A PORTABLE SYSTEM FOR ROBUST ACOUSTIC DETECTION OF ATYPICAL SITUATIONS

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

[1]Department of Electrical and Computer Engineering, University of Patras
Rion, 26500, Patras, Greece
[2]Department of Music Technology and Acoustics, Technological Educational Institute of Crete
Rethymno, 74100, Crete, Greece
phone: + (30) 2610 969806, fax: + (30) 2610 977 336, email: sntalampiras@upatras.gr
web: http://www.wcl.ece.upatras.gr/dalas

ABSTRACT

This work presents a practical, automatic and robust methodology for acoustic surveillance of hazardous situations. The proposed system efficiently identifies atypical situations which include scream, explosion and gunshot sound events under different kind of environments (e.g. metro station, urban etc). The main objective is to detect abnormal events which lead to life-threatening circumstances or property damage by helping an authorized officer to take the appropriate actions through a decision support interface. After extensive experimentations, a fully probabilistic structure of Gaussian mixtures was designed which incorporates task depended feature sets. A testing procedure under different SNR conditions was followed and we report high detection rates with respect to false alarm and miss probability rates.

1. INTRODUCTION

Lately automatic systems which monitor human daily activities are becoming increasingly common. The main aim is civil safety which is achieved through surveillance of public spaces for recognition of potentially hazardous situations. Atypical events are the ones that imply a threat to human life or property loss/damage. The basic purpose of our work is robust and reliable detection of such circumstances by exploiting solely the acoustic modality. In this type of situations extreme emotional manifestations, gunshots and explosions are usually encountered. Our goal is to build a system that detects on time a crisis situation and to provide this result to an authorized officer for further evaluation and action. Such a system should be characterized by accuracy, user friendliness and flexibility meaning that with slight alterations the system can work properly under different kind of environments.

The research area of acoustic surveillance has gained a lot of attention recently addressing various types of applications. It is a branch of generalized sound recognition technology, namely computational auditory scene analysis. This particular domain tries to understand the surrounding environment using the incoming audio as its only input, inspired by the respective property that humans exhibit in their everyday life quite effortless. In [1] an emotion recognition system is described that makes use of prosody and audio quality combined with spectral and cepstral parameters to train Gaussian mixture models (GMMs). Their database was the SAFE corpus. The classification task concerned fear and neutral speech while they achieved 30% error rate. Valenzise et al [2] presented a surveillance system for gunshot and scream detection and localization in a public square. Forty-nine features were computed in total for building two parallel GMMs in order to identify 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. An interesting application, crime detection inside elevators was described in [3]. A GMM for each one of the eight classes was trained using low-level features. The data set contained recordings of suspicious activities in elevators and some event free clips. A gunshot detection method under noisy environments was explained in [4]. Their corpus consisted of data which were artificially created from a set of multiple public places and gunshot sound events extracted from a CD of sounds for 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. Acoustic surveillance in a typical office environment was explored in [5]. The audio files were captured in a standard office room for a period of 48 days. The detection was based 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 [6] the issue of detection of audio events in public transport vehicles was addressed. The incoming audio data were first automatically segmented and then classified using GMM and SVM as shout or non shout using a hierarchical architecture. 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. Vacher et al [7] 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) as well as files from a commercial CD. They used wavelet based cepstral coefficients to train GMMs for eight sound classes while their system was evaluated under different SNR conditions.

This paper is an extension of a previous work of ours [8] while the system architecture has been altered so that typical speech parts are also considered and several important sound parameters are included for achieving better performance. We report on a complete, practical and real-time framework for acoustic surveillance of hazardous situations that is flexible towards working under different kind of environments (e.g. metro station and urban). Furthermore our dataset is thorough and concise after combining several well documented professional sound effect collections which contain audio of high quality. Our methodology can find use in many surveillance tasks such as in a military environment, banks, and public means of transportation.

2. SYSTEM OVERVIEW

The main goal of our system is to emphasize on detection of human vocal reactions and non-vocal atypical events associated with hazardous situations. To this end, the structure that was designed has the form depicted in Figure 1. The unknown sound class is predicted

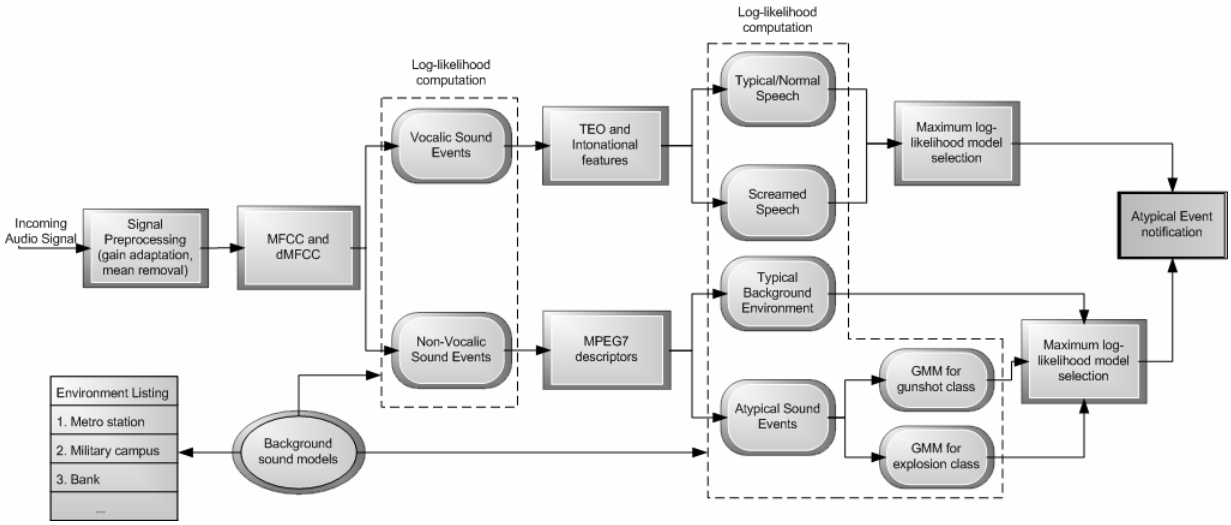


Figure 1 - The proposed probabilistic structure for atypical event detection.

through three discrete subsequent stages, where each stage depends on the previous one. The sound is classified as a vocalic (normal or screamed speech) or a non-vocalic (background environment, gunshot or explosion) event. In case it is found to be a vocalic event, another set of descriptors is computed and the sound is recognized as normal or screamed speech. In the opposite case, 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. The selected architecture comprises a fully probabilistic structure which utilizes diagonal GMMs to estimate the probability function of each sound category while their parameters were defined after exhaustive experiments regarding each classification phase.

2.1 Feature Extraction

The low level attributes that are extracted from the audio signals for constructing statistical models are explained in this paragraph. We exploit the property of Mel-scale filterbank to compress the dimensions of the Fourier transformed vector, while its output is logarithmically partitioned. We also employed a variety of LLDs provided by the MPEG-7 protocol since it currently comprises the state of the art in the area of content-based audio recognition [9]. We concluded into using the next four LLDs: *Waveform Min, Waveform Max, Audio Fundamental Frequency and Audio Spectrum Flatness*. Additionally we used a group of features which have been shown to provide accurate results as regards classification of speech under stress [10]. TEO autocorrelation envelope area was computed for discriminating between normal and screamed speech signals. Furthermore we searched for features that are indicative of the variations that intonation exhibits when it comes to atypical speech. Intonation can be expressed as the pattern of pitch alterations during speech, thus we used PRAAT [11] software to calculate pitch, pitch derivative as well as harmonic to noise ratio (HNR). It should be mentioned that all the signals were hamming-windowed using a frame size of 200ms with 75% overlap in order to smooth any discontinuities.

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

A great variety of standardized tools for automatic multimedia content description is incorporated into the MPEG-7 audio standard. Its main objective is to offer a degree of "explanation" of the information meaning. It eases navigation of audio data by providing 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. The following parameters were used:

- Audio Spectrum Flatness (ASF)

This descriptor is a measure of flatness of a particular portion of the signal and represents the deviation of the signal's power spectrum from a flat shape. The power coefficients are taken from non-overlapping frames while the spectrum is divided into ¼-octave resolution logarithmically spaced overlapping frequency bands. The ASF is derived as the ratio of the geometric mean and the arithmetic mean of the spectral power coefficients within a band. This feature can efficiently differentiate between noise (or impulse) and harmonic sounds and we should take into account that a large deviation from a flat shape generally depicts *tonal* sounds.

- Audio Waveform (AWF)

This constitutes compact description of the shape of an audio signal by computing the minimum and maximum samples within successive non-overlapping frames.

- Audio Fundamental Frequency (AFF)

For a given and assumed to be periodic portion of the signal AFF consists of an estimation of the fundamental frequency f_0. It can be used as an approximation of the pitch of musical sounds and voiced speech.

Intonation and Teager Energy Operator based features

The specific analysis concerns sixteen critical bands. Initially the signal is passed through Gabor filters for concentrating on a particu-

lar spectral band and each one's TEO profile is computed. Subsequently the autocorrelation envelope area of each frame is computed and then normalized by frame length/2. The output feature vector has sixteen coefficients like the number of the critical bands. They are used combined with pitch, pitch derivative and HNR 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.

The sound descriptors that are additionally calculated in both cases (vocalic and non-vocalic sound event detection) contain complementary to MFCC information and are specialized for serving the following classification step. While MFCC comprise a general description of the audio event, MPEG-7 LLDs reflect upon the flatness (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 modelling 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. 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. They are believed to be indicative of the alterations that the airflow pattern exhibits regarding the speech production under atypical circumstances. Pitch and harmonicity measurements are also included during this phase for offering information regarding the periodic character of the signal (in general, normal speech is to be more periodic than atypical speech). Comparative results regarding the addition of these groups of parameters are depicted in Table I.

Table I - Average 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	Number of Gaussian Components	Average Recognition Rate (%)
Vocalic vs. Non-Vocalic sound events (Subway environment)	64	100
Vocalic vs. Non-Vocalic sound events (Urban environment)	128	99.85
Typical vs. Atypical Non-Vocalic sound Events (Subway environment)	128	97.2 (87.6)
Typical vs. Atypical Non-Vocalic sound Events (Urban environment)	128	92.95 (88.2)
Explosion vs. Gunshot sound events	512	83.9 (76.4)
Normal vs. Screamed Speech	128	100 (89.1)

2.2 Classification Process

During classification procedure we used a generative approach, Gaussian mixture models. This approach is based on the assumption that the data belonging to a specific class follow a mixture of Gaussian distributions. This distribution can be approximated using the Expectation-Maximization (EM) algorithm which results to the creation of statistical models. The main characteristic of this type of classifiers is that they handle the samples of each class independently of the other classes. Subsequently the previously constructed models are used for computing a degree of resemblance (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 [12] implementation of GMM, written in C++ was used during the whole process. The maximum number of k-means iterations for initialization was 50 while the EM algorithm had and upper limit of 25 iterations with a threshold of 0.001 between subsequent iterations.

3. EXPERIMENTAL SET-UP

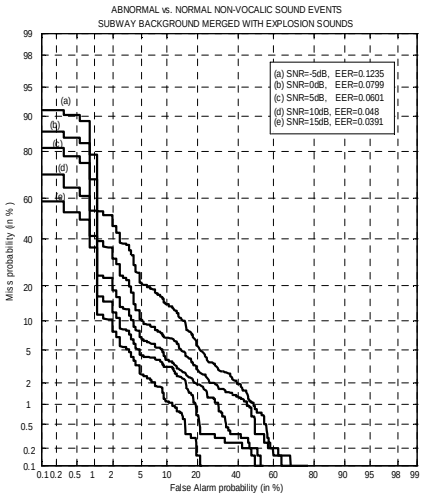
The audio data that were used for training the statistical models and testing the proposed system are reported in this paragraph. Natural corpora with extreme emotional manifestation and atypical sounds events for surveillance applications are not publicly available because of the private character of the data, their scarcity and unpredictability [13]. 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 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, (v) TIMIT and (vi) sound effects from internet sources were identified and isolated for putting together the final corpus. The concurrent usage of these datasets offers great variability and diversity regarding the *a-priori* knowledge which is to be incorporated to the probabilistic models.

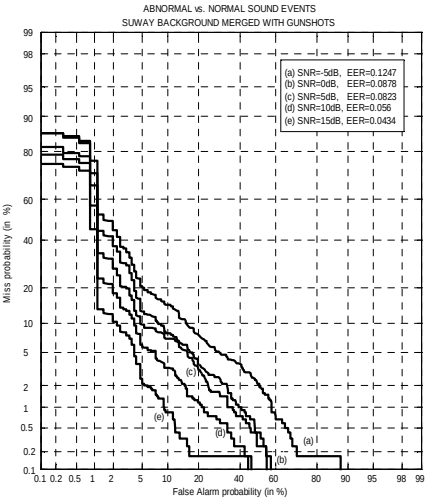
3.1 Model Construction and Average Recognition Rates

The data belonging to each class were splitted into 75% for training and 25% for testing in a random way. Content based audio recognition is based on the fact that every sound source is distributing its energy across different frequencies in a different way. A diagonal GMM was built for each category while testing consists of a simple comparison of log-likelihoods. Due to the system architecture we first constructed two kinds of models: vocalic (including screamed and normal speech) and non-vocalic (including explosion, gunshot and the respective environmental soundscape). After extensive experimentations on the number of Gaussian components we concluded to use the parameters tabulated in Table I. As it can be seen high recognition rates are achieved during every stage of the proposed implementation, showing the effectiveness of the selected feature sets and statistical method.

3.2 Detection of Atypical Situations in different kind of Environments

Emergency situations located in a metro station or urban environment were artificially created by merging abnormal sound events which indicate danger, crisis and high-risk in general, with subway and urban soundscape at different SNRs (from -5dB to 15dB with 5dB step). It should be mentioned that unlike previous studies [4] we utilized only clean sound samples to train our system. The proposed architecture depicted in Fig. 1 incorporating the respective probabilistic models that were described in the





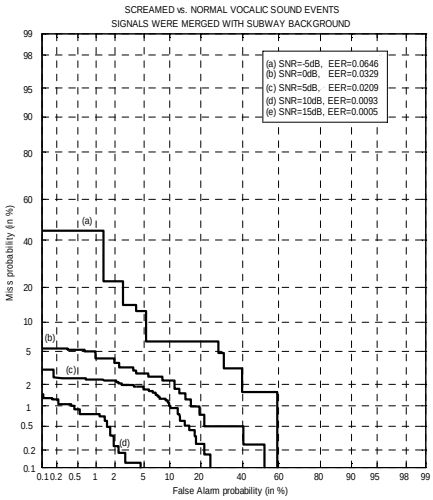
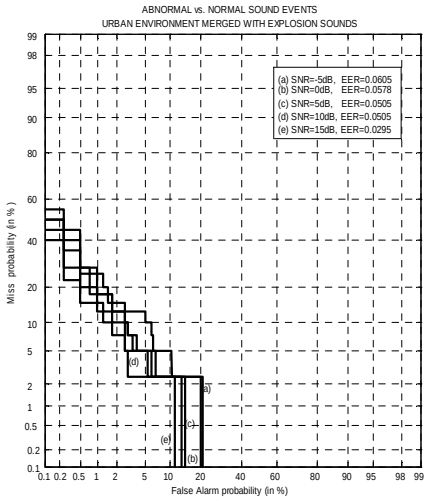
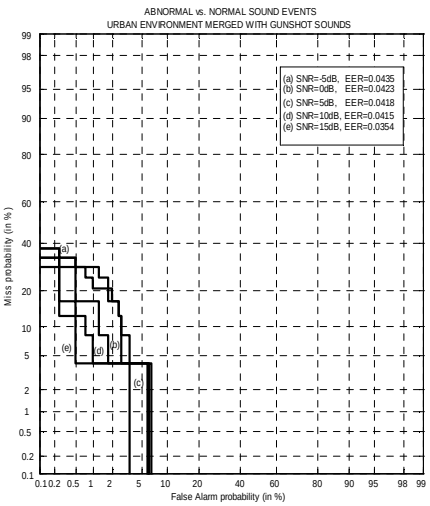


Figure 2 - DET curves regarding to atypical sound events as the target class when explosion, gunshot, screamed and normal speech sound events were merged with *subway* soundscape under different SNRs.





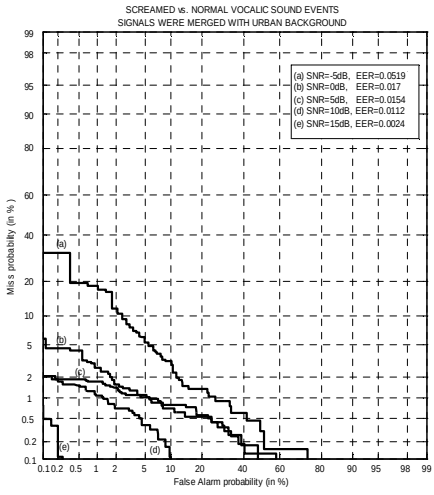


Figure 3 - DET curves regarding to atypical sound events as the target class when explosion, gunshot, screamed and normal speech sound events were merged with *urban* soundscape under different SNRs.

former paragraph, was tested using Detection Error Tradeoff (DET) curves, which have been shown to be effective for the evaluation of detections tasks [14]. The performance of such a system cannot be analyzed by a simple recognition rate because

of the underlying tradeoff error - detection of an atypical situation may fail or such an event may be declared when it is not present.

Fig. 2 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 -5dB SNR the average equal error rate (EER) of all types of events is 8.29% while the best detection rate concerns the abnormal vocalic sound events with 6% EER. This is an outcome of the structure of our implementation, each stage of which discriminates audio signals which have different spectral patterns and share only a few common characteristics. The audio signals that are most vulnerable to background noise corruption are the gunshot ones with 12.47% EER at -5dB SNR. At the energy ratio of 0dB which represent real-world conditions appropriately, the proposed system demonstrates high performance with EER of 6.68% and false alarm probability 2.26% which is of severe importance for this kind of applications.

Fig. 3 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 -5dB to 15dB. 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 -5dB SNR conditions being 5.19% in contrast to subway background. More precisely emergency situations at -5dB ratio are detected with EERs of 6.05%, 4.35% and 5.19% when 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 good false alarm probability with a mean value of 1% among the three sound event categories with 0dB SNR conditions. 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. Conclusively it can be observed that the results are quite promising and underline the effectiveness of the selected probabilistic structure, which incorporates features of high discrimination capabilities.

4. CONCLUSIONS

Threatening situations such as crime and terrorist acts in large urban areas are not fictitious scenarios but real facts that require special attention and measures. In this work we presented and evaluated a probabilistic framework for acoustic monitoring in a metro station and urban environment. Its main aim is to provide a practical, easily deployable, real-time system that can identify on time the sensed situation (unlawful act in progress, an accident and/or atypical behaviour) and deliver the necessary warning messages to an authorized officer. The proposed approach was tested against highly non-stationary metro station and urban background noise and demonstrated robust and reliable atypical event detection under adverse conditions. Future work includes one-channel signal separation as well as the incorporation of mixtures on-line adaptation in the recognition stage.

5. ACKNOWLEDGEMENTS

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] 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, June 2008.

[2] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci and A. Sarti, “Scream and gunshot detection and localization for audio-surveillance systems,” in *AVSS*, London, England, September 2007.

[3] R. Radhakrishnan and A. Divakaran, “Systematic acquisition of audio classes for elevator surveillance,” *SPIE Image and Video Communications Processing*, vol. 5685, pp. 64-71, March 2005.

[4] C. Clavel, T. Ehrette and G. Richard, “Event detection for an audio-based surveillance system,” in *IEEE International Conference on Multimedia and Expo*, Amsterdam, Holland, July 6-8. 2005, pp. 1306 - 1309.

[5] A. Harma, M.F. McKinney and J. Skowronek, “Automatic surveillance of the acoustic activity in our living environment,” in *IEEE International Conference on Multimedia and Expo*, Amsterdam, Holland, July 2005.

[6] J.-L. Rouas, J. Louradour and S. Ambellouis, “Audio events detection in public transport vehicles,” in *IEEE Intelligent Transportation System Conference*, Toronto, Canada, September 2006.

[7] M. Vacher, D. Istrate, L. Besacier, J.-F. Serignat and E. Castelli, “Sound detection and classification for medical Telesurvey,” in International Conference of Biomedical Engineering, Innsburg, Austria, February 2004.

[8] S. Ntalampiras, I. Potamitis and N. Fakotakis, “On acoustic surveillance of hazardous situations,” to be presented in *ICASSP 2009*, Taiwan, Taipei, April 2009.

[9] H.-G. Kim, N. Moreau and T. Sikora, *MPEG-7 Audio and Beyond: audio content indexing and retrieval*. Wiley Publishers, October 2005.

[10] G. Zhoun, J. H. L. Hansen and J.F. Kaiser, “Nonlinear Feature Based Classification of Speech Under Stress,” *IEEE Trans. on Speech and Audio Proc.*, vol. 9, no. 2, pp. 201-216, March 2001.

[11] PRAAT software provided at http://www.praat.org

[12] Torch Machine Learning Library at http://www.torch.ch

[13] C. Clavel, I. Vasilescu, L. Devillers and T. Ehrette, “Fiction database for emotion detection in abnormal situations,” in *ICSLP 2004*, Jeju, Korea, October 2004.

[14] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, “The DET Curve in Assessment of Detection Task Performance,” in *Eurospeech 97*, Rhodos, Greece, September 1997.