Audio Database in Support of Potential Threat and Crisis Situation Management

Stavros Ntalampiras[†], Ilyas Potamitis[‡], Todor Ganchev[†], Nikos Fakotakis[†]
[†]Department of Electrical and Computer Engineering,
University of Patras, 26500 Rion-Patras, Greece
[‡]Department of Music Technology and Acoustics, Technological Institute of Crete,
Daskalaki-Perivolia, 74100, Rethymno, Crete, Greece
E-mail: sntalampiras@upatras.gr, potamitis@stef.teicrete.gr, tganchev@ieee.org, fakotaki@wcl.ee.upatras.gr

Abstract

This paper describes a corpus consisting of audio data for automatic space monitoring based solely on the perceived acoustic information. The particular database is created as part of a project aiming at the detection of abnormal events, which lead to life-threatening situations or property damage. The audio corpus is composed of vocal reactions and environmental sounds that are usually encountered in atypical situations. The audio data is composed of three parts: Phase I - professional sound effects collections, Phase II recordings obtained from action and drama movies and Phase III - vocal reactions related to real-world emergency events as retrieved from television, radio broadcast news, documentaries etc. The annotation methodology is given in details along with preliminary classification results and statistical analysis of the dataset regarding Phase I. The main objective of such a dataset is to provide training data for automatic recognition machines that detect hazardous situations and to provide security enhancement in public environments, which otherwise require human supervision.

1. Introduction

The analysis of human behaviour in unrestricted environments, including localization and tracking of multiple people as well as recognition of their activities, currently constitutes a topic of intensive research in the signal processing and computer vision communities (Haritaoglou et al. 2000). This research is driven by different important applications, including unattended surveillance and intelligent space monitoring.

Despite the legitimacy of several privacy issues, many systems have been deployed for surveillance applications. These systems generate large amounts of data that need to be filtered out through automatic analysis, either for online detection of dangerous situations, or for offline information retrieval. This work is part of a project that conducts research on the characterization of atypical movements, crowd move and collective behaviour interpretation, with a view to identify and respond to unusual but critical actions (e.g. violent acts, accidents, potential threat or hazard). The main objective is to provide a decision support interface to enhance the performance of human experts by overlaying machine prediction on the experts' information to alert security officers to an unlawful act in progress, or accident.

An abnormal/atypical situation in the context of prediction and interpretation of human behaviour is an unplanned event which may lead to a human life threatening situation and requires prompt action to protect life or to limit its damages. The main idea behind this work is to describe the audio data, which will be used to perform automatic detection and categorization of such events. Automatic recognition of atypical events is of great importance for prediction, early detection and prevention of danger, emergency, crisis and other high-risk situations (Jaques, 2007). The state-of-art technology for automatic sound recognition (Couvreur, 1998; Aucouturiera, 2007) relies on the statistical comparison of the time-frequency characteristics of the unknown sounds with the ones that have been extracted from real-world recordings of events captured in such situations. Beside acoustic events (Coskun et al. 1999), human speech is also indicative source, which can facilitate the identification of crisis or danger (Yang & Rothkrantz, 2007). It is well-known that the accuracy and performance of automatic recognition systems are largely dependent on the availability of appropriate data sources, which are utilized for training statistical models of the events of interest. However, the data collection and annotation is not a trivial task and as a rule requires multidisciplinary expertise, and major investment of efforts, time and money.

As regards Phase I, we have gathered sounds that are usually encountered in the context of atypical situations such as footsteps, hits, explosions etc. In Phases II and III we assume that in atypical situations humans experience negative emotions such as fear, pain, anger and sorrow. Emotion detection for security applications based solely on the acoustic modality requires an appropriate real-life database.

However, natural corpora with extreme emotional manifestation for surveillance applications are not publicly available because of the private character of the data, their scarcity and unpredictability (Clavel & Vasilescu, 2004; Clavel et al. 2006). In the present work, we report on the current development and analysis of a dedicated database of speech and sounds, which are indicative of atypical events.

2. Objectives and Database Development

The primary step for the creation of a dedicated single-channel audio surveillance machine is the construction of a database, which provides comprehensive representation of atypical speech and sound events related to danger, emergency, crisis and high-risk situations in general. The basic purpose of this database is to train probabilistic models that describe different audio classes.

The database development is organized in three subsequent phases, each contributing to the diversity of the final corpus. Specifically, the phases of our project depend on the sources that are utilized:

Phase I contains 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. Phase I, which is now completed, is composed of two parts:

Part A: Collection and analysis of recordings of impulsive sounds related to acoustic monitoring of atypical situations (e.g. glass breaking, dropping of objects, fracture of material, hits, footsteps, door sounds, gunshots, weather phenomena, sounds related to cars and explosions). Ten different categories are included with 100 recording sessions for each category.

Part B: Human vocal sounds related to negative emotions. The categories we have focused are (i) pain, (ii) fear, (iii) sorrow and (iv) anger.

Phase II involves acquisition of speech and sounds from action and drama movies. In the latter case, our selection criteria rely on actors realistic playing and situation naturalness, and on the type of abnormal situation that will match the application field of acoustic surveillance. Moreover, we consider only these segments where the quality of vocalizations allows for proper modeling, i.e. fragments where noise or music dominates over speech are excluded. A corpus of 100 audio-visual sequences is in progress of acquisition. Annotation with respect to manifestations of emotional states in abnormal situations, either in individual, group or crowd situations of emotional states in abnormal situations (crowd surges may connote an emergency situation), is ongoing.

Phase III covers the acquisition of sounds and speech from real-world emergency, crisis and catastrophic events, as they are available in television and radio broadcast news, documentaries, etc. The audio-visual sequences are currently extracted from broadcast news that manifest human verbal/non-verbal audio reaction in real threat and hazardous situations.

3. Annotation Procedure

Phase I is annotated according to the situation (e.g. footsteps on a wet substrate) and emotional category (e.g. child crying) as the sound effects collection usage is based on the efficient annotation tags in order to allow for the quick retrieval of the sound effect.

In Phases II and III, the annotation of speech and audio events is performed by employing the Anvil tool (Kipp, 2001), following the methodology of (Clavel & Vasilescu, 2004). Anvil was selected for this phase, since it provides the means for annotating not only the emotions manifested in the specific segment, but also for capturing context and the situation in details. The picture/video capability facilitates for proper capturing of the underlying emotional state, and reduces the disagreement between annotators. As it was observed at the early stages of our work, when video is available (in addition to the audio) the annotators tend to reach consensus on nearly all controversial fragments. In contrast to the audio only case, when a fragment is heard out of the context, humans tended to disagree in approximately 15% of the segments. The annotation tags we considered in the database are presented in Table 1.

Tag	***Options***
Sex	Male / Female
Age	Child / Adult / Elderly
Verbal	Yes / No
Situation	Normal / Aggression / Fight / Murder / Panic / Natural disaster / Other
Emotional state	Neutral / Pain / Fear / Sorrow / Anger
Data source	File name
Position in file	Frame/time coordinates
Audio quality	Clean / Noise / Music / Other noise

Table 1: Annotation tags used in Phase II.

4. Statistical Description of Sounds Related to Atypical Events

We performed statistical and objective evaluation of numerous sound parameterization techniques with respect to their appropriateness for automatic detection of atypical events. The subject of our analysis focused on the divergence of normal speech characteristics from atypical events. The aim of this investigation is to identify a set of non-redundant and informative parameters that can be utilized as feature vectors in GMM/HMM-based classification framework.

Specifically, several sets of sound parameters such as: nonlinear Teager energy operator (TEO)-based features (Zhou et al. 2001), critical band based TEO autocorrelation envelope area (TEO-CB-Auto-Env) features (Zhou et al. 2001), genetic algorithm selected features (16-GA and 48-GA) (Casalea, 2007), MPEG-7 sound recognition descriptors (Casey, 2001), Mel-frequency cepstral coefficients (Slaney, 1998), wavelet packet based audio descriptors (Sarikaya & Hansen, 2000), pitch, duration, intensity, are evaluated with respect to their capability to discriminate audio events, which indicate divergence from normal situations. As a special case we study a redundant set of audio descriptors, suitable for GMM/HMM and ANN-based classifiers. The redundant feature vectors are believed to

guarantee noise-robustness in adverse conditions and when portions or certain frequency bands of the signal are corrupted or absent.

In the evaluation of the appropriateness of audio descriptors, besides the statistical and information or relative entropy-based measures, we also employ an objective performance measure that accounts for the accuracy of detection of atypical events for each feature vector. In this way, instead of evaluating the individual parameters independently one from another, we study an objective manner which feature set leads to the lowest classification rates. This approach is equivalent to identifying the best *set* of sound parameters for the specific task, instead of selecting the best individual parameters and combining them.

4.1 Recognition of Atypical Sound Events

As a first step, we considered the TEO-CB-Auto-Env feature set as well as four low level descriptors from the MPEG-7 standard. The former set demonstrated good performance in the task of stress recognition (Zhou et al. 2001) while the latter was chosen to provide basic properties of the classes. Sixteen Critical Bands were used for the derivation of the first set while *waveform min*, *waveform max*, *audio fundamental frequency* and *audio power* complete the vector. The frame size was 200 samples (25 ms) with 80 overlapping samples between two subsequent frames. The process of sound recognition is based on the fact that each sound source distributes its energy across different frequencies in a unique way. This property is captured by left-right HMMs composed of seven states. Each state is modeled by Gaussian mixture consisting of five components. The Baum-Welch algorithm is utilized for training. The final classification is consisted of a conventional, maximum log likelihood estimation. The categories of atypical sound events that were modeled are the following ten: glass breaking, dropping of objects, fracture of material, hits, footsteps, door sounds, gunshots, weather phenomena, sounds related to cars and explosions. In Figure 1, a characteristic sample taken from each sound category is depicted in both time and frequency domain. A number of differences can be observed among these categories. For instance, the energy of a recording belonging to category weather phenomenon is much higher than the one produced by a door opening/closing sound.

During the experimentations we performed ten-fold cross validation to better utilize that available dataset. A frame-based decision was adopted. The average recognition rate is 76.7% and the confusion matrix is tabulated in Table 3. These results are quite promising while the usage of additional feature sets is expected to improve classification performance.

4.2 Data Analysis for Part B

The ability of the same features sets to describe sounds that belong to Part B is investigated in this paragraph. The extraction process was identical with the previous one resulting to a vector with twenty dimensions. Several

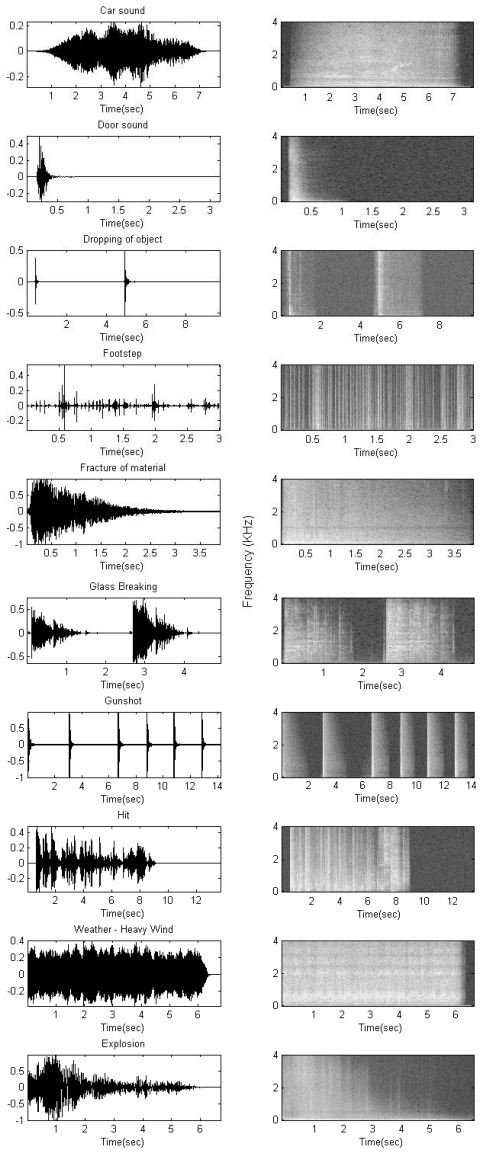

Figure 1: *Left column:* Characteristic time-domain plots of the categories of Phase I. *Right column:* Corresponding spectrograms of sounds.

statistical measures were calculated that characterize the distribution of the features among two classes atypical situations (anger, fear, pain and sorrow) and normal/typical situations (Table 2). Both male and female sound samples, obtained from the TIMIT database were used to represent normal speech. We present results

Category	Statistical Measure	TEO CB 2	TEO CB 4	TEO CB 6	WavMin
Normal Situation	Average	0.95	0.78	0.70	-0.20
	Variance	0.02	0.015	0.01	0.03
	Median	0.80	0.70	0.70	-0.07
	75^{th} percentile	0.90	0.80	0.80	-0.001
Atypical Situation	Average	0.90	0.88	0.78	-0.10
	Variance	0.035	0.025	0.02	0.01
	Median	0.95	0.90	0.76	-0.05
	75^{th} percentile	1.00	1.00	0.89	-0.01

Table 2: Statistical measures as regards Phase 1 – Part B

Presented \ Responded	Car sounds	Door sounds	Dropping of objects	Footsteps	Fracture of material	Glass breaking	Gunshots	Hits	Weather Phenomena	Explosions
Car sounds	**95**	0	0	0	5	0	0	0	0	0
Door sounds	0	**96**	0	4	0	0	0	0	0	0
Dropping of objects	0	0	**62**	15	0	0	0	0	23	23
Footsteps	0	0	0	**89**	0	0	0	11	0	0
Fracture of material	0	0	0	0	**97**	0	3	0	0	0
Glass breaking	0	0	0	0	0	**32**	26	42	0	0
Gunshots	0	0	0	0	0	0	**100**	0	0	0
Hits	0	0	0	0	40	0	30	**30**	0	0
Weather phenomena	0	0	0	12	0	0	0	0	**88**	0
Explosions	0	0	0	0	0	10	0	0	12	**78**

Table 3: Confusion matrix for Phase I – Part A (%)

regarding four representative descriptors, which were involved in the computations (TEO of Critical Bands 2, 4 and 6 as well as Waveform Min). As Table 2 presents there is a significant difference between the values of the parameters for the normal and atypical situations. For instance, 75% of the data for the normal category have a *Waveform Min* value lower than -0.001, while the same measure for atypical data is ten times higher. This is due to the fact that humans produce sounds with relatively high energy when it comes to abnormal situations. Furthermore, a significant difference among the average values can be noticed, considering the small dispersion that TEO-CB-Auto-Env features carry (0<TEO<1.16). Moreover, human sounds generated by atypical events produce higher values than normal sounds do, regarding the entity of the descriptors.

The present work reports preliminary results, and a further study is required to better understand the differences between normal and atypical events. We believe that the usage of more advanced audio parameterization techniques will contribute for a better separation of these categories.

5. Discussion and Conclusion

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. The main task of acoustic monitoring is to identify in time the sensed situation and deliver the necessary warning messages to an authorized officer. We have described the collection process and statistical analysis of extracted features on a database aiming specifically for acoustic surveillance of atypical events. We believe its construction and analysis is valuable for tasks as audio pattern recognition, signal separation, array signal processing, localization and tracking based on the acoustic modality. Future work includes the exploitation of more sophisticated descriptor sets and completion of Phase II and III.

6. 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”.

7. References

Aucouturiera, J.-J., Defreville, B., Pachet, F. (2007). The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *Journal of the Acoustic Society of America*, 122(2), pp. 881--891.

Casalea, S., Russo, A., Serranoa, S. (2007). Multistyle classification of speech under stress using feature subset selection based on genetic algorithms. *Speech Communication*, 49(10-11), pp. 801--810.

Casey, M. (2001). MPEG-7 sound-recognition tools. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6), pp. 731--747.

Clavel, Ch., Vasilescu, I. (2004). Fiction database for emotion detection in abnormal situations. In *Proceedings of the INTERSPEECH-2004*, pp. 2277--2280.

Clavel, Ch., Vasilescu, I., Devillers, L., Ehrette, T., Richard, G. (2006). Fear-type emotions of the SAFE corpus: annotation issues. In *Proceedings of the LREC-2006.*

Coskun, E., Hamzah, R., Kwong, C., Mesilogou, M., Mohan, R., Park, C., Yu, H.-Q., Grabowski, M. (1999). The Washington State Ferries Risk Assessment Project. Puget Sound Event Database Analysis. *Prepared for: Blue Ribbon Panel on Washington State Ferry Safety and Washington State Transportation Commission Olympia*, Washington.

Couvreur, C., Fontaine, V., Gaunard, P., Mubikangiey, C.G. (1998). Automatic classification of environmental noise events by hidden Markov models. *Applied Acoustics*, 54, pp. 187--206.

Haritaoglu, I., Harwood, D., Davis, L. (2000). W4: real-time surveillance of people and their activities. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22 (8), pp. 809--830.

Jaques, T. (2007). Issue management and crisis management: An integrated, non-linear, relational construct. *Public Relations Review*, 33, pp. 147--157.

Kipp, M. (2001). Anvil - A Generic Annotation Tool for Multimodal Dialogue. In *Proceedings of the INTERSPEECH-2001*, pp. 1367--1370.

Sarikaya, R., Hansen, J.H.L. (2000). High resolution speech feature parameterization for monophone-based stressed speech recognition. *IEEE Signal Processing Letters*, 7(7), pp. 182--185.

Slaney, M. (1998). Auditory Toolbox. Version 2. Technical Report #1998-010, Interval Research Corporation.

Yang, Z., Rothkrantz, L.J.M. (2007). Emotion Sensing for Context Sensitive Interpretation of Crisis Reports. In *ISCRAM-2007*, B. Van de Walle, P. Burghardt and C. Nieuwenhuis eds., pp. 507--514.

Zhou, G.-J., Hansen, J.H.L., Kaiser, J.F. (2001). Nonlinear Feature Based Classification of Speech Under Stress. *IEEE Transactions on Speech and Audio Processing*, 9(3), pp. 201--216.