Audio Event Classification using the Genetic Regulatory Network

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

Abstract

The massive multimedia data, including video, audio, and text can be used by users in content browsing, retrieval, classification management. In this study, audio event classification through a Genetic Regulatory Network (GRN) is considered. Genetic Regulatory Network is adopted as classification framework to detect audio events such as gunshot, explosion for surveillance system at important public places in a noisy environment. Selecting GRN for designing a classifier among the different classifier in event detection system can not only reduce cost and effort with high performance accuracy in varying nature of environments.

1. Introduction

Research in the area of automatic surveillance systems is mainly focused on detecting abnormal events based on the acquired video information [1]. Current implementations typically consist of a large number of cameras distributed in an area and connected to a central control room. While this kind of analysis provides valuable information we concentrate on detecting atypical events by exploiting only the acoustic modality. 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. Last but not least, 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 proposed system framework exploits three features (short-time energy, MFCCs and zero-crossing rate) combined with the GRN classifier. The main goal of this paper is to efficiently characterize the acoustic environment in terms of threatening/non-threatening conditions. The outcome of the system is to help/warn authorized personnel to take the appropriate actions for preventing crime and/or property damage. 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. In this article concentrate on detecting atypical two sound events (gunshot and explosion). The current methodology is inspired by the work of Wilpon et al [7] regarding keyword spotting. In our model the sounds which presents 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 to be included in the feature extraction process will be carried out. The rest of this paper is organized as follows. Section 2 describes the related work. In Section3 explains genetic regulatory network. Section 4 presents feature representation. Section 5 explains proposed system and Section 6 reports experiments. Finally, Section 7 concludes the paper.

2. Related Work

Wei-Ta Chu and group [8] proposed different event pairs are classified in their literature; engine and car-braking, gunshot and explosion at video scene. Signals are employed by SVM, GMM classifier. Overall accuracy in gunshot and explosion, engine and car-braking using SVM are found to be better than using GMM classifier.

Lie Lu and group[9] presented ten audio events(applause, laughter, cheer, car-braking, car crash, explosion, gun-shot, helicopter, plane, and siren) classified with 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.

Aggelos Pikrakis et al., [10] detected gunshot event using Baysian Network and dynamic programming. 12 dimensional features such as MFCC1, MFCC2, MFCC3, MFCC1 (max), spectrogram-based, 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%.

Stavros Ntalampiras and group [11] modeled acoustic surveillance of hazardous situation 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.

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. To represent the acoustic events in an environment, a set of signal characteristics is employed. In order to detect the events from these signal nature, a classifier is formulated using genetic regulatory network.

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. It is necessary to develop the models that adequately represent the classification tasks in audio events.

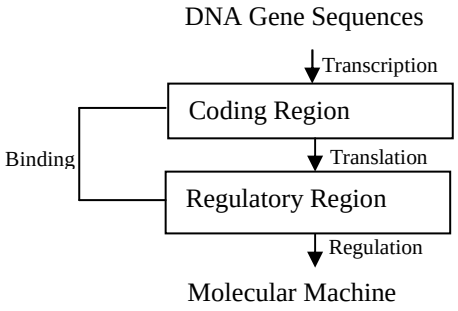

Figure 1(a). GRN network structure at Biological genes

In Figure 1(a) simplified representation of transcriptional regulation. DNA sequences transcript at coding region and translate at regulatory region. The artificial genomes in these two regions are binned with interaction map I_w in Figure 1(b). In Figure 1(b) the two regions are marked with tokens 'GN' and 'TE'. The possible pair of genes' weights are 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).

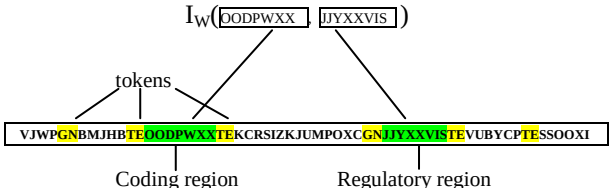

Figure 1(b). Region mark with tokens and calculate weight of gene with interaction map

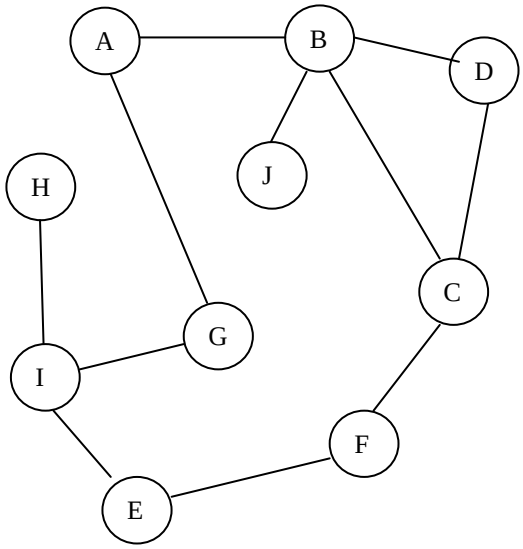

Figure 1(c). Gene regulatory network using 10 genes

In paper [12], Peter Durr et at 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. Can greatly reduced set of inputs via reducing computation time and improving 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 a first step, the audio stream is broken into a sequence of non 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.

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 short time energy for a sampled signal where h(n-m) is a windowing function. For simplicity a rectangular windowing function is used as defined in following equation

$$H(n) = 1 \qquad 0 \le n \le N-1 \qquad (2)$$
$$= 0 \qquad \text{Otherwise}$$

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 zero-crossing rate (ZCR) of a frame is defined as in equation (3) the number of times the audio waveform changes its sign in the duration of the frame.

$$ZCR = \frac{1}{N}|sign(x(n)) - sign(x(n-1))| \qquad (3)$$

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

N is the number of samples per frame

5. Proposed System

In proposed system framework, audio signals are pre-processed to have unique processing environment. All audio streams are re-sampled to 16 KHz with 16 bits resolution. Each audio frame is of 25 milliseconds, with 50% overlaps. Suitable acoustic features such as Short-time energy, zero-crossing rate and Mel-frequency spectral coefficients (MFCCs) are extracted.

In the classification stage, a classifier is designed through a genetic regulatory network. Two events: gunshot and explosion will be discriminated using this GRN classifier. In Figure 2, the flow of proposed event detection algorithm is presented with a block diagram.

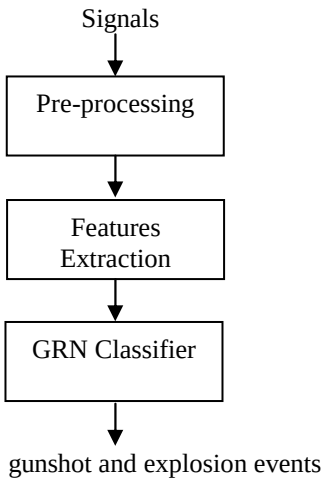

Figure 2. System Architecture of the proposed system

6. Experiments

The results from audio event classification become eagerly needed in many of cases. GRN classifier will be employed in detecting gunshot and explosion for surveillance system at important public places in a noisy environment. The experiments will

be conducted with less computational cost and less computational effort by using three distinct audio features. The system will improve the interaction between human and audio events and also influence on decision-making in genetic networks. GRN run as the based classifier for the whole process. All required data input will be received from the internet. With acquired audio collection, a corpus with varying event will be constructed. 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

The performance of the proposed framework could make it more applicable to the any problem of audio event classification. As GRN is used as the optimization classifier, a classifier with an optimized network structure is obtained. 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 regulations. 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. Thus, along with some future enhancements this work will boost the wide range of various system identification and acoustic monitoring research.

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" in 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] 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