Classification of Audio Events in Surveillance System using Genetic Regulatory Network

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

Abstract

The surveillance system at important public places in a noisy environment deals with audio events detection is essential and useful application. At surveillance systems aiming to detect abnormal situations based on visual clues while, in some situations, it may be easier to detect and 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 and explosion with low computational time and complexity. Selecting GRN in event detection system can not only reduces cost and effort but also aims to obtain high performance and accuracy in varying nature of environments.

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 offers a solution to detect abnormal audio events in continuous audio recordings in security of public places such as bank, subway, airport, mainline station, exhibition hall, stadium, market, etc. The robustness of detection against variable and adverse conditions and the reduction of the false rejection rate which is important 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 is useful especially in situations when other sensors such as video fail to reliably 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.

In the area of surveillance system in [1] consist of a large number of cameras distributed in an area and connected to a central control room. This approach offers several advantages such as: a) low computational needs, b) the illumination conditions of the space to be monitored and possible occlusion do not have an immediate effect on sound. Previous approaches on the subject 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 GMM classifier. They use different SNRs during the training phase for achieving 10% and 5% false rejection and false detection rate respectively. In [3] they present an emotional recognition scheme for public safety. Their 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 building a parallel classification system based on GMMs for discrimination of ambient noise, scream and gunshot sounds. After a feature selection algorithm they result in 90% precision and 8% false rejection rate. An audio-based surveillance system in a typical office

environment is described in [5]. The background noise model is continuously updated for serving interesting event detection while both supervised and k-means data clustering are inspected. In [6] audio data recorded using simultaneously 4 microphones are classified with two different approaches - GMM and 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 includes 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 detecting a set of events such as gunshot and explosion using the audio streams. In order for such an implementation will be useful and practical it must offer very low false alarm rate while keeping detection 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 and explosion). The rest of this 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.

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 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 network (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.

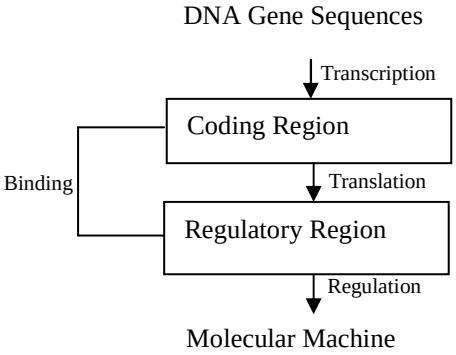

Figure 1(a). GRN network structure at Biological genes

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).

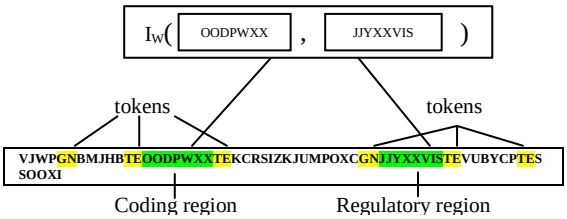

Figure 1(b). Region mark with tokens and calculate weight of gene with interaction map

In Figure 1(b), the two regions are marked with tokens ‘GN’ and ‘TE’. 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).

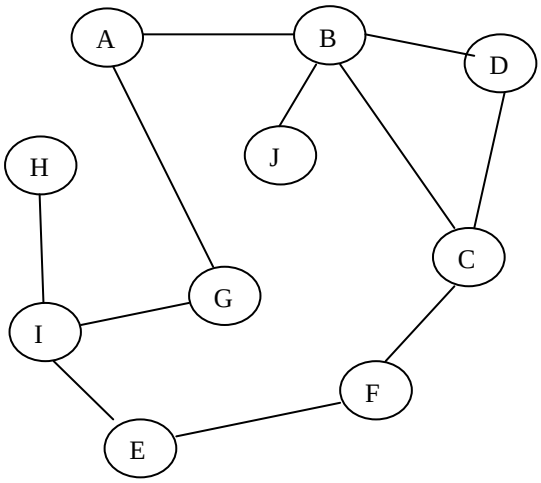

Figure 1(c). Gene regulatory network using 10 genes

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) 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. The equation (1) for the STE is defined as

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

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

Mel-frequency spectral coefficients (MFCC) 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 (MFCC).

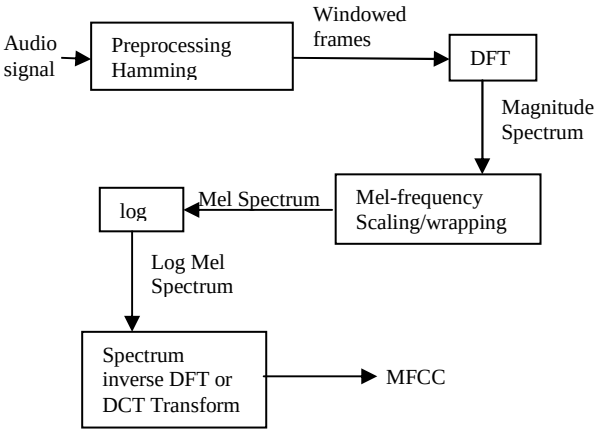

Figure 2. The Processing Steps to get MFCC

5. Proposed System

In proposed system framework, audio signals are pre-processed to have unique processing environment. We define an audio frame to be a fixed size audio segment which is extracted from the continuous audio stream. All audio streams are re-sampled to 22 KHz with 16 bits resolution. Each audio frame is of 40 milliseconds, with 50% overlaps. First, we extract various time-domain features (short-time energy and zero-crossing rate) spectral domain feature set as Mel-frequency spectral coefficients (MFCC) from input audio streams. Second, in proposed dimension reduced with novel multi-layer evolutionary trained neuro-fuzzy recurrent network (ENFRN) [15] applied to the problem of GRN reconstruction. Related to that, advantage of this approach is that it overcomes the need of prior data discretization, a characteristic of many computational methods which often leads to information loss. In this framework initially set fuzzy IF-THEN rules and composite scores. And then calculate the combinations of pair with desired composite score and number of regulation; from this extract only as less as possible combinations lower than the threshold (suitable desired composite score) as input for SVM classifier. Calculate the combination of regulators by Particle Swarm Optimization (PSO) of data within same rule. Lastly, SVM classify two events using the outcomes from GRN.

Some of the most effective approaches towards problems regarding temporal information processing are the recurrent neural networks (RNNs) and recurrent fuzzy neural networks (RFNNs) [14]. Recurrent networks, in general, can deal with temporal and spatial/ temporal problems by adapting self loops and backward connections to their topologies and structures, both of which are used to memorize past information. 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 Recently, RFNN combined with Particle Swarm Optimization (PSO) to capture the complex nonlinear dynamics of genetic regulatory networks. For audio events classification, generative GRN (Genetic Regulatory Network) and discriminative SVM (Support Vector Machine) approaches are investigated. The main purpose of this 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. In Figure 3 the flow of proposed event detection is presented with a block diagram.

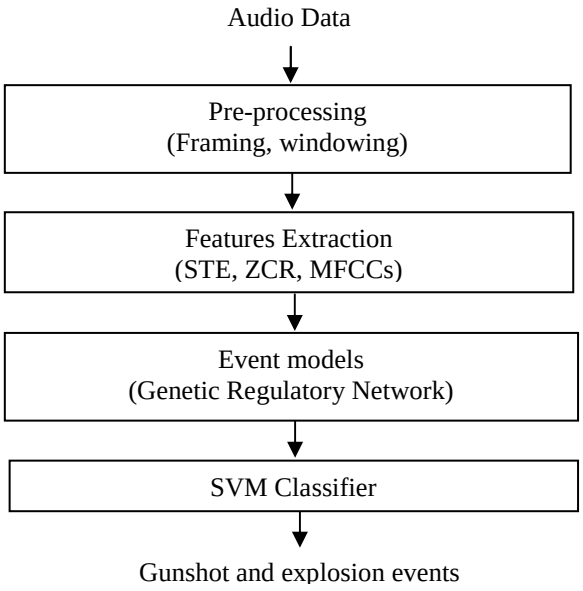

Figure 3. System Architecture of the proposed system

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 will 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) will be received from the internet and CDs. In order to detect the events from these signal nature, a classifier is formulated using genetic regulatory network. In the classification stage, SVM classifier is designed through GRN and will be discriminated two events. GRN run as the based classifier for the whole process. All experiments will be implemented using the MATLAB. According to the literature, any system has not been fulfilling the user requirement completely. By using GRN classifier upon any audio event will improve the accuracy and performance of the whole system. GRN will be used to design robust audio event classification system so we will get efficient event detection system. The audio event classification system will be expected to offer accuracy, correctness, less execution time and better performance. The performance of the proposed framework will be measured calculating precision and recall.

7. Conclusion

In this paper, a robust audio-based audio detection system was introduced. This system represents an essential building block of a complete acoustic surveillance system. It is based on a SVM classifier to classify the audio events and in order to reduce the false rejection and false detection rates. We show that the noise level of the training database has a significant impact on the performance of the system which allows to selecting the most appropriate noise level of the training database for a targeted false rejection rate. The performance of the proposed framework could make it more applicable to the 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. Its main aim is to identify on time the sensed situation and deliver the necessary warning messages to an authorized person. The proposed methodology is practical, can operate in real-time and elaborates on two abnormal sound events. The recognition results will achieve under a variety of background environmental noise.

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