Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2009, Article ID 594103, 15 pages
doi:10.1155/2009/594103

Research Article

An Adaptive Framework for Acoustic Monitoring of Potential Hazards

Stavros Ntalampiras,[1] Ilyas Potamitis,[2] and Nikos Fakotakis[1]

[1] *Electrical and Computer Engineering Department, University of Patras, Rio-Patras 26500, Greece*
[2] *Department of Music Technology and Acoustics, Technological Educational Institute of Crete, Daskalaki-Perivolia, Crete 74100, Greece*

Correspondence should be addressed to Stavros Ntalampiras, sntalampiras@upatras.gr

Received 7 March 2009; Revised 19 June 2009; Accepted 3 August 2009

Recommended by Vijay Parsa

Robust recognition of general audio events constitutes a topic of intensive research in the signal processing community. This work presents an efficient methodology for acoustic surveillance of atypical situations which can find use under different acoustic backgrounds. The primary goal is the continuous acoustic monitoring of a scene for potentially hazardous events in order to help an authorized officer to take the appropriate actions towards preventing human loss and/or property damage. A probabilistic hierarchical scheme is designed based on Gaussian mixture models and state-of-the-art sound parameters selected through extensive experimentation. A feature of the proposed system is its model adaptation loop that provides adaptability to different sound environments. We report extensive experimental results including installation in a real environment and operational detection rates for three days of function on a 24 hour basis. Moreover, we adopt a reliable testing procedure that demonstrates high detection rates as regards average recognition, miss probability, and false alarm rates.

1. Introduction

Lately automatic systems which monitor human daily activities are becoming increasingly common [1, 2]. The aim of this work is to contribute to civil safety by proposing and automatic acoustic surveillance system that monitors public spaces for potentially hazardous situations. These hazardous events imply a threat to human life or property loss/damage (e.g., gunshot, explosion and human reaction to this kind of situation) and usually entail a strong acoustic emission. This work reports on a practical system that exploits solely the acoustic modality. This modality is cost-effective compared to other kind of sensors (e.g., infrared and visual cameras as well as laser scanners) and can be used in a stand-alone mode as a recognizer of acoustic events or in a fusion process that combines the likelihood of detected events along with complementary cues of other sensors.

The research area of acoustic surveillance in particular, has gained a lot of attention recently addressing various types of applications [3–10]. It is a branch of generalized sound recognition technology, namely computational auditory scene analysis. This particular domain tries to interpret the surrounding environment using the incoming audio, inspired by the respective property that humans exhibit in their everyday life quite effortless. Previous efforts in the abnormal sound event detection domain consider a wide range of audio features combined with various classification techniques. Previous research on the subject is far from concluding on a common framework as, for example, in the case of speech/speaker recognition where the classifier and the feature extraction process is more or less established (i.e., GMMs and HMMs as classifiers and variations of spectral features as input). The difficulty lies on the fact that (a) an atypical situation is not a well defined category (e.g., laughter versus cry versus scream), (b) there are many cases where there is a thin line between a typical and an atypical situation (e.g., gunshot versus explosion), and (c) the microphone can be located far from the source of the acoustic incident therefore, reverberation and acoustic events belonging to an

almost unrestricted range of classes may become the input to the microphone.

Previous approaches on the task of acoustic monitoring focus on different aspects of the classifier, the feature extraction process, the training data and the number of classes. In [3] the authors used an emotion recognition system for detecting fear-type emotional manifestations which take place during abnormal situations. The extracted features described prosody and audio quality in combination with spectral and cepstral parameters to train Gaussian mixture models separately for voiced and unvoiced audio parts. Their database was based on fiction movies and consisted of seven hours of recordings organized into 400 audiovisual sequences (SAFE corpus). The classification task concerned fear and neutral speech while they achieve 30% error rate. Valenzise et al. [4] presented a surveillance system for gunshot and scream detection and localization in a public square. Forty-nine features were computed in total and given as an input to a hybrid filter/wrapper selection method. Its output was used to build two parallel GMMs for identifying screams from noise and gunshots from noise. Data were drawn out from movie sound tracks, internet repositories and people shouting at a microphone while the noise samples were captured in a public square of Milan. The resulted precision was 93% at a false rejection rate of 5% when the SNR condition is 10 dB. An interesting application, crime detection inside elevators was described in [5]. Their approach relied on time-series analysis and signal segmentation. By automatic clustering of audio data, consistent patterns were discovered and data were collected for training a GMM for each one of the eight classes using low-level features. The data set contained recordings of suspicious activities in elevators and some event free clips while they reported detection of all the suspicious activities without any misses. A gunshot detection method under noisy environments was explained in [6]. Their corpus consisted of data which were artificially created from a set of multiple public places and gunshot sound events extracted from the national French public radio. Widely used features were employed, including MFCC for constructing two GMMs with respect to gunshot and normal class using data of various SNR levels. The benefits of supervised and unsupervised clustering for acoustic surveillance in a typical office environment were described in [7]. Their system was based on a continuously updated background noise spectrum profile which served interesting event detection. Both k-means and manual selection of cluster centers methods were used for clustering audio files that were captured in a standard office room for a period of 48 days. The detection relied on two alternative criteria each of which put a threshold onto two quantities which were designed to detect loud onset and transients in the environment. In [8] the issue of detection of audio events in public transport vehicles was addressed by utilizing both a generative and a discriminative method. The audio data were recorded using 4 microphones during four different scenarios which included fight scenes, a violent robbery scene and scenes of bag or mobile snatching. They utilized GMM and SVM while their feature set was formed from the first 12 MFCC, energy, derivatives and accelerations. Vacher et al. [9] presented a framework for sound detection and classification for medical telesurvey. Their corpus consisted of recordings made in the CLIPS laboratory, files of the "Sound Scene Database in Real Acoustical Environment" (RCWP, Japan). They used wavelet based cepstral coefficients to train GMMs for eight sound classes while their system was evaluated under different SNR conditions. Last but not least, a hierarchical classification scheme which identified normal from excited sound events was described in [10]. The authors used four audio features for training GMMs, each one associated with one node of the classification tree. The audio was recorded for around two hours in the real environment (office corridor) and included talk, shout, knock, and footsteps.

To our point of view, previous approaches focus on different aspects of classification of general audio events trying to optimize a specific part of the problem and are mostly laboratory based experiments, that is, prerecorded well defined classes are presented to a classification algorithm. Our approach targets to a practical system that operates in a real space on a 24/7 basis and does not use isolated events but rather a continuous of acoustic events as in the case of real life. Our emphasis is on making an integrated acoustic surveillance *system* that is self-adaptive to different acoustic environments (e.g., metro station, urban etc). This paper is focused on abnormal situations characterized by specific sound events—screams, explosions, and gunshots—which take place in (a) metro station, (b) urban environment and (c) a setting suited for military applications. Furthermore special care has been taken so that our dataset is thorough and concise after combining several well documented professional sound effect collections which contain audio of high quality. We provide a thorough investigation of features for the detection of three types of atypical sound events and a self-adaptive functionality that retrains its models during its activity.

Table 1 provides a comparison of the aspects which are considered in this work and several existing approaches. Screamed speech is a vocal sound directly related to human negative emotions (e.g., fear, pain, anger) and its detection can help minimizing the cost or even avoiding threatening circumstances. Our methodology was tested in a real internal space where both typical and atypical situations were played at random for three consecutive days through loudspeakers while a computer analyzed the emitted sound every 2 seconds. A video which demonstrates the proposed acoustic surveillance system can be downloaded from http://www.wcl.ece.upatras.gr/dalas/doku.php?id=demos. The feature extraction part is Matlab© based while the classification part is written in C++. The system is practically real time since it reports an event with an average delay of 2 seconds. For the type of the application we are investigating this delay is not critical and, moreover, the processing delay can be further reduced since the code is not optimized as regards the feature extraction process.

The rest of this paper is organized as follows: in Section 2 a complete overview of the system is given along with a short description of all sets of sound parameters. Sections 3 and 4 explain the experimental procedures and report detailed

Table 1: Research approaches on the task of acoustic surveillance.

Reference	Atypical sound classes	Model adaptation	Environments	Classifier	Features	Database
Proposed approach	Scream, gunshot and explosion	*MAP* adaptation of GMMs	Metro station, urban and military setting	GMM	MFCC, MPEG-7, CB-TEO, Intonation	Large audio corpora from professional sound effects collections
Clavel et al. [3]	Fear-type emotions	—	—	GMM	Prosody, audio quality, spectral, cepstral	SAFE corpus
Valenzise et al. [4]	Scream and gunshot	—	Public square	GMM	Temporal, spectral, cepstral, correlation	Movie soundtracks, internet and people shouts
Radhakrishnan and Divakaran [5]	Banging and non-neutral speech	—	Elevator	GMM	MFCC	Elevator recordings
Clavel et al. [6]	Gunshot	—	Public space	GMM	MFCC, spectral moments	CDs for the national French public radio
Harma et al. [7]	Interesting events in an office	Noise spectrum update	Office	Threshold, clustering	Temporal, spectral	Recordings from an office room
Rouas et al. [8]	Shout	Adaptive threshold for sound activity detection	Railway	GMM, SVM	Energy, MFCC	Recorded during 4 scenarios
Vacher et al. [9]	Scream and glass break	—	Apartment	GMM	Wavelet based cepstral coefficients	Laboratory recordings and RCWP
Atrey et al. [10]	Shout	—	Office corridor	GMM	ZCR, LPC, LPCC, LFCC	Recorded in office corridor

detection results obtained at different SNR levels while our conclusions are drawn in the last section.

2. General Architecture of the System

The main goal of our system is to detect human vocal reactions (i.e., screams, expressions of pain) and non-vocal atypical events associated with hazardous situations (gunshot and explosions). To this end, the structure that was designed has the form depicted in Figure 1. We employ a hierarchy of three discrete subsequent stages, where each stage depends on the previous one, for processing the incoming audio sequence before its class is determined. In brief, after preprocessing and feature extraction, the sound is classified as vocalic (normal or screamed speech) or non-vocalic (background environment, gunshot or explosion) event. Based on this decision a different path is chosen to further characterize the audio signal. In case it is found to be a vocalic event, a different set of descriptors is computed and the sound is classified as normal or screamed speech. In the case of nonvocalic event, an additional feature extraction phase follows and the signal is classified as non-threatening background environment or as an atypical sound event, while during the third stage the systems proceeds into specifying the type of the hazardous situation. Normal situations encompass all contexts that do not demonstrate life and/or property hazard.

2.1. Feature Extraction Analysis. In this paragraph we explain the-low level attributes that are extracted from the audio signals for constructing statistical models which represent the basic *a priori* knowledge we have about the audio categories. They were chosen because they capture diverse aspects of the audio structure. Furthermore they are not too sensitive to the SNR conditions, like energy or loudness. The first stage discriminates between vocalic and non-vocalic events, thus we used Mel frequency cepstral coefficients (MFCC) which provide a gross description of how the energy is distributed on frequencies. Subsequently, vocalic events are characterized upon their abnormality using critical band based Teager energy operator (TEO) autocorrelation envelope area [11], pitch and harmonic to noise ratio (HNR). They are indicative of the variations that intonation exhibits when in comes to atypical speech. Non-vocalic events are processed by computing Waveform Min, Waveform Max, Audio Fundamental Frequency and Audio Spectrum Flatness (ASF) as defined by the MPEG-7 audio standards [12], which capture the time-domain shape, periodicity and flatness of the spectrum in different bands.

Since we try to spot specific sound events we experimented with larger frame sizes than the ones commonly used (10–30 milliseconds) for speech/speaker recognition and based on the highest recognition rate after extensive experimentation we concluded to frames of 200 milliseconds with 75% overlap. Next, we briefly explain the feature extraction processes.

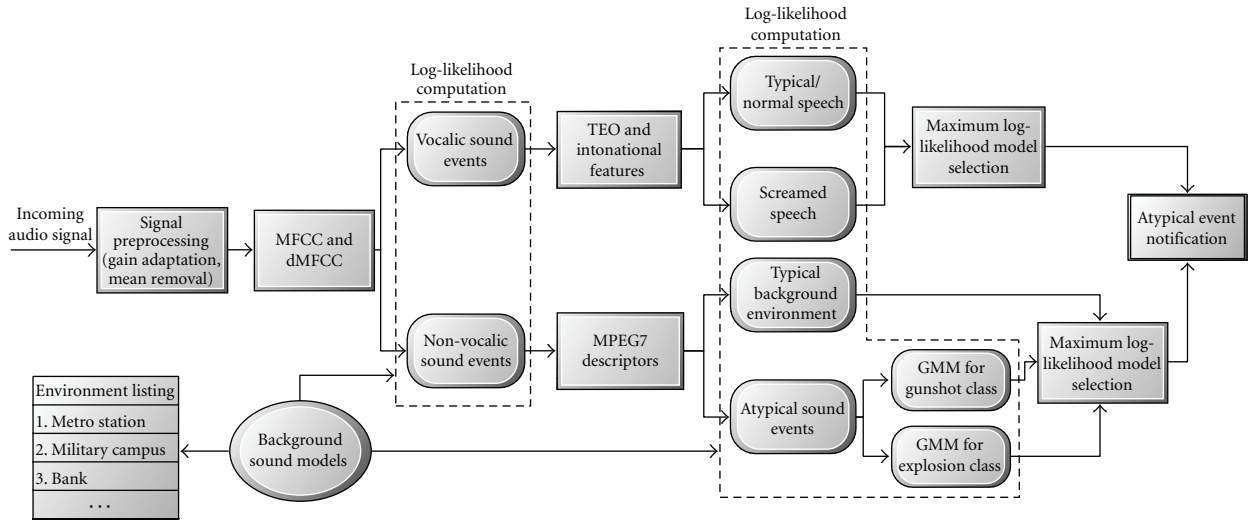


Figure 1: Block diagram of the acoustic surveillance system.

2.1.1. Mel-Frequency Cepstral Coefficients. The energy of short time Fourier transform associated with each frame is first computed and then filtered using Mel scale which lowers the number of dimensionality and emphasizes spectral bands that are important to human perception. Subsequently the logarithmic operator is applied to the extracted coefficients and finally the discrete cosine transform is used to decorrelate them. We retain the thirteen most important vectors, including the 0th one which expresses the energy of the signal. Furthermore, cepstral mean normalization is employed throughout the paper. The rate at which they change over time is also calculated to result to a feature vector with twenty six dimensions. MFCCs are used alone to discriminate vocalic sound events from the non-vocalic ones as well as during the other phases of system's topology combined with the descriptors explained next.

2.1.2. MPEG-7 Audio Protocol Descriptors. The idea behind MPEG-7 audio protocol is the creation of a large set of standardized tools for automatic audio content characterization. The sound descriptors that are calculated in both cases (vocalic and non-vocalic sound event detection) contain information that is complementary to MFCCs. While MFCC comprise a general description of the audio event, MPEG-7 LLDs reflect upon the flatness, ratio of the geometric and arithmetic mean of the spectral power coefficients associated to a given band (ASF), the envelope's structure (AWF) and the periodicity (AFF) of the specific sound, thus characterizing it at a higher level. In the case of non-vocalic sound events this information is crucial and needs to be taken under account during the modeling procedure. On the contrary when a vocalic sound event appears in the audio stream, the needed features are the ones with capabilities to identify whether a vocalic segment is typical or atypical. Pitch and harmonicity measurements are also included to characterize the periodic character of the signal. Comparative results regarding the addition of these groups of parameters are depicted in Table 2.

2.1.3. Intonation and Teager Energy Operator Based Features. TEO analysis takes place on the sixteen critical bands. Gabor band pass filters are used to focus on a particular spectral band while each one's TEO profile is windowized into frames of 200 milliseconds with 75% overlap. Subsequently the autocorrelation envelope area is computed and then normalized by half the frame length. The output feature vector has sixteen coefficients like the number of the critical bands. The audio analysis which relies on Teager energy operator can reveal aspects of verbal or non-verbal human reactions which are not captured by MFCC and are related to stress expression. It has been reported to be indicative of the alterations that the airflow pattern exhibits regarding the speech production under atypical circumstances. They are appended to pitch, pitch derivative and HNR (based on a forward cross-correlation analysis) which depict the variation of intonation regarding typical and atypical speech. Together with the already computed MFCC they form a vector for discriminating between normal and screamed speech audio events. For their calculation we used PRAAT software [13] which is optimized for speech signals.

2.2. Classification Process. We utilized diagonal Gaussian mixture models (GMM) for modeling the distribution of each sound class. They are based on the underlying assumption that the distribution of the data belonging to each class can be described statistically by a linear combination of Gaussian distributions. Torch [14] implementation of GMMs, written in C++ was used while the maximum number of k-means iterations for initialization was 50 and the EM algorithm had and upper limit of 25 iterations with a threshold of 0.001 applied to the difference between two consecutive iterations. The probabilistic models were stored

Table 2: Recognition rates achieved regarding each stage of system's topology for different kinds of environments. The recognition score without the additional feature extraction stage is depicted in parenthesis for comparison.

Classification problem	No. of mixtures	Feature set	Recognition rate (%)
Vocalic versus non-vocalic sound events (subway environment)	64	MFCC+dMFCC	100
Vocalic versus non-vocalic sound events (urban environment)	128	MFCC+dMFCC	99.85
Vocalic versus non-vocalic sound events (military environment)	128	MFCC+dMFCC+MPEG-7 LLDs	100
Typical versus atypical non-vocalic sound events (subway environment)	128	MFCC+dMFCC+MPEG-7 LLDs	97.2 (87.6)
Typical versus atypical non-vocalic sound events (urban environment)	128	MFCC+dMFCC+MPEG-7 LLDs	92.95 (88.2)
Typical versus atypical non-vocalic sound events (military environment)	32	MFCC+dMFCC+MPEG-7 LLDs	100 (91.6)
Explosion versus gunshot sound events	512	MFCC+dMFCC+MPEG-7 LLDs	83.9 (76.4)
Normal versus screamed speech	128	MFCC+dMFCC+intonation +CB-TEO-auto-Env	100 (89.1)

and consequently used for determining the log-likelihood that a specific sample was generated by the specific model. Finally this type of score with respect to every model was computed and the origin of the sound sample was identified by selecting the model with the maximum log-likelihood.

3. Experimental Set-Up

This section contains the description of the organization of the experiments that were conducted. In Section 3.1, we explain classification tests regarding every stage of the proposed system for determining the number of Gaussian components that offers the highest recognition rate. In Section 3.2 the system incorporating the previously constructed models was evaluated in terms of false alarm and detection rates through an artificial kind of experiment at various SNR conditions. Atypical sound events were randomly merged with background noise and tested for detection. The last experimental phase, discussed in Section 4 aimed at simulating an ongoing atypical situation. The system including the feedback loop for model adaptation was tested upon the detection of simulated atypical and typical situations.

Atypical audio data including extreme emotional manifestations and abnormal sound events are not publicly available because of their private character, their scarcity and unpredictability [15]. Data that indicate catastrophic situations as regards the particular kind of dangers mentioned in this article were identified and isolated from large scale professional sound effect collections. These types of collections are mainly used by the movie industry due to their vast variety, massive quantity and high quality. It is almost always the case that a movie's audio stream (e.g., footsteps, door knocking, etc.) is not the one that was actually recorded at the scene but a different one imposed to the movie. Thus, there exists an enormous corpus of almost any kind of audio events including non-vocal as well as vocal atypical reactions for building up statistical recognition models. The final corpus was acquired 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, (v) TIMIT and (vi) sound effects from various internet sources. By using these datasets simultaneously a high degree of variation as well as diversity regarding the entity of the audio classes was incorporated to the models. The sampling rate of all sound samples was 16 kbps with 16 bit analysis while the average duration for each category and in total is given in Table 3. Furthermore, in Figure 2 the spectrograms of representative samples taken from each category are illustrated. Atypical sound events were artificially merged with highly non-stationary background noise of the under study environment for simulating abnormal situations and conducting detection experiments. Details concerning the evaluation of the system under different SNR conditions are provided in the next section.

Table 3: The audio corpus.

Category	Number of sound records	Average duration of each record in seconds
Explosion	131	13.77
Gunshot	187	32.94
Scream	270	4.04
Normal speech	1680	3.08
Subway environment	32	44.88
Urban environment	106	83.35
Environment suited for military applications	31	68.69
Total	2437	35.82

3.1. Statistical Model Construction and Classification Experiments. We employed 75% of the data belonging to each class for training a statistical model to represent the corresponding sound class. The remaining 25% were used for testing while splitting was done in a random way. Audio pattern recognition is based on the underlying assumption that each audio effect has a unique way in which is spreading its energy across different frequency bands. This constitutes its so-called audio signature which can be revealed and subsequently identified automatically using statistical pattern analysis techniques. GMMs with diagonal covariance matrix were built for every category while testing consisted of a simple comparison of log-likelihoods. Following the topology of the system two types of models were created first: *vocalic* (including normal and screamed speech) and *nonvocalic* (including explosion, gunshot and the respective background environment). Subsequently we build up models regarding normal speech, screamed speech, typical background environment and atypical sound events (including explosion and gunshot). After extensive experimentations on the number of Gaussian mixtures and on the feature sets used during each classification problem, we achieved the recognition rates which are given in Table 2 associated with every stage of the system's architecture for different kind of environments. During this experimental phase we classified each sound sample of the corpus by summing the log-likelihoods obtained from each model with respect to all the frames of the specific recordings. Finally the model which presented the highest summed log-likelihood was selected and its class was assigned to the particular sound.

As it can be observed, the average recognition rates achieved during every processing step of the system are relatively high and in some cases such as normal versus screamed speech, the discrimination rate reaches 100%. In the specific task the dissimilarity regarding the spectral/energy distribution that screamed vocal reactions exhibit when compared to normal speech is reflected. This kind of difference is illustrated in Figure 2 among with characteristic spectrograms of each audio category. The classification of explosions and gunshot sound events presents the lowest rate which is 83.9%. At the particular stage, the errors occur due to the great variability among sound recordings of the same class. Additionally several sound clips are acoustically similar even though they belong to different categories, meaning that some explosion sounds like gunshots and vice versa. Further incorporation of a series of sound descriptors not only did not provide better performance, but raised the computational cost. After extensive experimentations, we managed to capture distinctive and characteristic information of the audio classes using a feature vector of rather low number of dimensions for every phase of the system's topology.

3.2. Detection of Abnormal Situations in Different Acoustic Environments. During the second experimental phase, threatening situations were generated artificially by merging abnormal sound events of three types with subway, urban and military background environment under different SNR conditions (varying from −5 dB to 15 dB with 5 dB step). Normal/typical audio events (background soundscape and normal speech) were also provided as input for measuring the abnormality level that our topology characterizes them with. This is an early indicator of the false alarm rate produced by our methodology, which should be kept to a minimum.

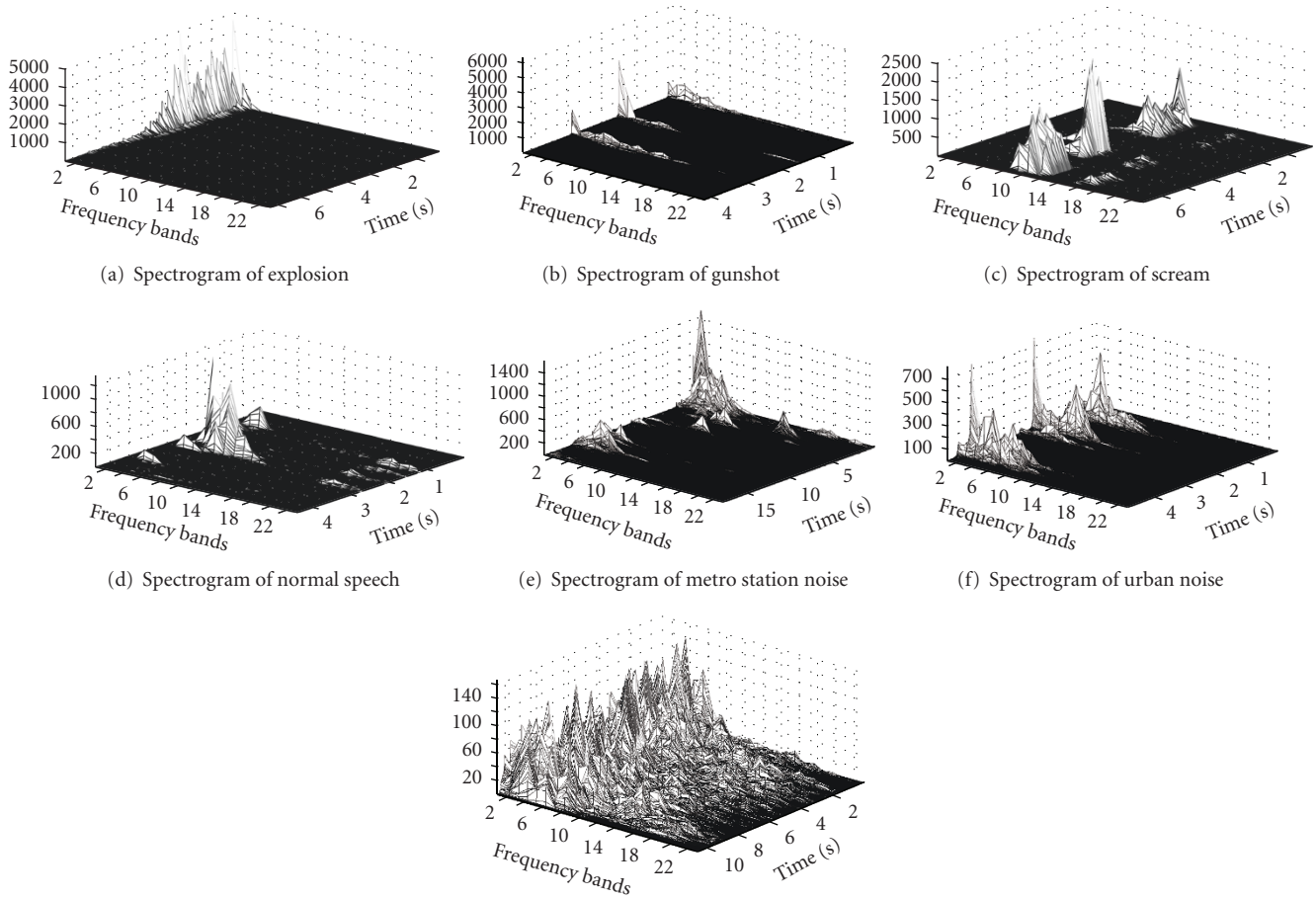
Our efforts are towards constructing a system that can efficiently work under different kind of environments. We demonstrate the performance of the proposed methodology operating in three different kinds of audio environments: subway, urban and military. The subway soundscape includes horns, opening/closing doors, people talking in the background, train locomotion and so forth. The urban one is dominated by movements of transport vehicles (e.g., cars, motorcycles, buses, etc.) and crowd noise but it also contains wind, rain, thunder as well as other sounds associated with weather conditions. The background setting suited for military applications is characterized by a great deal of sounds that appear during a military operation [16].

Recognition rates and/or confusion matrices do not constitute a sufficient way for reliable evaluation of an event detection task. In the particular type of problems there exist two kinds of errors that one should be aware of and try to minimize: (i) the case where an atypical sound event *is present* but it is not identified and (ii) the case where an atypical sound event *is not present* but it is falsely detected. The needed testing platform is provided by Detection Error Tradeoff (DET) curves which have been shown to be effective for the evaluation of detections tasks [17]. In our case special attention was placed upon achieving as low false alarm rates as possible for creating a useful and practical system.

The DET curves detect abnormal situations under three kinds of environments and SNR conditions. Fifty representative atypical sound events were selected from each category (scream, explosion, and gunshot) which were artificially merged with a part of the same size belonging to the normal environmental background soundscene chosen randomly. This procedure was iterated for each atypical event 50 times for obtaining reliable results, thus each class of abnormal situations was tested for identification 2500 times.

Two series of experiments were conducted for atypical sound event detection under metro station and urban environment. Each recording is normalized by its maximum value (gain normalization). The DET curves for both types of environment are illustrated in Figures 3 and 4. The log-likelihood produced by the abnormal non-vocalic Gaussian mixture was employed in the case of explosion and gunshot detection while the log-likelihoods produced by the atypical vocalic mixture were used to generate the screamed versus normal speech DET curves. These values were each time normalized with the respective normal one.

Figure 3 depicts results of atypical sound event detection for all three different sound categories under metro station background environment. A rapid degradation is observed when the SNR condition of the test signals decreases. However emergency situations are adequately detected even at very low SNR conditions. In the case of −5 dB SNR the average equal error rate (EER) of all types of events is 8.29% while the best detection rate concerns the abnormal



(a) Spectrogram of explosion

(b) Spectrogram of gunshot

(c) Spectrogram of scream

(d) Spectrogram of normal speech

(e) Spectrogram of metro station noise

(f) Spectrogram of urban noise

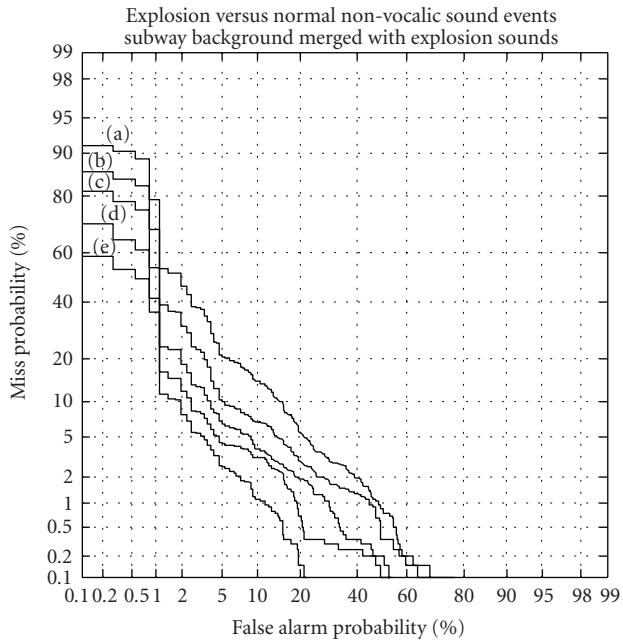
(g) Spectrogram of military enviroment

Figure 2: Representative figures of all background environments and atypical sound events in the Mel-spectral domain.

vocalic sound events with 6.46% EER. This is an outcome of the structure of our implementation, where each stage discriminates audio signals that have different spectral patterns and share only a few common characteristics. The audio signals that are most vulnerable to background noise corruption are the "gunshots" with 12.47% EER at −5 dB SNR. At the energy ratio of 0 dB which represent real-world conditions appropriately, the proposed system demonstrates high performance with average EER of 6.68%, average miss detection rate of 16.4% and average false alarm probability for abnormal events of 2.26% which is of severe importance for this kind of applications.

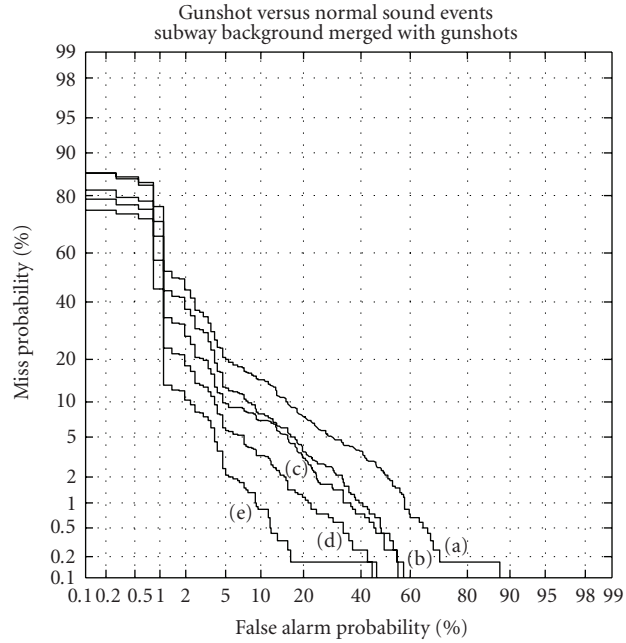
Figure 4 illustrates the capabilities of our implementation under urban environment. At this stage we used the statistical models that were created with the inclusion of urban audio data. As expected, miss detection probability falls as the SNR conditions increase from −5 dB to 15 dB. Atypical sound events are detected with relatively low EERs across all SNR values when the audio signal is corrupted by urban background environmental noise. We observe that better performance is achieved with average EER at −5 dB SNR being 5.19% in contrast to subway background. More precisely emergency situations at −5 dB SNR are detected with EERs of 6.05%, 4.35% and 5.19% where the abnormality refers to explosion, gunshot and scream sound events respectively. The events that are less affected by background noise are scream sounds while explosion detection presents the highest EERs across all SNR conditions. Additionally, our implementation provides very low false alarm probability with a mean value of 1% for the three abnormal sound events at 0 dB SNR conditions while the respective average miss detection rate is 13.2%. The corresponding EERs achieved by the system regarding to abnormal situation expressed as explosion, gunshot and screams are 5.78%, 4.23%, and 1.7%, respectively.

The respective DET curves regarding to the case of an environment suited for military applications are shown in Figure 5. As it can be visually verified, the results in this case are significantly improved. They achieve very low EERs with respect to the detection of all three different kinds of atypical situations. The best detection rates appear in the screamed speech case, which was expected due to the different structure of the specific audio signals compared to the variations that the particular environment exhibits. Abnormal situations are well detected even at low SNR conditions. More specifically in the case of −5 dB SNR the



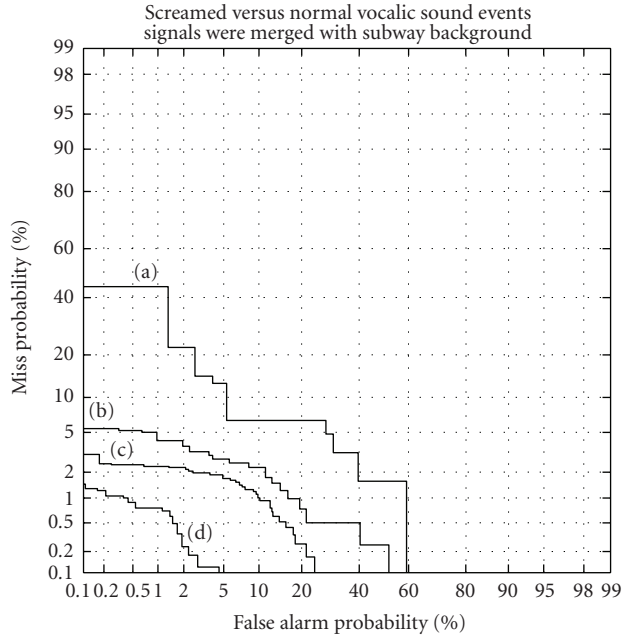
(a) SNR = −5 dB, EER = 0.1235
(b) SNR = 0 dB, EER = 0.0799
(c) SNR = 5 dB, EER = 0.0601
(d) SNR = 10 dB, EER = 0.048
(e) SNR = 15 dB, EER = 0.0391

(a)



(a) SNR = −5 dB, EER = 0.1247
(b) SNR = 0 dB, EER = 0.0878
(c) SNR = 5 dB, EER = 0.0823
(d) SNR = 10 dB, EER = 0.056
(e) SNR = 15 dB, EER = 0.0434

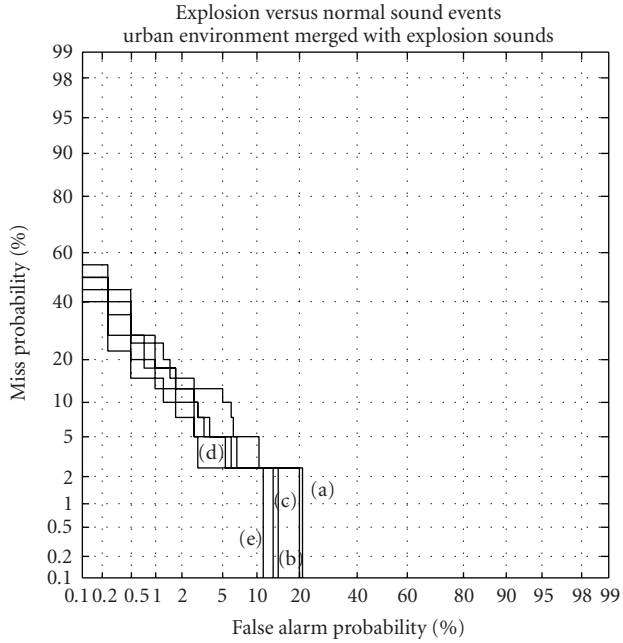
(b)



(a) SNR = −5 dB, EER = 0.0646
(b) SNR = 0 dB, EER = 0.0329
(c) SNR = 5 dB, EER = 0.0209
(d) SNR = 10 dB, EER = 0.0093
(e) SNR = 15 dB, EER = 0.0005

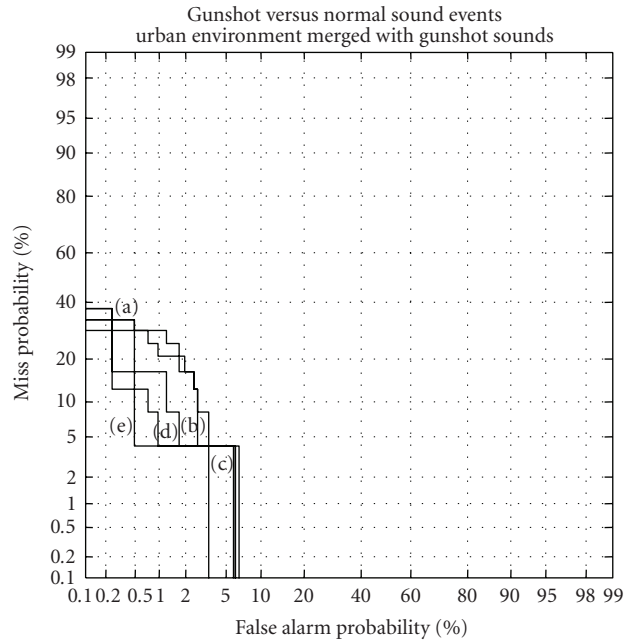
(c)

Figure 3: DET curves. Target classes are explosion, gunshot, screamed, and normal speech. Background is subway noise under different SNRs.



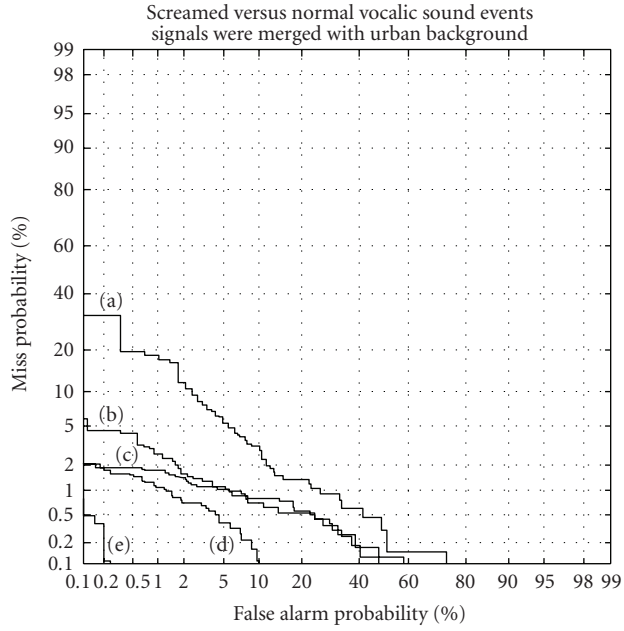
(a) SNR = −5 dB, EER = 0.0605
(b) SNR = 0 dB, EER = 0.0578
(c) SNR = 5 dB, EER = 0.0505
(d) SNR = 10 dB, EER = 0.0505
(e) SNR = 15 dB, EER = 0.0295

(a)



(a) SNR = −5 dB, EER = 0.0435
(b) SNR = 0 dB, EER = 0.0423
(c) SNR = 5 dB, EER = 0.0418
(d) SNR = 10 dB, EER = 0.0415
(e) SNR = 15 dB, EER = 0.0354

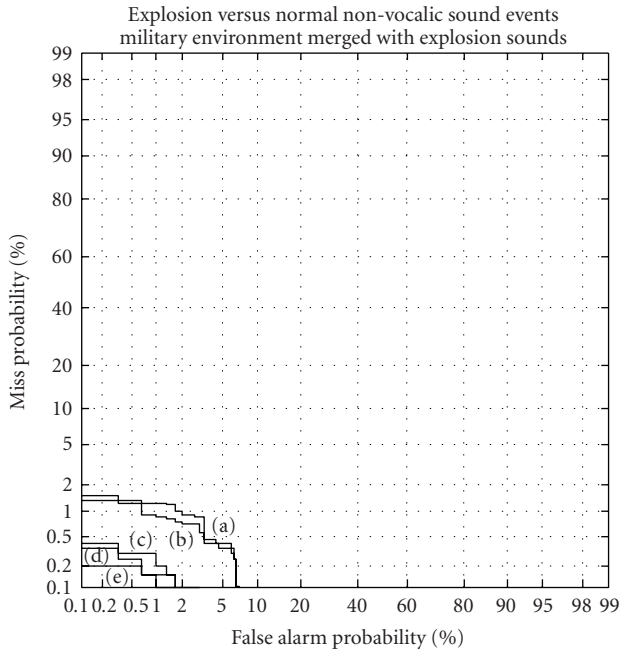
(b)



(a) SNR = −5 dB, EER = 0.0519
(b) SNR = 0 dB, EER = 0.017
(c) SNR = 5 dB, EER = 0.0154
(d) SNR = 10 dB, EER = 0.0112
(e) SNR = 15 dB, EER = 0.0024

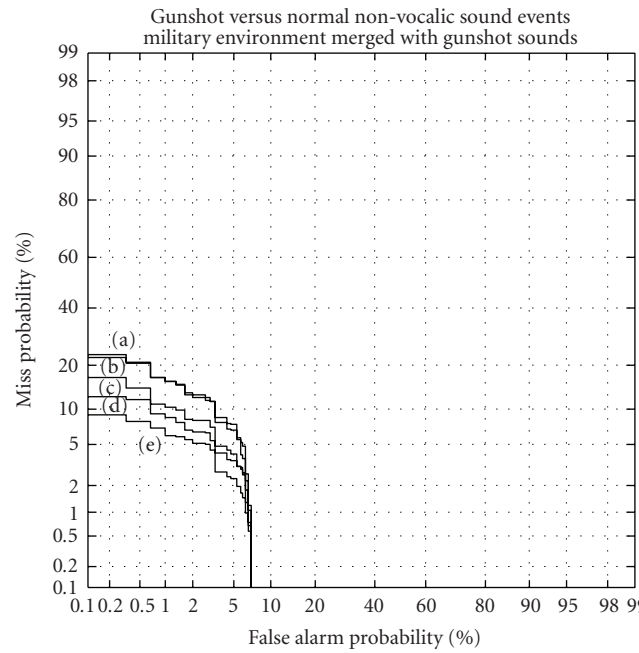
(c)

Figure 4: DET curves. Target classes are explosion, gunshot, screamed, and normal speech sound events. Events are merged with urban noise under different SNRs.



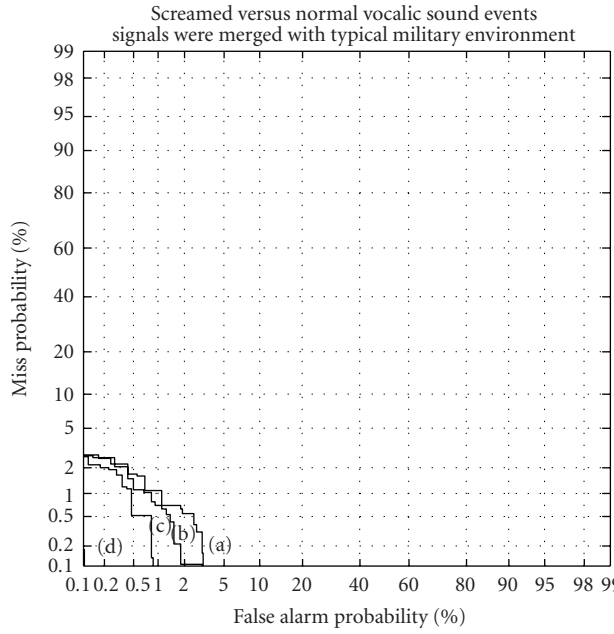
(a) SNR = −5 dB, EER = 0.0193
(b) SNR = 0 dB, EER = 0.0123
(c) SNR = 5 dB, EER = 0.0011
(d) SNR = 10 dB, EER = 0.0009
(e) SNR = 15 dB, EER = 0.0005

(a)



(a) SNR = −5 dB, EER = 0.0524
(b) SNR = 0 dB, EER = 0.0469
(c) SNR = 5 dB, EER = 0.0418
(d) SNR = 10 dB, EER = 0.0312
(e) SNR = 15 dB, EER = 0.0306

(b)



(a) SNR = −5 dB, EER = 0.0103
(b) SNR = 0 dB, EER = 0.0081
(c) SNR = 5 dB, EER = 0.0053
(d) SNR = 10 dB, EER = 0.0011
(e) SNR = 15 dB, EER = 0.0003

(c)

Figure 5: DET curves target classes are explosion, gunshot, screamed, and normal speech sound events. Events are merged with military environment under different SNRs.

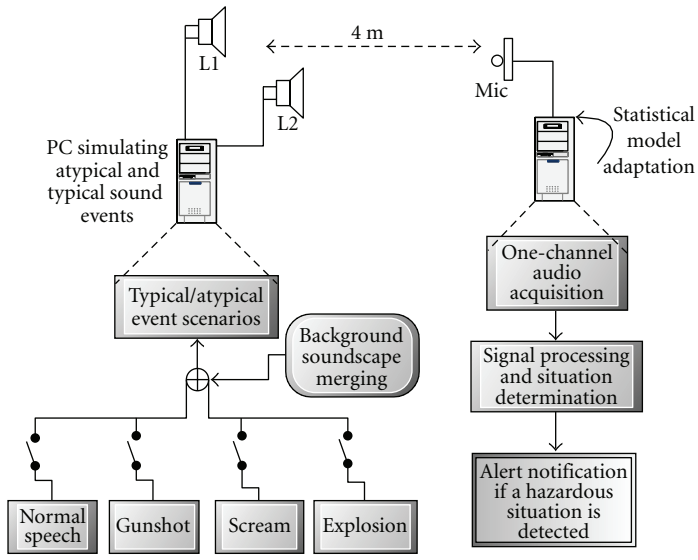


Figure 6: Diagram of detection and classification experiments inside a real internal space.

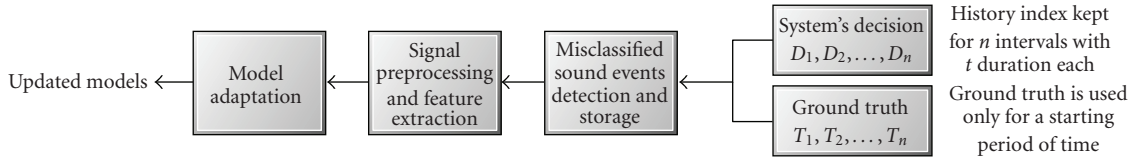


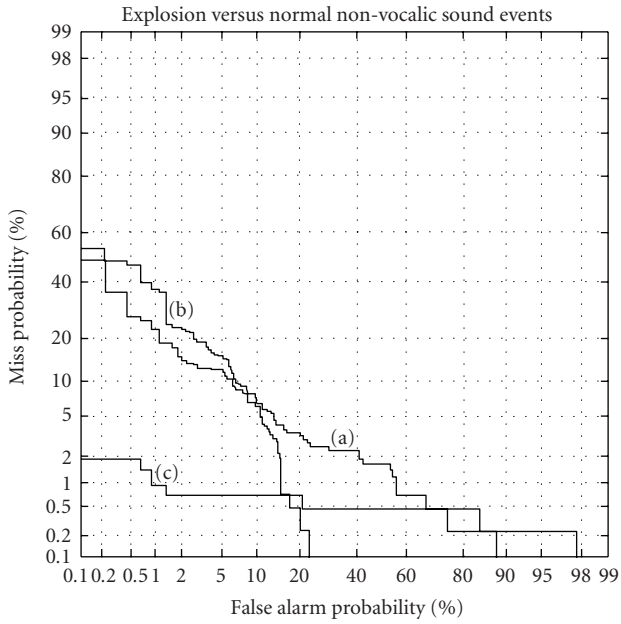
Figure 7: Feedback loop for GMM adaptation.

obtained EERs are 1.93%, 5.24% and 1.03% for explosion, gunshot and scream sound event detection respectively. An important objective of such a system is to limit the probabilities of false alarms. Care was taken regarding this aspect and in the case of 0 dB SNR we had 0.93% average false alarm probability and 2.67% average miss detection probability for the three kinds of atypical situations and 2.24% average EER (it should be mentioned that the corresponding EERs for every atypical sound event under all three different types of environment are shown on the upper right corner of every figure). We conclude that the results analyzed in this section are very encouraging and underline the importance of the selected statistical architecture in which features that capture different aspects of the audio structure were incorporated.

4. Experimentations in Real Internal Spaces

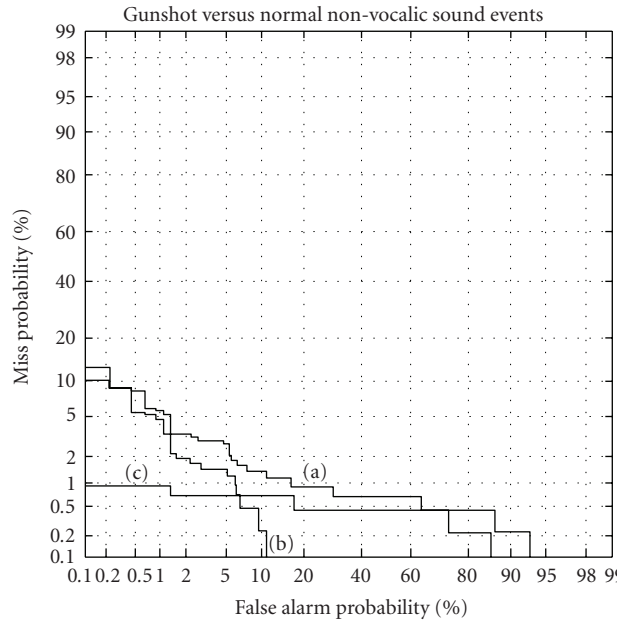
Our main goal during the third experimental phase was to approximate real-life operational conditions and to evaluate the statistical model adaptation which is provided via the feedback loop (see Figure 7). Atypical situations were artificially created at a random way, as described in Section 3 and played through loudspeakers while a microphone was placed in another part of a real $6.75 \times 4.9 \times 3$ room with reverberation time of 0.3 second. We employed two personal computers, one was reproducing abnormal sound events at predetermined time instances (so as to have *a-priori* knowledge of the ground truth) through two conventional loudspeakers while the second one was constantly capturing audio data with a simple microphone. Subsequently these data were processed and classified by the proposed system. This personal computer was also used to carry out both supervised and unsupervised model adaptation while the entire set-up is depicted in Figure 6.

4.1. The Problem of Sliding Windows. Several issues came up in the specific type of experiment including the sliding windows and the rareness in which atypical events appear. Windowing the incoming audio signal into chunks of a predefined size was not adequate to provide satisfying results because the duration of explosions, gunshots and screams sound events varied greatly. Neither the start time nor the duration of a keysound effect was known to the system, thus it may be cut into parts belonging to different sound classes. To this end we decided to process the incoming audio signal on a frame by frame basis (200 milliseconds with 75% overlap), while the consecutive frames with the same label were merged into one segment with the start time corresponding to the first frame and the duration corresponding to the total number of frames in the segment. Both pieces of information were saved for helping future investigation of the scene at the particular time as well as the sum of log-likelihoods normalized by the number of frames for each sound event. Moreover, a smoothing process was then applied to remove unreasonable inconsistencies among neighboring frames. Basically we removed single frame



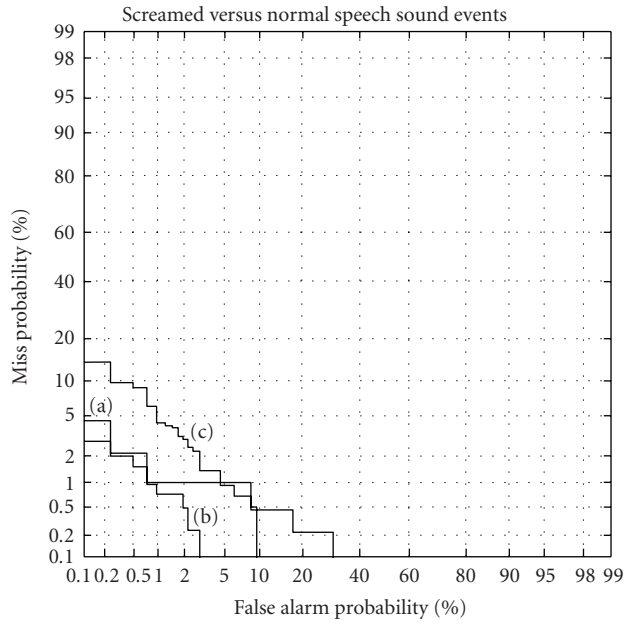
(a) Metro, EER = 0.0798
(b) Military, EER = 0.0774
(c) Urban, EER = 0.0102

(a)



(a) Metro, EER = 0.0247
(b) Military, EER = 0.0173
(c) Urban, EER = 0.0079

(b)



(a) Metro, EER = 0.0099
(b) Military, EER = 0.0084
(c) Urban, EER = 0.0223

(c)

Figure 8: DET curves representing the ability of the adapted system to detect both nonvocalic and vocalic abnormal sound events under three different kind of environments.

Table 4: Equal error rates for the three phases of the experiment.

Adaptation phase	Environment	Average EER (for three atypical events)	Average EER for three environments (% improvement)
No Adaptation	Subway	0.32461	0.26253 (−)
	Urban	0.2901	
	Military	0.17289	
Supervised	Subway	0.11226	0.09676 (63.14%)
	Urban	0.12673	
	Military	0.0513	
Both supervised and unsupervised	Subway	0.03786	0.02856 (70.48%)
	Urban	0.01346	
	Military	0.03436	

detections (outliers) which correspond to 200 milliseconds and do not comprise a logical duration regarding to atypical sound events.

4.2. Dealing with the Rareness of Atypical Situations. The second issue consisted of not only the rareness which characterizes the existence of abnormal situations but also the fact that another *non interesting* sound event may take place and cause a misdetection (false alarm). It is a rather difficult task to create accurate models to represent these sounds mainly because their existence is hard to be known a-priori (e.g., jiggling keys, horns, dog barking, etc.) while it is important not to be misinterpreted by the system as an atypical events. To this aim, we collected data from various types of sources making sure that the background soundscape of each scenario is described by a high degree of variance (e.g., the subway soundscape includes horns, opening/closing doors, people talking in the background, train locomotion, etc.).

Although the audio samples are representative of the categories we need, they do not provide an accurate description of all possible realizations of such events. Consequently we incorporate a technique which gives our system the ability to *adapt itself* to the acoustic conditions of the operational environment. Adaptivity is provided by the application of the *maximum a posteriori* (MAP) method to the statistical models of each class and is implemented via a feedback loop [18]. Initially this is a supervised semiautomatic process which takes into account the ground truth as an input from authorized personnel. After this starting period of time the system adapts in an unsupervised manner exploiting its own decisions.

The feedback loop is depicted in Figure 7. A history index is kept that contains the series of decisions made by the system regarding n intervals with duration of t seconds, where each duration depends on the outcome of the matching process, that is, when the system predicts the same class for consecutive frames, these specific frames then comprise one audio sequence while a new sequence is formed when the prediction changes. The ground truth was also kept in parallel for the same periods of time and the audio data were stored at 16 KHz with 16 bit analysis. Subsequently the misclassified data were analyzed and the respective feature sequences were used for adapting the corresponding models. This phase takes place during an inactive period of time. Afterwards the system was adapted in an unsupervised manner using its own decisions to replace the manually reported ground truth. More specifically the process of unsupervised adaptation works in the following way: for a given period of time all the segments including the corresponding predictions are stored. Subsequently the appropriate sound parameters are extracted by the system (e.g., MFCC and dMFCC for adapting the vocalic/non-vocalic model). These parameters are then employed to adapt the respective model according to system's prediction. This way the system is capable for autonomous model refinement and thus further adapting *itself* to the environmental conditions.

4.3. DET Curves of the Adapted System. At the first phase, the particular experiment was conducted for three subsequent days while the ground truth was known. Half of these data were manually analyzed for classification errors and then used for adaptation of the respective models (supervised *MAP* adaptation). At the second phase the Gaussian models were adapted in an unsupervised manner based on decisions made automatically by the system, utilizing audio data that were captured during one day. The results reported in this section were obtained using the log-likelihoods that were given as outputs by the models that were adapted after both phases. During the adaptation process the parameters of the Gaussian components (weights, variances and means) were learnt from the adaptation data while the value of the prior weight during update was set to 0.5 (changes to this parameter did not provide better performance). In contrast to other adaptation algorithms (e.g., maximum likelihood linear regression), the specific methodology requires more adaptation data since it works at the component level. However because of this low-level approach, when large amount of data are available *MAP* method tends to perform better, something that holds to our approach since we collected 72 hours (3×24 hours) of data. Half of these data were used for supervised model adaptation. The next twenty

four hours served unsupervised adaptation while the rest of the data (12 hours) were employed for testing the adapted system. It should be mentioned that this experimental stage exploits the reverberation reduction properties that cepstral mean normalization offers.

The DET curves with respect to each one of the three environments are depicted in Figure 8 (explosion, gunshot and atypical speech detection). As we can see the best detection rates are achieved in the urban environment (average EER = 0.01346) followed by the environment suited for military applications (average EER = 0.03436) and the subway (average EER = 0.03786) one. Additionally false alarm rates are kept to low values regarding the typical conditions of all three environments. It should be noted that the system is to continue the adapting process in an unsupervised manner, thus achieving even better performance. In Table 4 the EERs which correspond to every stage of the particular experiment are tabulated. As we can see there exist significant improvements at both adaptation phases. We conclude that the results of the adapted system are quite promising and show the portability and flexibility that the proposed structure offers.

5. Conclusions

We proposed an integrated system for acoustic surveillance of atypical situations. We investigated a large number of feature sets in order to conclude to the best representatives of an atypical situation that involves audio expressions of pain, stress, gunshots and explosions. We constructed a hierarchical system that is based on probabilistic models which was trained using a large amount of high quality sounds from professional sound effects collections. The system was evaluated under adverse conditions containing highly nonstationary background noise under three different kinds of environments. The system is adaptable to the internal or external space where it is installed by using online model adaptation. The latter can become an indispensable part of a practical acoustic surveillance system.

Our future work includes the incorporation of blind channel equalization methods to deal with the problem of reverberation. Furthermore we will work on the adaptation module for making it faster and more efficient by using a thresholding technique on the confidence score. This will produce a weighting measure on the adaptation data so that the system exploits in a different way data with different confidence measures. Finally, we intend to fuse the likelihood of the atypical event with information from visual and infrared sensors so as to provide enhanced detection of hazardous events.

Acknowledgment

This work was supported by the EC FP 7th Grant PROMETHEUS 214901 "Prediction and Interpretation of human behaviour based on probabilistic models and heterogeneous sensors."

References

[1] S. Park and H. Kautz, "A hierarchical recognition of activities in daily living using multi-scale, multi-perspective vision and RFID," in *Proceedings of the IET International Conference on Intelligent Environments*, Seattle, Wash, USA, 2008.

[2] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809–830, 2000.

[3] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Communication*, vol. 50, no. 6, pp. 487–503, 2008.

[4] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS '07)*, pp. 21–26, London, UK, September 2007.

[5] R. Radhakrishnan and A. Divakaran, "Systematic acquisition of audio classes for elevator surveillance," in *Image and Video Communications and Processing 2005*, vol. 5685 of *Proceedings of SPIE*, pp. 64–71, March 2005.

[6] C. Clavel, T. Ehrette, and G. Richard, "Event detection for an audio-based surveillance system," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '05)*, Amsterdam, The Netherlands, July 2005.

[7] A. Harma, M. F. Mckinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '05)*, pp. 634–637, Amsterdam, The Netherlands, July 2005.

[8] J.-L. Rouas, J. Louradour, and S. Ambellouis, "Audio events detection in public transport vehicle," in *Proceedings of the IEEE International Conference on Intelligent Transportation Systems (ITSC '06)*, pp. 733–738, Toronto, Canada, September 2006.

[9] M. Vacher, D. Istrate, L. Besacier, J.-F. Serignat, and E. Castelli, "Sound detection and classification for medical telesurvey," in *Proceedings of the International Conference on Biomedical Engineering*, pp. 395–399, Innsburg, Austria, February 2004.

[10] P. K. Atrey, N. C. Maddage, and M. S. Kankanhalli, "Audio based event detection for multimedia surveillance," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, vol. 5, pp. 813–816, Toulouse, France, May 2006.

[11] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 201–216, 2001.

[12] S. Quackenbush and A. Lindsay, "Overview of MPEG-7 audio," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 725–729, 2001.

[13] PRAAT software, http://www.fon.hum.uva.nl/praat/.

[14] Torch Machine Learning Library, http://www.torch.ch/.

[15] C. Clavel, I. Vasilescu, L. Devillers, and T. Ehrette, "Fiction database for emotion detection in abnormal situations," in *Proceedings of the International Conference on Spoken Language Processing*, Jeju, South Korea, October 2004.

[16] F. Chen, "Speech technology in military applications," in *Designing Human Interfaces in Speech Technology*, Springer, New York, NY, USA, 2006.

[17] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech '97)*, Rhodos, Greece, September 1997.

[18] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.