety of background environmental noise are very encouraging.

5. ACKNOWLEDGMENTS

This work is under the EC FP 7th grant Prometheus 214901.

6. REFERENCES

[1] I. Haritaoglu, D. Harwood, and L. Davis, "W4: real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 809-830, 2000.

[2] C. Clavel, T. Ehrette, and G. Richard, "Event detection for an audio-based surveillance system," in *IEEE International Conference on Multimedia and Expo*, Amsterdam, July 2005.

[3] C. Clavel, I. Vasilescu, L. Devillers, G. Richard and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Communication*, Elsevier, pp. 487-503, 2008.

[4] L. Gerosa, G. Valenzise, F. Antonacci, M. Tagliasacchi and A. Sarti, "Scream and gunshot detection in noisy environments," in *EURASIP*, Poznan, Poland, September 2007.

[5] A. Harma, M.F. McKinney, J. Skowronek,, "Automatic surveillance of the acoustic activity in our living environment," in *IEEE International Conference on Multimedia and Expo*, 2005.
[6] J.-L. Rouas, J. Louradour and S. Ambellouis, "Audio Events Detection in Public Transport Vehicle," in *IEEE Intelligent Transportation Systems Conference*, Toronto, September 2006.
[7] J. G. Wilpon, L. R. Rabiner, C.-H. Lee and E. R. Goldman, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 1870-1878, November 1990.
[8] G. Zhoun, J. H. L. Hansen and J.F. Kaiser, "Nonlinear Feature Based Classification of Speech Under Stress", *IEEE Transactions on Speech and Audio Processing*, pp. 201-216, March 2001.

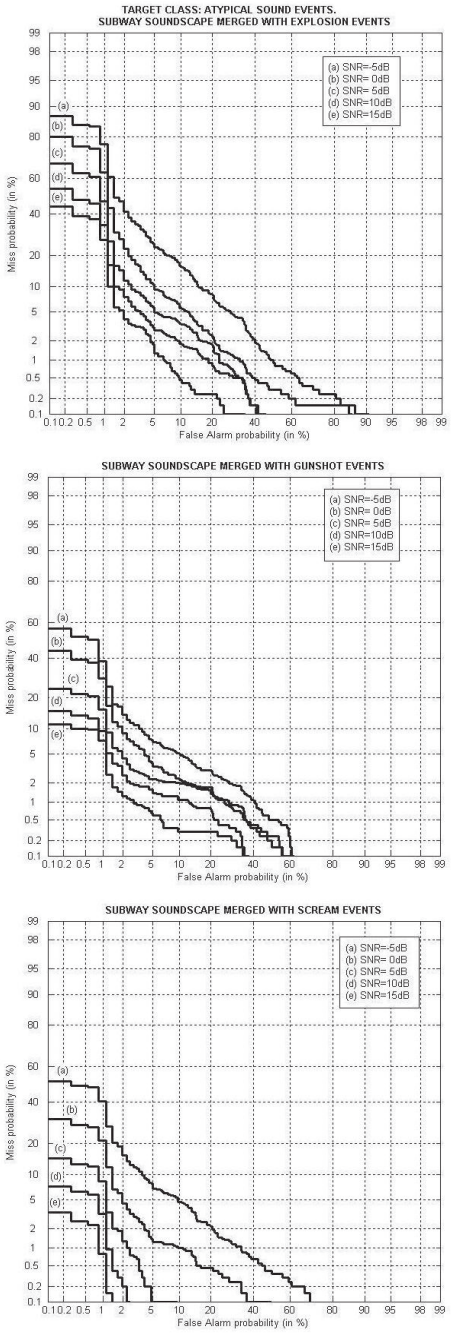

Fig. 2. 1st stage DET curves regarding to atypical sound events as the target class under different SNRs. Each sub-figure corresponds to results obtained by mixing the subway signal with one of the three atypical sound events.

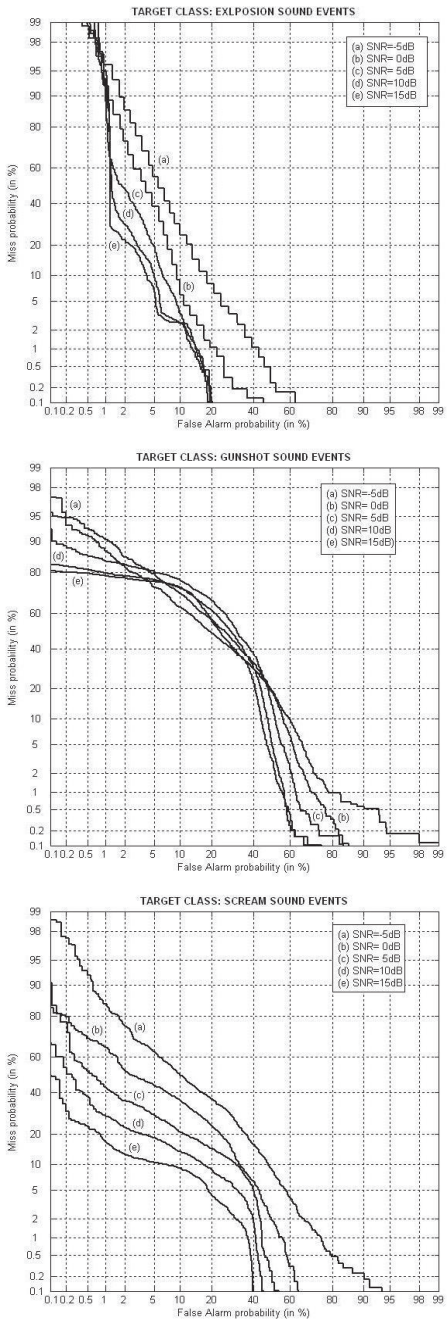

Fig. 3. 2nd stage DET curves each one corresponding to a specific atypical sound event as the target class under different SNRs. The sub-figures from up to bottom refer to results obtained for explosion, gunshot and scream audio categories respectively.