Audio Events Classification in Threatening Conditions at Surveillance System using Genetic Regulatory Network

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

Abstract

The threatening environments in acoustic surveillance system at important public places in a noisy environment deals with audio events detection is essential and useful application. At surveillance system intends to detect abnormal situations based on visual scene while, in some conditions, it may be easier to classify events using the audio information. Audio events classification for threatening environment through Genetic Regulatory Network (GRN) is considered. GRN is adopted as classification framework and greatly reduced input dimensions. Thus using the results from GRN framework as inputs for Support Vector Machine (SVM) can correctly classify audio events such as gunshot or explosion with low computational time and complexity. SVM applies as novel discriminative approach for a supervised sound and event classification task. 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.

Key Words: Audio events, Audio Features, classification tasks, Genetic Regulatory Network, Support Vector Machine.

1. Introduction

The area of surveillance system and what kind of the event is mainly focused on detecting abnormal events based on the acquired audio information. The system provides a solution to classify abnormal audio event in continuous audio recordings in security of public places such as bank, subway, airport, mainline station, exhibition hall, stadium, market, etc. The correctness of classification depends on different conditions and the less false rejection rate which is critical in surveillance applications. The use of audio sensors in surveillance and monitoring applications 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 sensors are very useful than other sensors such as video sensors fail to sense the event in some condition. For example, when the object is in the dark or not easy to sense, the audio sensors can be more suitable in detecting the existence of objects that the objects makes some sound. Audio event 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.

In the area of surveillance system in [1] consists of a large number of cameras placed in distributed areas and connected to a central control room. This approach gives several advantages such as: a) the computational needs are very low; b) the illumination conditions of the space to be monitored and possible occlusion do not have an immediate effect on sound. Previous approaches upon the matter of acoustic monitoring include cases such as in [2] where a gunshot detection system is presented based on features derived from the time-frequency domain and Gaussian Mixture Model (GMM) classifier. During the training phase, they use different SNRs as 10% and 5% false rejection and false detection rate respectively. In [3] represent an emotional recognition scheme for public safety. The main objective is fear vs. neutral classification and by using different models for voiced and unvoiced speech they reach 30% error rate. In [4] they report on a parallel classification system based on GMMs for ambient noise, scream and gunshot sounds discrimination. After a feature selection algorithm they result in 8% false rejection rate and 90% precision. An audio-based surveillance system in office environment is described in [5]. The model of event detection is constructed with both supervised and k Means data clustering. In [6] audio data recorded using simultaneously 4 microphones are classified with two different approaches - GMM and Support Vector Machine (SVM) for shot detection in a railway environment. The work of Wilpon et al [7] regarded keyword spotting. In this model the sounds which present highly non-stationary properties (it

contains horns, opening/closing doors, people talking in the background, train movement etc). Extensive experimentation regarding the best set of features is carried out by feature selection process.

In this paper, we focus on classifying event such as class 1 for gunshot and class 2 for explosion using the audio streams. In order for such an implementation will be useful and practical if it must offers classification accuracy as high as possible under noisy conditions. Our approach is basically motivated by the fact that sound provides information that is hard or impossible to obtain by any other means. On top of that, such a method comprises a low cost and relatively easy during setup, solution. This article concentrates on detecting atypical two sound events (gunshot or explosion). The paper is continued to organize as follows. Section 2 describes 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

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 band energy ratio, volume, frequency centroid, zero-crossing rate, bandwidth, and 8-order MFCC. Advantages are robustness of detection performance and bridge the gap between audio features and semantic concepts. This system has 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 (laughter, applause, cheer, car-braking, explosion, car crash, gun-shot, siren, helicopter and plane) 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(harmonicity prominence and sub-band spectral flux) 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 and 2nd chroma-based feature, zero-crossing rate, pitch and energy entropy 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 decision 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 Hidden Markov Model (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 in the worst case false detection rate is reaching 43%.

3. Genetic Regulatory Network

Genetic Regulatory Network is used in biology, it aims to understand the manner of an organism interact in complex networks, and in medicine, it aims at diagnosis and treatment on a system's level understanding of both intra and inter-cellular interaction. 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. The complex relations within gene regulatory network can highlight inhibitory or excitatory interactions and how intracellular or extracellular factors affect gene. It is necessary to develop the model that adequately represents the classification tasks in audio events.

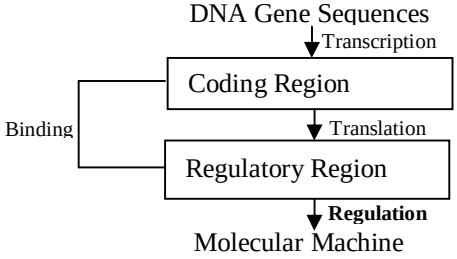

Figure 1(a). GRN network structure at Biological genes

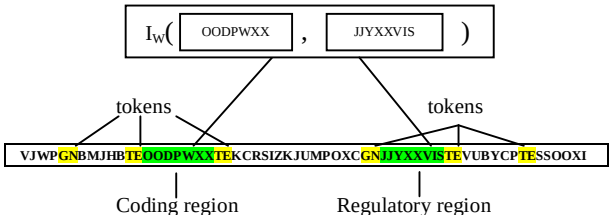

Figure 1(b). Region mark with tokens and calculate weight of gene with interaction map

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 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. Then get the best combinations of genes and connect these pairs. Lastly get the network among these genes shown in Figure 1(c).

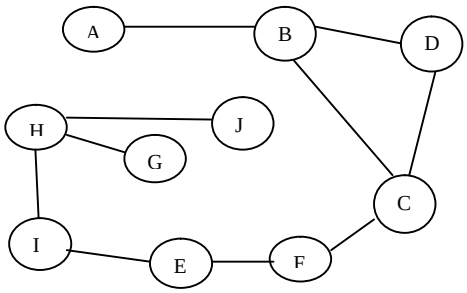

Figure 1(c). Gene regulatory network using 10 genes

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) and neural classifier to classify sleep and wake condition. Their system is achieved 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

The selection of right and suitable features is important for event classification that characterizes original data adequately and can select a set of features from a larger set of available features. In the audio sequences, several audio features are extracted and utilized from time-domain amplitude and frequency-domain spectrogram. The crucial task for successful classification is using the right features. It is highly accurate, robust and efficient for real-time implementation. The features representation in both time and frequency domain parameters achieves good results in minimizing misclassification.

At the first step, the audio wave files are 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, loudness or volume of an audio signal. In event detection, silence regions in sound play a very small role and hence can be removed by using STE method. The energy less than a certain threshold are considered to be silence and discarded before MFCCs are extracted. STE can also be used to detect the transition from unvoiced to voiced speech and vice versa. The energy of voiced speech is always greater than the energy of unvoiced speech. The 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 and m is the sample that window is centered on.

The zero-crossing rate (ZCR) is defined as in equation (2). ZCR calculates the number of times of the audio waveform changes in the duration of the frame.

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

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

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

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.

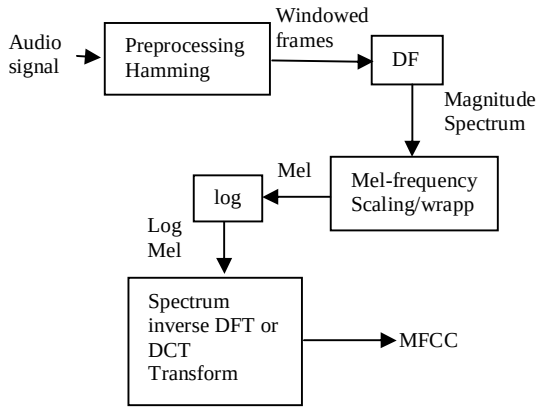

Figure2. The Processing Steps to get MFCC

5. Proposed System

In proposed system framework, audio signals are pre-processed to have unique processing environment. All audio streams are re-sampled to 22 KHz with 16 bits resolution. Each audio frame is 40 milliseconds with 50% overlaps. 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 various time-domain features (short-time energy and zero-crossing rate) spectral domain feature set as Mel-frequency spectral coefficients (MFCCs) from input audio streams. Second, neuro-fuzzy recurrent network (ENFRN) [15] applied to the problem of GRN reconstruction. In this framework setting fuzzy IF-THEN rules and 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. Lastly, SVM classify two events using the outcomes from GRN.

Recurrent neural networks (RNNs) and recurrent fuzzy neural networks (RFNNs) [14] are some of the most effective approaches regarding spatial or temporal problems by depending self loops and backward connections to their structures. Fuzzy-based approaches are better in the uncertainties of modeling noisy data and avoided noise-related problems.

Audio events classification system used Genetic Regulatory Network (GRN) and Support Vector Machine (SVM) approaches for generative discriminative respectively. SVM have been used successfully at the sounds (events) classification. The main purpose of this paper is to efficiently characterize threatening conditions using acoustic information only. For preventing crime and/or property damage, the outcome of the system is to warn authorized personnel to take the appropriate actions. In Figure 3 shows the block diagram of the proposed event classification.

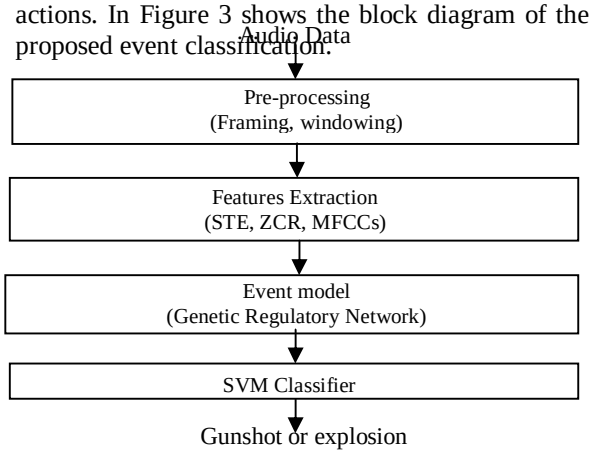

Figure3. System Architecture of the proposed system

6. Experiments

The results from audio event classification become eagerly needed in many of cases. GRN based classifier will be employed in detecting gunshot and explosion for surveillance system at important public places in a noisy environment. The system can improve the interaction between human and audio events and also influence on decision-making in genetic networks. To represent the acoustic events in an environment, a set of signal characteristics is employed. All required audio data streams (sound of gunshot and explosion) can be received from the internet and CDs. In order to detect the events from these signals nature, a classifier is formulated using genetic regulatory network. GRN run as the based classifier for the whole process. All experiments will be implemented using the MATLAB. GRN can be used to design robust audio event classification system. The audio event classification system can be expected to offer accuracy, correctness, less execution time and better performance. The performance of the proposed framework is measured with calculating precision and recall.

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). Calculate regulators for each one of the incoming audio wave file based on composite scores. The initial ENFRN is created. Optimize of ENFRN (e.g. reducing the number of rule and output nodes) utilizing Binary Particle Swarm Optimization (BPSO). Get final ENFRN structure after training process of ENFRN utilizing Particle Swarm Optimization (PSO). Moreover, ENRFN employed with Particle Swarm Optimization (PSO) to capture the complex nonlinear dynamics of genetic

regulatory networks. The system also calculates the following values: 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, 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 inputs greatly reduced as expected. For the current test bed, 225 audio files, the average percentage of 3.987 is get so that 96% feature sets can be reduced. The following table shows as example record of GRN test results.

Table1. Use percentage of feature set and that greatly reduced dimension after using GRN

Name (.wav)	Input Feature Sets	Total input	Total inter-action	Percentage Upon Input and total interaction	Run time (minutes)
bomb1	84x15	1260	38	3.02	10
bomb2	97x15	1455	17	1.17	11
Explosion (1)	195x15	2925	237	8.102	3
Dynamite	304x15	4560	76	1.67	16
6-shots	97x15	1455	32	2.20	4
auto_gun	38x15	2070	101	4.88	5

The necessary values are extracted depend upon the index of regulator numbers from ENFRN calculation. This matrix contains a few numbers of values comparing with original feature matrix. After getting reduced matrix and then unique this matrix. SVM used these 3 matrixes (Feature vector matrix, reduced matrix and unique matrix) with linear kernel in order to classify class 1 for gunshot or class 2 for explosion.

Ten different two fold cross validation experiments were conducted for evaluation. Based on the 3 matrix sets, the average accuracy rate 64.65% is get by using 10 times running of SVM. Based on GRN outcomes, the average accuracy rate 67.17% is get by using 10 times running of similar SVM. Based on unique GRN outcomes, the average accuracy rate 63.21% is also get by same process. The proposed framework's performance can be measured with precision and recall at following table 2. The next following bar chart shows 3 results from SVM and GRN based SVM classifier.

Table2. Precision and recall of events using SVM classifier based on GRN outcomes

Event	Precision	Recall
Gunshot	83.33	77.21
Explosion	85.89	74.32

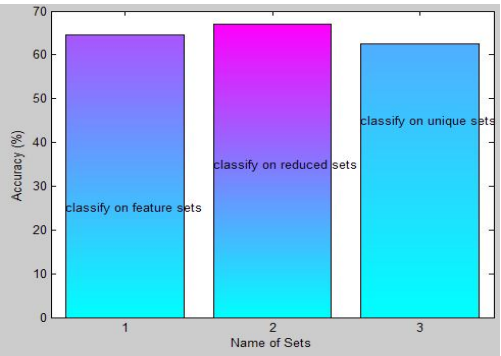

Figure4. Getting 3 results based on SVM and GRN based SVM classifier

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

SVM perform a good out-of-sample generalization, if the parameters *C* and *r* are appropriately chosen. Even when the training sample may have some bias, SVMs can be robust. Since the optimality problem is convex, SVM gives a unique and good solution. It is independent of the feature dimensionality. Different discriminate functions can be obtained by using different kernel functions. Different SVM classifiers can be constructed by using polynomial, RBF or linear kernels. But also have disadvantages of SVM:

As a supervised classifier, it requires large amount of labeled training data. If the large number of support vectors, system testing speed can be slow.

7. Conclusion

The system represents an essential building block of a complete acoustic surveillance system. GRN and SVM can be used as the optimization classifier in an optimized classification framework. Proposed method intends to provide higher user satisfaction with lower computational cost. The performance of the proposed framework will makes more applicable to any problem of audio event classification. Currently, the two fold cross validation is applied as only 225 input data is used for the system. More labeled training data will be required in order to get higher accuracy rate. 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. Its main aim is to classify on time the sensed situation and deliver the necessary information to dedicated person. The present work is expected to be succeeded in classifying audio events. 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] 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.

[3] C. Clavel, I. Vasilescu, L. Devillers, G. Richard and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Communication*, Elsevier, pp. 487-503, 2008.

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

[5] 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.

[6] J.-L. Rouas, J. Louradour and S. Ambellouis, "Audio Events Detection in Public Transport Vehicle," in *IEEE Intelligent Transportation Systems Conference*, Toronto, September 2006.

[7] J. G. Wilpon, L. R. Rabiner, C.-H. Lee and E. R. Goldman, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 1870-1878, November 1990.

[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.