Automatic detection of audio events indicating threats

Matúš Pleva, Eva Vozáriková, Stanislav Ondáš, Jozef Juhár, Anton Čižmár
Department of Electronics and Multimedia Communications
Technical University of Kosice
Kosice, Slovakia
{matus.pleva, eva.vozarikova, stanislav.ondas, jozef.juhar, anton.cizmar}@tuke.sk

Abstract— **This paper is focused on the area of audio event classification and detection for the purpose of citizens' security in the urban environment. There are various acoustic/audio events, which occur during the possibly dangerous situations. The main goal of our work was to build a simple audio event detection system trained on a database of recordings and to test the often used approach based on MFCC parametrization and HMM-based representation of acoustic information.**

Keywords - audio event, automatic detection, HMM, SVM

I. INTRODUCTION

The area of audio events detection and classification is nowadays an often solved problem [1] [6] [7] [8] [12]. Audio event detection and video event detection are a subtask of multimodal detection systems. Applications of these systems are used for particular tasks e.g. surveillance of public places, stadiums or on the stations of public transport [13] [14]. In these kinds of systems, the audio event detection system could generate alerts, when some possibly dangerous situation occurred. Such detection systems can be very important due to the fact, that human is not able to watch several audio signals simultaneously (man is not able to distinguish to which input source this event appropriates). Another situation is for video signals, where human is partially able to watch several screens in the same time, and he is able to identify the screen on which occurs some dangerous situation.

Also the Closed-Circuit Television - CCTV operators do not want to hear the audio from the surveillance cameras, because it can cause the increase of tiredness and the only reasonable way is the alert system, which will alert and replay the detected audio event together with notice about the source camera. It is also important to lower the false detections, because the alerts could become also a "noise" source for the operators.

The proposed work has the background in the European FP7 INDECT Project: "Intelligent information system supporting observation, searching and detection for security of citizens in urban environment", which started in year 2009 and its duration is 5 years. The main objectives of the INDECT project are to develop a platform for the registration and exchange of operational data, acquisition of multimedia content, intelligent processing of all information and automatic detection of threats and recognition of abnormal/strained behavior (cry, scream) or violence [4].

The main goal of this work is to build an audio event detection system, which will be able to indicate threat in urban environment. There were a lot of papers, which were focused on audio event detection in meeting rooms [5], [10]; or indoor environment, but only a few of them are related for outdoor or public places [7], [1].

At the beginning it is necessary to analyze the sounds, which the system may detect, and then to choose the appropriate feature set and the detection approach. It is also important to collect audio event corpus for training and testing purposes.

a) Feature set

A suitable feature set for representation of sounds plays an important role [1]. The often used feature set consists of MFCC (Mel Frequency Cepstral coefficients) parametrization and their time derivations. This feature sets well represents speech spectral structure but it has limited (but usable) performance for audio event, too. The same situation is with PLP (Perceptual Linear Prediction) - based feature sets.

Another approach is using of a set of several spectral parameters appended with other characteristics like short-time energy, zero-crossing rate, pitch, autocorrelation parameters, spectral entropy, spectral centroid, spectral roll-off, spectral kurtosis, spectral slope, spectral flatness, etc. [5], [6].

b) Detection methods

Different approaches based on the supervised, semi-supervised and unsupervised learning classification methods can be used for this task. For example generative methods like HMM, semi-supervised HMM, GMM and a large margin GMM were also studied. Other approaches are based on the SVMs.

The semi-supervised HMM-based approaches partially addressed the lack of labeled training data for events. Basic principle of this method is as follow: Usual models are trained (learned) from a large amount of data. Unusual (event) model is derived from usual model of an iterative process via Bayesian adaptation, where the number of iterations is equal to the number of unusual models [8].

The sound classes in large margin GMMs are modeled by ellipsoids - which induce nonlinear decision boundaries in the input space - as opposed to the half-spaces and hyperplanes in SVMs. Because the "kernel trick" is not necessary to induce nonlinear decision boundaries, large margin GMMs are more readily trained on large and difficult data sets, as arise in [9].

As an unsupervised approach we consider to test SVM [10], [11]. SVMs compute the linear decision boundary, which maximizes the "margin" of correct classification - that is the distance of the closest example(s) to the separating hyperplane.

If the labeled examples are not linearly separable, the "kernel trick" can be used to map the examples into a nonlinear feature space and to compute the maximum margin hyperplane in this space. We distinguish "soft" and "hard" SVM on dependency of solving misclassified examples. "Soft" SVM is more suitable for classification in noisy environment. Both approaches give very similar results.

Next, the section II. describes the classification of audio events, which may mean threat. The section III. informs about preparation of database of audio events for training of the acoustic models and section IV. deals with the training of HMM (Hidden Markov Models) models and with the building of the simple audio event detection system. The results of the experiments can be found in section V.

II. AUDIO EVENTS

The analysis of the audio event detection area started with a need of categorization of intended acoustic events. They can be divided in to three main categories:

A. *Speech-based audio events*

This group of events consists of all events, which are produced by human beings in a form of spoken words and phrases and they are related to threats, violence or dangerous situations. Calling for help, warning shouts, profanities, vulgarisms, etc. could be considered as the main items of this group.

B. *Non-speech audio events*

Non-speech audio events can be divided in to several groups:

- *Inarticulate sounds belonging to (coming from) human* beings: crying, screaming, fans crowd shout, etc.
- *Sounds belong to mobile objects /cars, trams, planes/*: traffic accidents, alarms and honks, sound of shear
- *Sounds accompanying threats or abnormal behavior:* broken glass (show-windows, bottles), explosions, fun pyrotechnics, shooting, car brake abruptly
- *Audio events produced by crowd of people:* audio events produced by crowd are a special group of sounds, which can indicate threats, a special attention should be dedicated to crowd tendencies, which indicates increasing of bad emotions
- *Other sounds:* sounds of battle/fighting, animal sounds like dog snarl, bark, yelp

C. *Ambient noises*

Audio input of the surveillance system in outdoor environment contains also the noise of the ambient: included music, sounds produced by the weather conditions (like strong rain, thunder storms, strong wind), trams, trains, buses etc.

III. DATABASE PREPARATION

The corpus of recordings, which have to include a large number of each sound realization, is necessary for training of models for audio events detection. In this experiment a reduced set of specific audio events was chosen, because of detection algorithms testing. It contains audio events which are interesting for surveillance systems such as explosion, broken glass, shot and shout. Also a background model, silent model and the "other" model for other loud events in the recordings were trained. Data were collected in a real environment (recorded near the stations, rush street, etc.) and from the affordable sources like youtube.com, movies, and recordings of TV shows "Odsúdené", "Prvé oddelenie"– from prison and police work). The collection is still in progress.

Acquired data were analyzed according to the duration, type and quality point of view. The next step was the processing of this data (pre-processing, annotation, normalization). The recording was demuxed and the audio stream was then re-sampled to 24 kHz, 16 bit, mono-channel format (using approximation and filtering algorithms). The sampling rate was selected for easy resampling from 48 kHz recordings, which is a frequency often used in a higher quality videos. Also, the frequency analysis of audio events, which we have realized shown, that the energy differences between the events is located up to 12 kHz. The recordings were manually labeled using Transcriber [3]. Particular events are not overlapped. In the training and testing process were used recordings with noise. The composition of the corpus is following:

TABLE I. ACTUAL STATE OF THE DATABASE

	Train (files)	Test (files)	Total (files)
Explosions	9	3	12
Broken glass	40	18	58
Shot	50	23	73
Shout	34	15	49
Other	20	6	26
Background	40	10	50
Silent	60	24	84

IV. BUILDING OF AUDIO EVENT DETECTION SYSTEM

During the analysis of this task we selected a basic set of potentially dangerous sounds, which includes sounds (audio events) like: shouting, the sound of broken glass, explosions and gunshots. This group is of course not complete and we will extend it.

In our experiment, HMM classifier was used to classify the sound of audio events. The goal of acoustic processing is provided by appropriate method to determination of the conditional probability *P(O/W)*, which means a probability, that a word/event *W* will represent acoustical vector/observation *O*. We used continuous ergodic HMM-s using from 2 to 16 PDF (Power Density Function) mixtures and different number of states. We used MFCC (Mel-frequency cepstral coefficients), their first and second time derivation (delta (D), delta-delta - acceleration (A)). The vector size has initially 36 parameters. Then a more feature extraction configurations were tested. The energy (0), log-energy (E), and cepstral mean subtraction (Z) parameter were tested and the recognition results were compared. The best configuration was when computing thirteen MFCC, energy, delta and acceleration coefficients with cepstral mean subtraction (E_D_A_Z).

Next, the window size was set to 20 ms and the step was 10 ms. The high frequency contents of many types of sounds require a higher sampling frequency than usually used in speech processing. In this work 24 kHz was used. The sampling frequency 8 kHz decreased the audio events detection results.

The audio recordings were manually annotated and then the models for each event (shot, shout, explosion, broken glass) were trained by HTK toolkit [2]. Models for silent, background and other sounds/events were also trained.

The Fig. 1. depicts the architecture of the proposed audio event detection system.

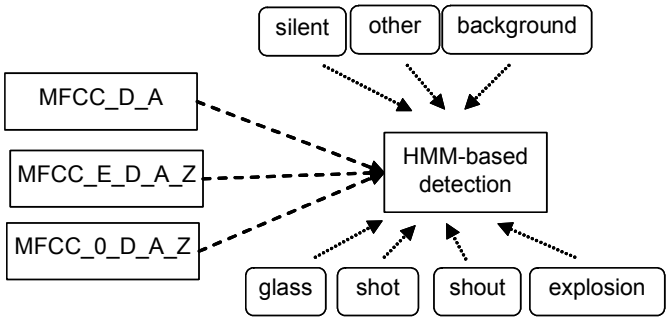

Figure 1. Audio event detection system

V. RESULTS OF THE EXPERIMENT

Three types of parametrizations were compared (MFCC_D_A, MFCC_0_D_A_Z, MFCC_E_D_A_Z) using from 1 to 5 number of states and from 2 to 16 PDF mixtures. For evaluating the percentage of labels correctly recognized '%correct' was used. In [2] correctness is defined using:

$$\%Correct = \frac{H}{N} * 100\%,$$

where H is the number of correct recognized labels and N is the total number of labels in the reference transcription files.

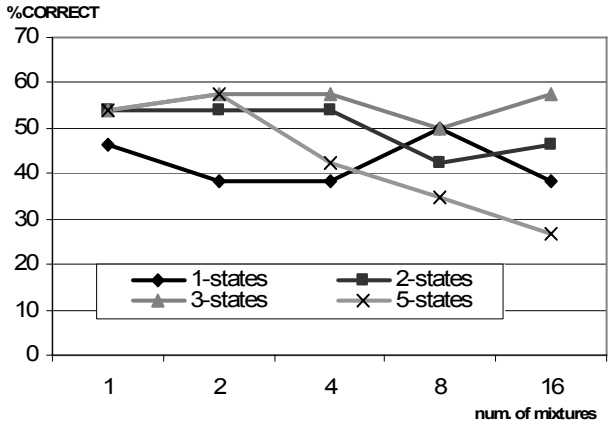

Figure 2. Correctnes for MFCC_0_D_A_Z HMM models

Our goal was to determine the model witch results in the best correctness during the detection task.

As we can see on Fig. 2, 3-states model reach better results then 5-states model. Decreasing of the correctness for the higher numbers of mixtures (5-state model) could indicate a lack of training data. Therefore, in the next experiments, we did not use higher number of states than 3.

On the Fig.3 there are depicted values of %Correct for the MFCC_D_A models with 1, 2 and 3 states in HMM model prototype and using a different number of PDF mixtures. The highest values of %Correct were for 3-states model.

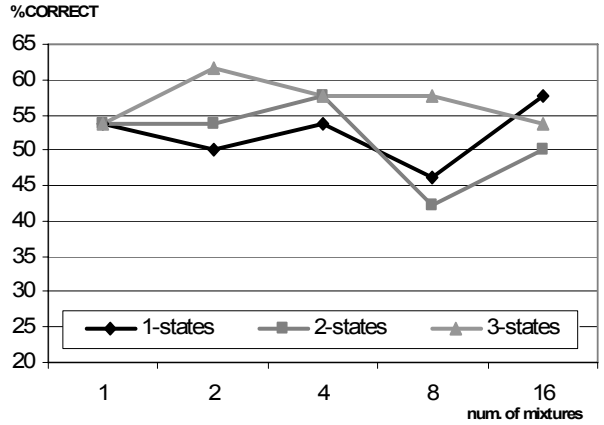

Figure 3. Correctness for MFCC_D_A HMM models

The different situation is for MFCC_E_D_A_Z models, where the best result gives the 1-state HMM model. But using energy for audio events detection is misleading, because in this experiment the audio events which are loud and which could lead the system to false alarms were not trained, or the amount of this data was small (laughing, car accelerating on the crossroad, strong wind or rain, thunders, etc.).

In the next phase, these events will be trained using the "other" model. Or there is also a possibility to train a special model for each of them, but only the dangerous events will produce the alarm after post-processing phase.

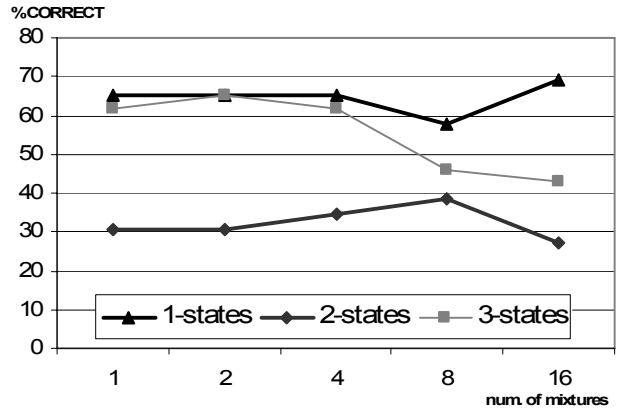

Figure 4. Correctness for MFCC_E_D_A_Z HMM models

The Table II. brings the comparison between the best results tests of tested parametrization feature sets. It could be said that models trained on such small database as used in this tests give very similar results around 65 %Correct. It seems

that models with energy parameter (energy or log energy) are more successful in the task of audio event detection. However there is the risk of false detections of audio events with similar energy characteristics, which should be evaluated in the post-processing phase or using special models, which should be trained to prevent the false alarms.

TABLE II. THE BEST %CORRECT VALUES FOR ALL MODELS

	Num. of states	Num. of mixtures	%Correct
MFCC_D_A	3	16	61,54
MFCC_0_D_A_Z	3	2	57,69
MFCC_E_D_A_Z	1	16	**69,23**

VI. CONCLUSIONS

The interesting applications of audio events detection systems are systems assigned for camera operators. The CCTV operators during the work watch the images from tenth of cameras with no sound. In this situation may occur that they are not able to catch all threats or unusual behavior. The intended audio event detection system could force the attention of the operator to the specified camera and replay the sound, which was detected as a special event (explosion, car crash, scream, gunshot, etc.), with some record before and after the event. Then the operator could check the recording and make a conclusion, if the event was a threat or any other unusual situation. But this process takes a lot of time for the operator, so it is important, that the system will not produce a significant amount of "false alarms". But this additional information could give the user of the audio event detection system additional information about the observed situation on the place covered by the camera, and also about events which are not in the camera visual field.

These preliminary results give us the knowledge for building the audio events database, and a database of acoustic backgrounds in different places – school, crossroad, railroad, city, village, bank, stations, bus stops, and so on.

In the future we would like to tune the algorithms of training and the feature extraction process. Also the small amount of the current events database reflected in the fluctuating results, so in the next step we are planning to record a real background database of the crossroads close to our laboratory, and simulated audio events in quiet environment, and then we will combine them and evaluate the detection algorithm.

Also, as described in the introduction, we will try to combine another feature sets in one vector and try to find a better feature vector for audio events detection. The feature extractions algorithms of not-MFCC feature sets are in simulating phase.

Next, also the transcription process of the collected audio events will be updated, and the events will be described more precisely and then the reference transcriptions should increase the resulted audio event models and the result computed from reference transcriptions of the test files, too. After that, the system of alarm post-processing will be evaluated, to decrease the false alarms occurrence.

The SVM tests are also in the progress, and also the results from SVM classifiers and HMM toolkit will be compared later. During SVM tests also a new high quality recordings of gunshots was captured in the quiet environment, and then we plan a different types of noise to be mixed to the test and training events database.

ACKNOWLEDGMENT

This work has been performed in the framework of the EU ICT Project INDECT (FP7-No.218086), and also supported by the Slovak Ministry of Education under research projects AV 4/2016/08 and MVTS COST2102/07.

REFERENCES

[1] Valenzise, G., Gerosa, L., Tagliasacchi, M., Antonacci, F., and Sarti, A. 2007. "Scream and gunshot detection and localization for audio-surveillance systems." In Proceedings of the 2007 IEEE Conference on Advanced Video and Signal Based Surveillance (September 05 - 07, 2007). AVSS. IEEE Computer Society, Washington, DC, 21-26.

[2] Young, S., Odell, J., Ollason, D., Valtchev, V., Woodland, P., Evermann, G., Hain, T., Kershaw, D., Moore, G.: "The HTK Book." Cambridge University 2009, online http://htk.eng.cam.ac.uk/prot-docs/htk_book.shtml

[3] Trancsriber tool website: http://trans.sourceforge.net/en/presentation.php

[4] EU ICT Project INDECT website: http://www.indect-project.eu/

[5] Temko, A., Malkin, R., Zieger, Ch., Macho, D., Nadeu, C., and Omologo, M.., "Acoustic Event Detection and Classification in Smart-Room Environment: Evaluation of CHIL Project Systems," in The IV Biennial Workshop on Speech Technology, November 2006.

[6] X. Zhuang, X. Zhou, T.S. Huang, M. Hasegawa-Johnson, "Feature analysis and selection for acoustic event detection", Proc. ICASSP, pp. 17-20, 2008.

[7] Baillie, M. and Jose, J.M. "Audio-based event detection for sports video." Lecture Notes in Computer Science, 2728 ., 2003, pp. 61-65. ISSN 1611-3349

[8] Zhang D., Gatica-Perez D., Bengio, S, McCowan, I.. "Semi-Supervised Adapted HMMs for Unusual Event Detection." Proceedings of the 2005 IEEE CVPR'05, Vol. 1, pp: 611 – 618, 2005, ISBN:1063-6919 , 0-7695-2372-2.

[9] Sha, F. and Saul, K. L. "Large margin Gaussian mixture modeling for phonetic classification and recognition." In Proceedings of ICASSP 2006, pages 265–268, Toulouse, France, 2006.

[10] Zieger, C. and Omologo, M.. "Acoustic event classification using a distributed microphone network with a GMM/SVM combined algorithm." Proc. Interspeech '08, Brisbane, Australia, September 2008.

[11] Chu, W., Cheng, J., Wu, J., and Hsu, J., "A study of semantic context detection by using SVM and GMM approaches", Proc. IEEE Int. Conf. on Multimedia and Expo, 2004.

[12] Zieger, Ch., Brutti, A., and Svaizer, P., "Acoustic Based Surveillance System for Intrusion Detection" In Proceedings of the 2009 IEEE Conference on Advanced Video and Signal Based Surveillance (September 02 - 04, 2009). AVSS. IEEE Computer Society, Washington, DC, pages 314-319.

[13] J.-L. Rouas, J. Louradour, S. Ambellouis, "Audio events detection in Public Transport vehicle," Proc. IEEE Intelligent transportation systems conference, Toronto, Canada, September 2006, pp. 73 -738.

[14] I. Trancoso, J. Portêlo, M. Bugalho, J. Neto, A. Serralheiro, "Training audio events detectors with a sound effect corpus," Proc. INTERSPEECH, Brisbane, Australia, September 2008, pp. 2546-2549.