A Comparative Study Using Two Classifiers For Hazardous Audio Event Classification

Tin Ei Kyaw
University of Computer Studies, Yangon
tineikyaw79@gmail.com

Abstract

The hazardous acoustic event classification system is presented and tested in threatening environments. The system is based on classified with Support Vector Machine (SVM), k Nearest Neighbor (kNN) and modeled with Genetic Regulatory Network (GRN). GRN is adopted as classification framework and greatly reduced input feature dimensions. Setting the results that have already reduced the inputs dimensions from GRN framework as inputs for SVM and kNN can correctly classify audio event with low computational time and cost. Comparative and classification tests are carried out using three kinds of input sets with SVM and kNN classifier. These input sets are original feature set, reduced dimension feature set by GRN and unique feature set. SVM applies as novel discriminative approach for dissimilarity measure in order to address a supervised sound-classification task and then shows good performance in the task of acoustic event classification. Selecting GRN in event classification system can not only reduces cost and effort but also aims to obtain high performance and accuracy in varying nature of environments.

Keywords: Acoustic Surveillance, Audio Features, Audio events, Classification tasks, k Nearest neighbor, Genetic Regulatory Network, Support Vector Machine.

1. Introduction

Audio signals which include speech, music and environmental sounds are important types of media. The problem of distinguishing audio signals [3] into these different audio types is thus becoming increasingly significant. A human listener can easily distinguish between different audio types by just listening to a short segment of an audio signal. However, solving this problem using computers has proven to be very difficult. Nevertheless, many systems with modest accuracy could still be implemented. For instance, audio event classification and detection [5] is broadly used in the entertainment industry, audio archive management, commercial music usage, surveillance [6], etc.

The system is very essential and useful application at important public places such as bank, subway, airport, mainline station, exhibition hall, stadium, market, etc in a noisy environment. At acoustic surveillance systems aim to detect abnormal situations based on video scenes while, in some conditions, it may be easier to detect and classify events using the audio information.

The area of surveillance system [1] and what kind of the event is mainly focused on detecting abnormal event based on the acquired audio information. The system offers a solution to detect abnormal audio events in continuous audio recordings in security of public places, industrial zones, markets, etc. The use of audio sensors in surveillance and monitoring applications [4] is becoming increasingly important. To know the abnormal situation, audio sensors are applied in distributed area at the place of video sensors because the former is cheaper and more convenient than the latter. Audio is useful especially in situations when other sensors such as video fail to detect the events. For example, when the object is occluded or is in the dark, the audio sensors can be more appropriate in

detecting the presence of objects assuming that the existence of the objects makes some sound. One of the major difficulties of an audio detection system is linked to the environmental noise that is often non stationary and that may be loud compared to the audio event to detect.

Like many other pattern classification tasks, audio classification [2] is made up of two main sections: a signal processing section and a classification section. The signal processing part deals with the extraction of features from the audio signal. The various methods of time-frequency analysis developed for processing audio signals, in many cases originally developed for speech processing, are used. The classification part deals with classifying data based on the statistical information extracted from the signals. Two different classifiers, k Nearest Neighbor (kNN) and Support Vector Machine (SVM), were trained and tested to classify audio signals into gunshot or explosion. The audio features used for classification were the Mel Frequency Cepestral Coefficients (MFCC), Zero Crossing Rates (ZCR) and Short Time Energy (STE).

In this situation, we use feature reduction technique called Genetic Regulatory Network (GRN) that is most successfully use in biological and medical domain. In signal processing environment, there are common practice to use feature reduction techniques like Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA), which map the features into a new vector space where the greatest variance by any projection of the data lies on the first coordinate, the second greatest variance lies on the second coordinate, and so on. Our proposed new dimension reduction technique (GRN) can be very much reduced feature dimension that is around 96% upon total feature vectors. The performance accuracy of the proposed system is compared with the two results from SVM and kNN classifier.

The rest of paper is organized as follows. Section 2 describes the related work. In Section 3 explains the structure of genetic regulatory network. Section 4 presents feature representation. Section 5 explains proposed system and Section 6 reports on experiments. Finally, Section 7 concludes the paper. Section 8 describes references.

2. Related Work

Zhu Liu and Qian Huang [7] exposed the experimental data from NBC Nightly News, sampled at the rate of 16 kHz with 16 bits per sample. Audio features are extracted at both frame level and clip level. Each frame consists of 512 samples and adjacent frames are shifted by 256 samples. Currently, four features are extracted from each frame. They are volume (root mean square of amplitude), zero crossing rate, pitch period, and energy ratios in sub band: 630 ~ 1720 Hz. Clip is a group of adjacent frames with the time span of 1 to 3 seconds, Nine clip level features are extracted: these features are 1) Non silence ratio (NSR). 2) Standard deviation of zero crossing rates (ZSTD). 3) Volume standard deviation (VSTD). 4) Volume dynamic range (VDR): the difference of maximum and minimum volume of a clip normalized by the maximum volume in that clip. 5) Volume undulation (VU). 6) 4 Hz modulation energy (4ME): the frequency component around 4Hz of the volume contour. 7) Smooth pitch ratio (SPR). 8) Non pitch ratio (NPR). 9) Energy ratio in sub band (ERSB) and then used two classification methods: simple hard threshold classifier and fuzzy classifier. Fuzzy logic based classifier outperforms the simple threshold classifier achieving an overall classification accuracy of above 94%.

Wei and group [8] proposed different event pairs are classified in their literature; they focused audio event and semantic context detection in video scenes are classified with SVM and GMM. Different events are engine and car-braking, gunshot and explosion. Overall accuracy in gunshot and explosion, engine and car-braking are precision of SVM is over 70% to 83% and GMM is 67% to 90% recall of SVM is 65% to 80% and GMM is 57% to 65%. In their survey, SVM found to be better than using GMM classifier. Features used are volume, band energy ratio, zero-crossing rate, frequency centroid, bandwidth, and 8-order MFCC. Advantages are robustness of detection performance and bridge

the gap between audio features and semantic concepts. In this system have two advantages: 1. Performance of semantic context detection is data-dependent 2. The feature values modeled by GMMs are too sensitive to the variations of different test data.

Lie and group [9] presented ten audio events(applause, laughter, cheer, car-braking, car crash, explosion, gun-shot, helicopter, plane, and siren) classified with Bayesian Network-based approach, HMM classifier and using features such as short-time energy, zero-crossing rate, band-energy ratio, brightness, bandwidth, MFCC, and two new features(sub-band spectral flux and harmonicity prominence) get high recall and precision. Domain focused on scenes and event detection at various TV shows and movies.

Aggelos et al [10] detected gunshot event vs. all other audio types using Bayesian Network and dynamic programming. 12 dimensional features such as MFCC1, MFCC2, MFCC3, MFCC1 (max), spectrogram-based feature, spectrogram, spectral roll of, 1st chroma-based feature, 2nd chroma-based feature, zero-crossing rate, energy entropy, pitch are used in this method. The experimental study of the paper reports that this method achieves overall precision with 78.8% and overall recall with 90.6%. The combination of decisions taken from an ensemble of one-vs-all BNs outperforms a single gunshots-vs-all BN by solving with dynamic programming and Bayesian Network.

Stavros and group [11] modeled acoustic surveillance of hazardous situation in metro station environment by GMM and HMM classifiers and using MFCC features set. This method reaches to highest average recognition accuracy of 93.05%. Three acoustic events considered to be classified are explosion, gunshot and scream.

Clavel and group [12] studied on sound detection produced by different gunshot. Shot and normal event classified with GMM and binary classifier. Features are short-time energy, first-eight MFCCs, spectral centroid and spectral spread. Result as false rejection rate falls from 18% to about 10% and can reduce the false rejection and false detection rates but false detection rate which, in the worst case, is reaching 43%.

3. Genetic Regulatory Network

Genetic Regulatory Network is used in biology that aims to understand the manner in which the parts of an organism interact in complex networks, and in medicine that aims at basing diagnosis and treatment on a systems level understanding of molecular interaction, both intra and inter-cellular. In biomedical system, use artificial genes at possible interaction with each other and get the link (strength) among them is also the structure of the network. By understanding the complex relations within this gene regulatory networks (GRN) can highlight inhibitory or excitatory interactions, as well as how intracellular or extracellular factors affect gene products. It is necessary to develop the models that adequately represent the classification tasks in audio events.

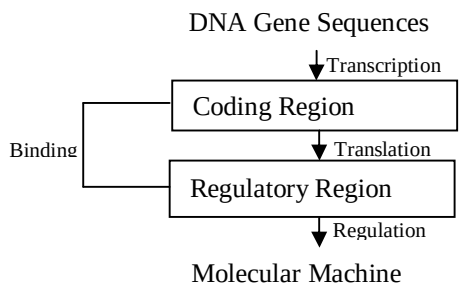

Figure 1(a). GRN network structure at Biological genes

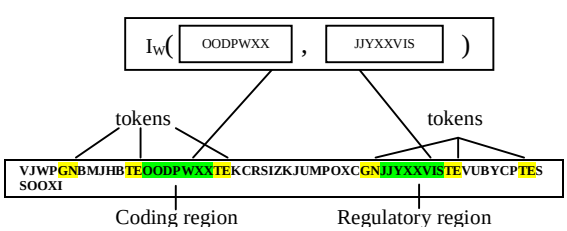

Figure 1(b). Region mark with tokens and calculate weight of gene with interaction map

In Figure 1(a) simplified the representation of transcriptional regulation between gene sequences. DNA gene sequences transcript at

coding region and translate at regulatory region. The artificial genomes in these two regions are binned with I_w (interaction map).

In Figure 1(b), the two regions from figure 1(a) are marked with two kinds of tokens 'GN' and 'TE' are the one with high occurrence in sequence. The possible pair of genes' weights is calculated with interaction map I_w. Depending on the one suitable weighted threshold then select the best combinations of genes and connect these pairs. Lastly get the network among these genes shown in Figure 1(c).

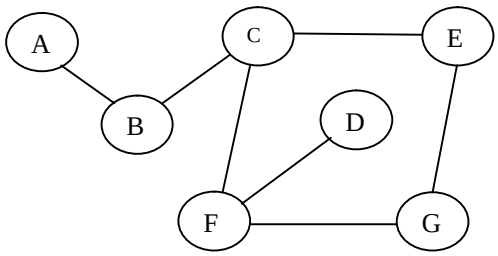

Figure 1(c). Gene regulatory network using 10 genes

In paper [13], Peter Durr et al proposed sleep/wake discrimination by using input features from both ECG and RSP data in biomedical domain. Used Analog Genetic Encoding (AGE) for the evolutionary synthesis of a neural classifier to classify sleep and wake condition. Achievement of similar performance to the hand-designed networks and accuracy of 88.49% and reduction the computational cost of almost 95% by reducing the input feature sets. In their proposed method used only 15 of the 736 input features comparing with the hand-designed network. So this can also reduce computation time and improve the energy efficiency of the mobile system.

4. Feature Representation

One important factor for event detection is the selection of suitable features that characterize original data adequately and can select a set of features from a larger set of available features. In the audio sequences, several audio features from time-domain amplitude and frequency-domain spectrogram are extracted and utilized. The crucial task for successful classification is using the right features. It is highly accurate and robust, and on the other hand, simple, efficient, and adequate for real-time implementation. It achieves excellent results in minimizing misdetection of voice, due to a combination of the feature choice in both time domain and frequency domain parameters.

At the first step, the audio stream is broken into a sequence of overlapping short-term frames and three features are extracted per frame. In our study, we use Short-time energy, zero-crossing rate and Mel-frequency spectral coefficients (MFCCs). Short-time energy (STE) is the total spectrum power of an audio signal at a given time and is also referred to loudness or volume.

Silence regions in sound play a very small role in event detection and hence can be removed. Here this fact is tested by removing silence regions of the waveforms by using Short Term Energy (STE) method. Basically the STE of the signal is computed and frames with energy less than a certain threshold are considered to be silence and discarded before MFCCs are extracted.

Short time energy can also be used to detect the transition from unvoiced to voiced speech and vice versa. The energy of voiced speech is much greater than the energy of unvoiced speech. Equation (1) for the STE is defined as

$$E_n = \sum_{-\infty}^{\infty} x^2(m)h(n-m) \qquad (1)$$

Where, h (n-m) is a windowing function.
m is the sample that window is centered on.
x is original signal and n is index

The zero-crossing rate (ZCR) of a frame is in equation (2). It is the number of times the audio waveform changes its sign in the duration of the frame.

$$\text{ZCR} = \frac{1}{N}\sum_{n=2}^{N}\left|\text{sign}(x(n)) - \text{sign}(x(n-1))\right| \qquad (2)$$

$$\text{where sign}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}$$

N is the number of frames,
n is the number of samples per frame,
x(n) and x(n-1) are values of signals.

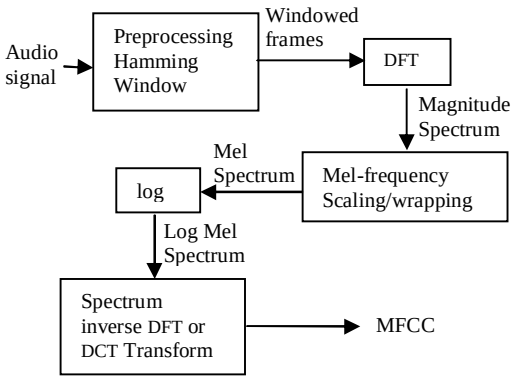

Figure 2. The Processing Steps to get MFCC

MFCCs are also increasingly finding uses in diverse areas of speech and audio signal processing application. In MFCC calculation, input signals are pre-processed with hamming windowing. The windowed frames are then transformed into transform domain with Discrete Fourier Transform (DFT). After getting magnitude spectrum, that are scaled by mel-frequency scales. Mel spectrums receiving from this stage are then changed using log function to obtain log mel spectrum. Finally, these spectrums are inversed with DFT or DCT to get MFCC coefficients. The following figure 2 shows the processing steps to get Mel-frequency spectral coefficients.

5. Proposed System

In proposed system framework, different audio wave files are pre-processed with framing and windowing functions to have unique processing environment. All audio streams are re-sampled to 22 KHz with 16 bits resolution. Each audio frame is segmented as 40 milliseconds, with 50% overlaps. In Figure 3 below shows the flow of proposed event classification of audio into two classes is presented with a block diagram.

In figure (3), the incoming audio wave file is duration of about 1to 3 seconds long and extracts 3 types of features from it as input for GRN model. First, we extract time-domain features (short-time energy (STE) and zero-crossing rate (ZCR)) and spectral domain feature set as Mel-frequency spectral coefficients (MFCCs) from input audio wave streams. Second, using with these feature values and then construct the event model with novel multi-layer evolutionary trained neuro-fuzzy recurrent network (ENFRN) [15] applied for the dimension reduction of the feature set. Lastly, SVM and kNN classify two events (gunshot or explosion) using the outcomes from GRN.

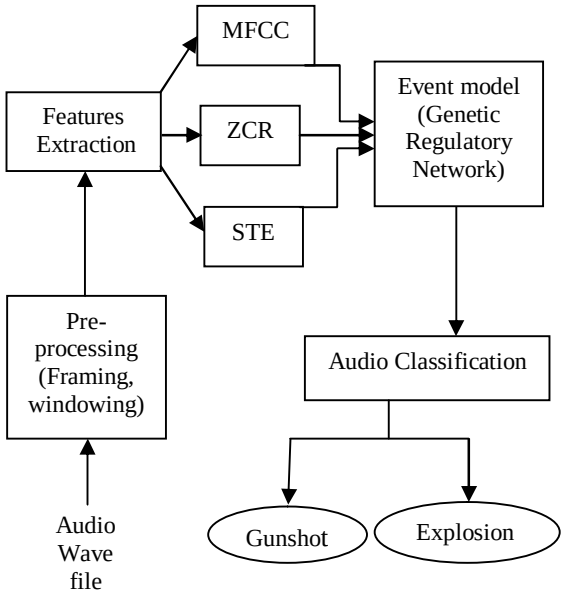

Figure 3. System Architecture of the proposed system

For audio events classification, generative (Genetic Regulatory Network) and discriminative (Support Vector Machine) approaches are investigated. Support vector machines (SVMs) have gained great attention and have been used extensively and successfully in the field of sound recognition. The main purpose of the paper is to efficiently characterize the environment in terms of threatening conditions while using acoustic information only. The outcome of the system is to help/warn authorized personnel to take the appropriate actions for preventing crime and/or property damage.

6. Experiments

The results from audio event classification become eagerly needed in many of cases. The system will improve the interaction between human and audio events and also influence on decision-making in genetic networks. There are many digital audio databases on the World Wide Web nowadays. All required audio data streams (sound of gunshot and explosion) can be received from the internet and CDs. All experiments will be implemented using the MATLAB. In this framework setting fuzzy IF-THEN rules and calculating composite scores then gives output (reducing the number of rule and outputs) as combination (by utilizing Particle Swarm Optimization (PSO)) of data within same rule. Additionally, fuzzy-based approaches are better candidates in dealing with the uncertainties of modeling noisy data and its fuzzy nature avoids noise-related problems. Furthermore RFNN combined with Particle Swarm Optimization (PSO) to capture the complex nonlinear dynamics of genetic regulatory networks.

The incoming audio wave file is about 1to 3 seconds long and extracts 3 features from it for GRN model. The first step is to construct GRN based on an Evolutionary trained Neuro Fuzzy Recurrent Network (ENFRN). Regulators for incoming features set can be calculated with their composite scores. Firstly, the initial ENFRN is created. Secondly, ENFRN optimization (e.g. reducing the number of rule and output nodes) utilizing Binary Particle Swarm Optimization (BPSO). Lastly, we get final ENFRN structure after training process of ENFRN utilizing Particle Swarm Optimization (PSO) and also calculate the following values in this stage: Mean Square Error (MSE) of every structure. Mean Score (MS) of every interaction. Composite scores (MSE and MS) of every interaction. From among the composite scores, we select the value less than 0.3. Then extract combinations of regulators, it is based on setting at initial maximum number of regulators.

After calculating, the number of input greatly reduced as expected. The following table shows as example record of GRN test results.

Table 1. Use percentage of feature sets that greatly reduced dimension after using GRN

Name (.wav)	Input Feature Sets	Total input before using GRN	Total inter-action after using GRN	Percentage Upon Input and total interaction
Bomb7	338x15	5070	89	1.716%
bomb drop	108x15	1620	103	6.358%
rifle	90x15	1350	43	3.18%
Dynamite	304x15	4560	76	1.67%
6-shots	97x15	1455	32	2.20%
auto_gun	38x15	2070	101	4.88%

Currently test upon different 225 audio wave files. After reading audio file, extracts features from it and sets as inputs to GRN. After calculating 225 audio files, gets total percentage of used input as 402.706 and average of about 3.987 inputs. So GRN can greatly reduce inputs as about 96% and also at computation time.

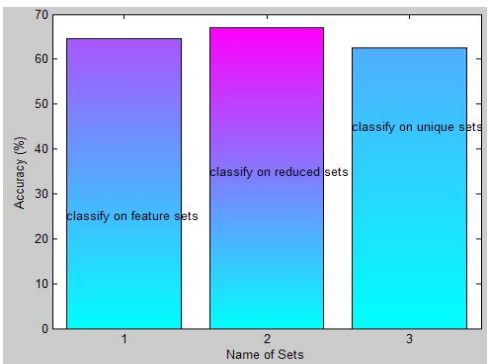

Figure 4. Average percentage of accuracy on three data values with SVM classifier.

After getting results from GRN, use SVM classifier with linear kernel. An internal fivefold cross validation on the training set in order to automatically select these parameters. Ten different fivefold cross validation experiments were conducted for evaluation. Based on 3 feature sets, the average accuracy rate is 64.65% by using 10 times running of SVM. Based on GRN outcomes, the average accuracy rate is 67.17% by using 10 times running of similar

SVM. Based on unique GRN outcomes, the average accuracy rate is 63.21% by using same process. These results are shown on figure 4. Therefore using GRN, the results are high accuracy and less execution time.

Figure 5 shows the feature values of input wave files, dimension reduced feature values by GRN and unique feature values upon after dimension reduction. Also shows the classification result of selected wave file upon three types of features sets.

File Name: g1.wav

Manual Tagging: gunshot

Classified Result: gunshot,gunshot,gunshot

Table 1. Original Feature Set

	11	12	13	14	15
24	1.5588	0.9498	0.3667	0.0120	0.1102
25	1.7073	1.0403	0.4017	0.0112	0.1034
26	1.4614	0.8905	0.3438	0.0093	0.1148
27	1.4673	0.8941	0.3452	0.0114	0.1352
28	0	0	0	0.0123	0.1284
29	1.1131	0.6783	0.2619	0.0123	0.1170
30	1.6266	0.9912	0.3827	0.0107	0.0898

Table 2. Reduced Feature Set by GRN

	1	2	3	4
22	1.6835	1.2893	0	0
23	0	0	0.4437	0
24	0	0	0	0
25	0	0	0	0
26	0	0	0	0
27	0	0	0	0
28	0	0	0	0
29	1.8281	1.4000	0	0

Table 3. Unique Feature Set

	1	2	3	4
7	0	1.4867	1.1386	3.4551
8	0.2018	0.4165	4.8869	0
9	0.9159	0.7014	0	4.2347
10	1.2505	2.4231	2.9994	0
11	1.5732	0	0	4.4762
12	1.6835	1.2893	0	0
13	1.8281	1.4000	0	0
14	1.9089	1.1302	0	3.1288

Figure 5. Sample record of original feature values, reduced values by GRN and make unique values upon GRN outcomes

In figure 6 provides the classified results of SVM and kNN classifier with original feature set, reduced dimension feature set by GRN and unique feature set. In that, SVM results is very much higher than kNN about 50% higher on test with reduced dimension feature set by GRN and around 15% higher on unique feature set.

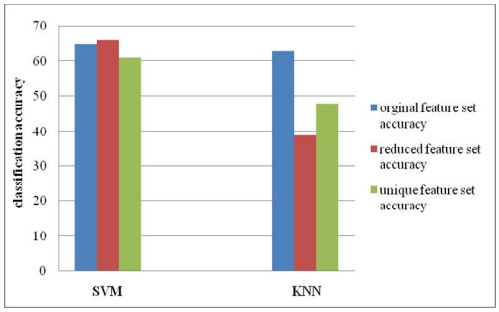

Figure 6. Comparison results with SVM and KNN classifier upon original feature, reduced feature and unique feature.

The advantages of the SVM technique can be summarized as follows:

-performs a good out-of-sample generalization and robust against changes of all vectors.

-provides a unique and good solution compared to Neural Networks.

-choice of suitable kernel (polynomial, RBF, linear), such as the Gaussian kernel, one can put more stress on the similarity.

-different discriminate functions can be achieved by using different kernel functions.

But also have disadvantages of SVM:

-requires large amount of labeled training data.

-the best choice of the kernel function is still a research problem.

-the testing speed also depends on the number of support vectors, so it could be slow.

Advantages of kNN are:

-it is robust to noisy training data.

-it is effective if the training data is large.

Disadvantages also have kNN:

-need to determine value of parameter K.

-computation cost is quite high because we need to compute distance of each query instance to all training samples.

7. Conclusion

In this paper, a robust audio-based audio classification system was introduced. This system represents an essential building block of a complete acoustic surveillance system. The

performance of the proposed framework can make more applicable to any problem of audio event classification. In real network analysis, the present work is expected to be succeeded in finding several reasonable audio events as compare to the other existing methods. In all the cases, even with the presence of noise, the current work has been designed to meet almost all the correct detections and classifications. Currently, the system is used 225 input data as current test bed and run five-fold cross validation. More labeled training data will be required in order to get higher accuracy rate. The proposed audio event classification system offer accuracy, less execution time and better performance. Proposed method intends to support higher user satisfaction with lower computational cost. According to the literature, any system has not been fulfilling the user requirement completely.

8. References

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

[2] H. Meindo and J.Neto, "Audio Segmentaion, Classification and Clustering in a Broadcast News Task", in Proceedings ICASSP, Hong Kong, China, 2003

[3] J. R. Deller, J.H.L. Hansen and J.G. Proakis, "*Discrete-Time Processing of Audio Signals*", IEEE Inc. 2000

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

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

[6] A. Harma, M.F. McKinney, J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *IEEE International Conference on Multimedia and Expo*, 2005

[7] Z.Liu, Q.Huang. "Classification of Audio Events in Broadcast News" AT&T Labs – Research.

[8] W.Chu, W. Cheng, J. Wu, J. Y. Hsu " A Study of Semantic Context Detection by Using SVM and GMM Approaches" *IEEE International Conference on Multimedia and Expo* (ICME), 2004

[9] L. Lu, R. Cai, A. Hanjalic "*Towards a Unified Framework for Content-based Audio Analysis*" in Microsoft Research Asia, Department of Computer Science and Technology, Tsinghua University, Beijing, P.R. China, ICASSP 2005

[10] A. Pikrakis, T.Giannakopoulos, S. Theodoridis "*Gunshot Detection in Audio Streams from Movies by means of Dynamic Programming and Baysian Networks*" in Department of Informatics University of Piraeus, Greece, ICASSP 2008

[11] S.Ntalampiras, I. Potamitis, N. Fakotakis "*On Acoustic Surveillance of Hazardous Situation*" in Department of Electrical and Computer Engineering, University of Patras, Greece, ICASSP 2009

[12] C. Clavel, T. Ehrette , G. Richard "*Events Detection for an Audio based Surveillance System*" *in Thales Research and Technology* France , 2005 IEEE

[13] D¨urr, W.Karlen, J.Guignard, C.Mattiussi, and D.Floreano "*Evolutionary Selection of Features for Neural Sleep/Wake Discrimination*" in Laboratory of Intelligent Systems, Switzerland, Volume 2009

[14] E.O.Dijk "*Analysis of Recurrent Neural Networks with Application to Speaker Independent Phoneme* Recognition" Department of Electrical Engineering

[15] I.A. Maraziotis, A.Dragomir, D.Thanos "*Gene Regulatory Networks Modeling using a Dynamic Evolutionary Hybrid*" BMC Bioinformatics 2010