Simulation of Audio Classification for Event Detection Using Adaptive Neuro Fuzzy Inference System for a Public Transport Vehicle

Cristina P. Dadula[12] and Elmer P. Dadios[1]
[1]Gokongwei College of Engineering
De La Salle University
Manila, Philippines
[2]Mindanao State University-General Santos City
cristina_dadula@dlsu.edu.ph
elmer.dadios@dlsu.edu.ph

***Abstract*—This paper presents the simulation of audio surveillance system in a public transport vehicle that detects event like screams and gunshots by classifying signals as normal or in crisis condition using adaptive neuro fuzzy inference system (ANFIS). Audio signals were divided into frames and represented by its feature. Feature is extracted using mel frequency cepstral coefficients. Eight audio files were used in the simulation where half of the files represent the normal condition and another half denotes the crisis condition. One hundred data sets from each file were used in training and another 100 data sets from each file were used in validation. The fuzzy inference system was created using the data centers produced using subtractive clustering given the range of influence. Different values of range of influence near the default value of 0.5 were simulated in order to observe the accuracy of the system. The system's validation accuracy is greater than 82% except for one file under normal condition that simulated a very high speed bus passing by.**

***Index Terms*— audio classification, ANFIS, audio event detection, adaptive neuro fuzzy inference system**

I. Introduction

Event detection in a surveillance system is an important feature of an intelligent transport system that monitors safety and security concerns. Surveillance systems typically use audio, video or both. One aspect of audio surveillance monitors possible occurrence of a crime that may include audio events such as screams and gunshots. In the case that at least one of the event occurs the system may locate the position of the source and steer the video camera to monitor further details.

Audio classification techniques for event detection are important processes in the analysis of audio signals. Various methods to classify signals have been developed. The two main steps of audio classification are feature extraction and classification. Feature extraction is done in order to reduce the amount of data to be processed. Features can be extracted from audio signals in time, frequency of coefficient domains. One of the most well known feature of an audio signal in coefficient domains is mel frequency cepstrum coefficient (MFCC). MFCC has been used as feature vector in audio classification and speaker recognition. Buket D. et. al. (2012) classifies non-speech normal and abnormal audio events using MFCC and artificial neural network classifiers. Normal audio events include engine noise and rain while abnormal audio events include glass breaking, dog barking, scream, gunshots.

The study of Liu, et. al (2002) classify audio clips using content-based technology using fuzzy logic system. Features were extracted from sound signals. Features were extracted from time, frequency, and coefficient domains. Time domain features include root mean square, silence ratio, zero-crossing. Frequency domain features include spectral centroid, bandwidth, and pitch. Coefficient features such as mel-frequency cepstral coefficient and linear prediction coefficient are widely used. There are various classifiers that have been used for sound classification such as Gaussian maximum a posteriori (MAP), Gaussian mixture model, spatial partitioning scheme based on a k-d tree and a nearest neighbor classifier.

Liu, M. et al classify audio signals using content-based technology. Features were extract

The paper is organized as follows: section 2 presents the extraction of audio features using MFCC; section the concept of ANFIS; sections 3, 4, 5 are the simulation, results and conclusion, respectively.

II. Audio Feature Extraction

The first step in audio processing is feature extraction. The audio feature used is Mel Frequency Cepstral Coefficient (MFCC). This is the most popular and widely used feature extraction technique in speech processing particularly speech recognition. The technique is based on

the model of how sounds are generated in human. This technique was introduced by S.B. Davis, and P. Mermelstein in 1980 (Muda, L. et al., 2010).

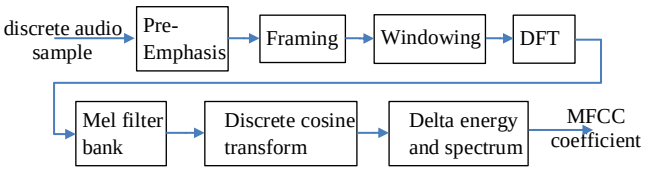

Figure 1. Steps in Computing MFCC

Pre-emphasis is the process of passing the signal to a filter in order to emphasize the energy of higher frequencies.

2. Framing - the signal is divided into frames with a duration of 20 msec. The voice signals is divided into frames of N samples with a separation of M samples. Usually M <N and the typical value of M is 100 and N=256.

3. The process of windowing is necessary to correct the problem called leakage in the power spectrum of a non-periodic signal. This is applied in order to determine the frequency content of the signal. In this step hamming window is used with the equation defined by

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N \qquad (1)$$

4. Fast Fourier Transform

The N sample signal is converted to frequency domain represented by *Y(w)*. *Y(w) = X(w)H(w) = x[n] * h[n]*, where *h[n]* and *H(w)* represent the transform pair of the vocal tract impulse response, *x[n]* and *X[w]* represent the transform pair of the sample signal, and the symbol * represents convolution operation.

5. The spectrum is fed into the array of mel filter banks. Each filter has a center frequency called mel frequency computed as,

$$mel\ frequency = 2595\log(1 + \frac{f}{700}) \qquad (2)$$

The output of mel filter bank is the sum of the spectral components. The aggregate of the output forms an array called mel spectrum.

6. Discrete cosine transform (DCT) is applied to mel spectrum. The output coefficients of DCT is called the Mel Frequency Cepstrum Coefficients (MFCC). There are twelve coefficients obtained in this process. This forms the feature vector of the audio signal.

III. Adaptive Neuro Fuzzy Inference System (ANFIS)

Fuzzy inference system (FIS) is the process of formulating the mapping from the given input and output using fuzzy logic. There are two known types of FIS: mamdani and sugeno. ANFIS is a fuzzy inference system that utilized neuro adaptive learning techniques in contructing fuzzy models. It is a sugeno-type fuzzy inference system which was first introduced by Jang in 1993. Sugeno-type fuzzy inference system is more compact and computationally efficient than mamdani-type (Mashrei, 2012). Its output is a function of weights computed as,

$$output = \frac{\sum_{i=1}^{n} w_i z_i}{\sum_{i=1}^{N} w_i} \quad . \qquad (3)$$

In a conventional fuzzy inference system (FIS) commonly referred to as mamdani-type FIS, the input is mapped to output using membership functions provided by the expert. This is a limitation of mamdani type, it requires an expert knowledge. ANFIS resolve this weak point. The network type structure of ANFIS is similar that of neural network. The neural network maps the input to the desired output by adjusting the weights of the network using a learning algorithm. It utilize neural network to adjust the parameters of the membership functions.

The output of ANFIS is illustrated in Figure 2 for a two-input sugeno-type fuzzy inference system. An equivalent figure is shown in Figure 3 for a two-input ANFIS based on neural network. ANFIS network structure similar to that of a neural network, which maps inputs through input membership functions and associated parameters, and then through output membership functions and associated parameters to outputs. The parameters of the membership function is adjusted through learning or training process. The learning process uses either back propagation or combination of least square estimation.

ANFIS structure based on neural network consists of five layers (Khoshsaadat, A., et al. , 2012). In the first layer, each node generates a membership grade of a linguistic label. In the second layer, each node has an output known as "firing strength" of each rule of fuzzy logic. In the third layer, the output of each node is the ratio of the strength of each rules to the sum of all the rule strengths. The output in this layer the normalized firing strength. Lastly, in the fifth layer, the output is the sum of the overall incoming signals.

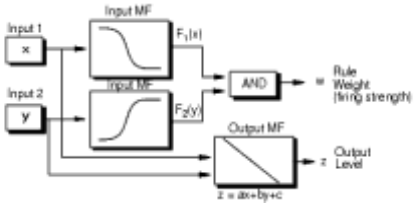

Figure 2. Two input sugeno-type fuzzy inference system (Matlab)

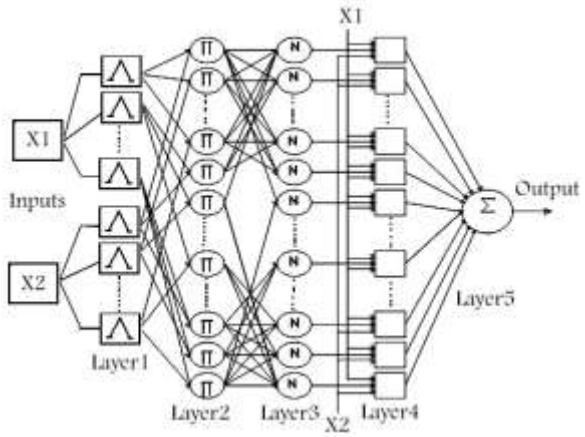

Figure 3 Structure of 2-input ANFIS (Khoshdaat, et al., 2014)

IV. SIMULATION

To simulate the normal and crisis condition in a public transport bus, 8 audio files were downloaded from the internet. The description of the files are shown in Table 1. The audio signals were edited to remove idle positions or silent part in the files. The first four files represent the crisis condition of the public transport bus.

Table 1. Audio files used in the simulations

Audio File	Audio signal	Frames	Training Data	Checking Data
1	Garan gun	378	100	100
2	Machine gun	628	100	100
3	Crowd panic	5227	100	100
4	Gunshots and crowd panic	1991	100	100
5	Bus inside noises 1	12290	100	100
6	Bus passed by	4443	100	100
7	Bus load passenger and go	8309	100	100
8	Bus inside noises 2	19439	100	100
Total			800	800

The simulations and processing were done using Matlab™ software. The audio signals were read and the steps in Figure 1 were performed. The MFCC parameters used in obtaining MFCCs is shown in Table 2. The number of frames obtained in each file is indicated in Table 1 under column "frame". Twelve MFCCs were obtained in each frame representing the feature vector. One hundred sets of MFCCs were used for training the ANFIS and another 100 for validation. They were selected randomly. The first four audio files represent the signals in crisis condition, and the output in this case is set to 1. The other four audio files represent the signals in normal condition and its output is set to -1.

Subtractive clustering was used to determine the cluster centers and the number of clusters for the generation of FIS. The important parameters of subtractive clustering are the range of influence represented by the radius of the circle *r*, squash factor, acceptance ratio and rejection ratio. Initially, these were set to the default value 0.5, 1.25, 0.5 and 0.15, respectively. Also simulated are the different values of range of influence represented by *r*: 0.49, 0.50, 0.51, 0.52, and 0.53. The range of influence is varied near the default value of 0.5 in order to observe its effect on the accuracy of designed NFIS with other parameters held constant.

The maximum required error in training was set to 0.005 and the maximum epoch was set to 3000 that means the trainings stops when one of these setting is reached.

Table 2 .MFCC parameters

Analysis Frame Duration (ms)	Tw	30
Analysis Frame Shift (ms)	Ts	5
Pre-emphasis Coefficient	alpha	0.97
Frequency Range to consider	R	[100 5000]
Number of Filterbank Channels	M	20
Number of Cepstral Coefficients	C	13
Cepstral Sine Lifter Parameter	L	22

V. RESULTS

The designed ANFIS for 12 inputs has a structure shown in Figure 4. It has 12 inputs and 1 output. Each input corresponds to each of the 12 MFCCs.

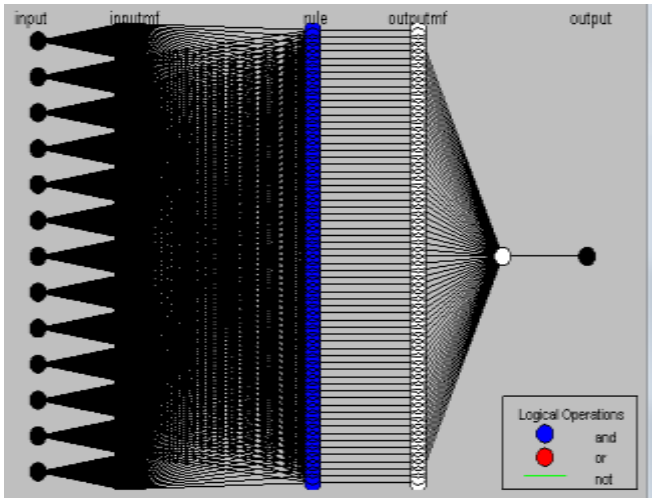

Figure 4. ANFIS structure for 12 inputs

The number of rules generated is dependent on the range of influence set in generating the FIS using subtractive clustering. The number of rules generated using a given range of influence is shown in Table 3. As the range of influence increases, the number of rules decreases.

Table 3 Number of rules given the radius of influence set in subtractive clustering.

Radius of influence	Number of rules
0.49	38
0.50	36
0.51	31
0.52	28
0.53	27

The result of the validation for each audio file is shown in Table 4. The output of ANFIS is crisp, a real number. In the tabulation of results, when the output is greater than 1, it is classified as crisis condition otherwise, it is classified as normal condition. The same result is plotted in Figure 5.

Table 4. Accuracy of validation

audio file	Range of influence (r)				
	0.49 (%)	0.5 (%)	0.51 (%)	0.52 (%)	0.53 (%)
1	74	82	89	89	88
2	96	92	97	100	87
3	97	98	98	98	95
4	96	100	98	100	100
5	100	100	100	100	100
6	6	9	16	16	13
7	99	95	97	96	97
8	100	100	100	100	100
Average	83.5	84.5	86.875	87.375	85.5

It can be observed in Figure 5 that most of the validation result has an accuracy greater than 82 %. There is one audio files whose validation accuracy is very low, that is audio file 6. This audio file is a sound of a very high speed bus that passed by. This is supposed to be classified as normal condition but the system can identified it as crisis condition. This implies that the feature of this sound is closely similar to gunshots or gun sounds and panic crowd. In the actual or real time implementation of this system, this error can be corrected by including a timer in the system. If the system detects a crisis condition it may send a status that a crisis condition occurs only after a few seconds to avoid false detection. The audio files 5 and 8 containing inside bus noises were perfectly validated. This implies that inside bus noises containing passenger talking and some occasional coughs of passengers are very different from crisis conditions containing gunshots and panic crowd.

Moreover, the results shows that the validation accuracy varies with range of influence, the parameter of subtractive clustering. The best value of the range of influence with other parameters held constant is 0.52. This entails that in the actual implementation of the system, the range of influence can be tuned in order to get the optimum validation accuracy.

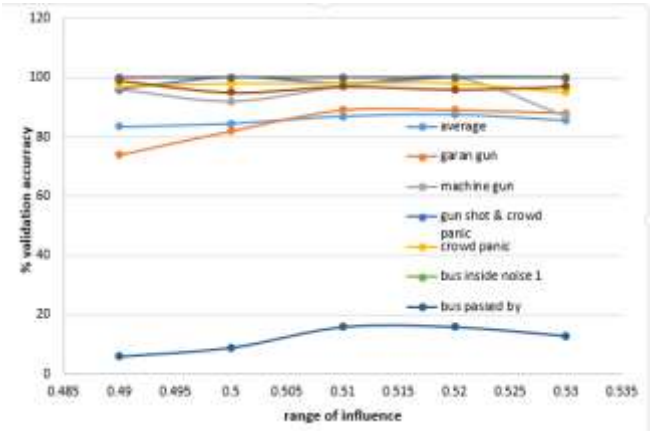

Figure 5. Validation accuracy for different values of range of influence

VI. CONCLUSION AND RECOMMENDATION

The ANFIS was able to detect the crisis and normal condition simulations. The result of the validation accuracy of ANFIS is greater than 82% except for the audio file that contains the sound of a very high speed bus that pass by. This sound is identified by the system as in crisis condition where in fact the data is classified as normal condition. To improve the accuracy in the actual implementation of ANFIS for audio classification it is important to gather as many data as possible for training so that it represents the data that ANFIS was trained for.

ACKNOWLEDGMENT

The authors would like to acknowledge the financial support extended by the Commission on Higher Education (CHED), De La Salle University-Manila, and Mindanao State University-General Santos City.

REFERENCES

[1] Meena, K. , et al. "Gender Classification in Speech Recognition Using Fuzzy Logic and Neural Network". The International Arab Journal of Information Technology, Vol. 10 , No. 5, September 2013.

[2] Silva, W. , Serra, G. "A Novel Intelligent System for Speech Recognition ". International Joint Conference on Neural Networks. Pp 3599-3604, July 2014.

[3] Liu M. , et. al(2002). Content-based Audio Classification and Retrieval Using a Fuzzy Logic System: Towards Multimedia Search Engines. Soft Computing 6, pp 357

[4] Yan, H. et al. "Adaptive neuro fuzzy inference system for classification of water quality status". Journal on Environmental Science. 22(12) 1891–1896 , 2010
Han Yan, Zhihong Zou∗, Huiwen Wang

[5] Muda, L., et al. "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, Vol 2, No.3, pp.138-143. March 2010

[6] Mashrei, M. "Neural Network and Adapative Neuro-Fuzzy Inference System Applied to Civil Engineering Problems". Fuzzy Inference System - Theory and Applications Edited by Dr. Mohammad Fazle Azeem pp. 471-504. Intech: May 2012

[7] Nazmy, T.M., El-Messiry, H., Al-Bokhity."Adaptive Neuro-Fuzzy Inference Inference System for Classification of ECG Signals". Journal of Theoretical and Applied Information Technology, Vol. 12, No. 2, Feb 2010.

[8]Thipparat,T. "Application of Adaptive Neuro Fuzzy Inference System in Supply Chain Management Evaluation". Fuzzy Logic-Algorithms Techniques and Implementations, chap 6. pp 115-126. Edited by Dadios, E. , Intech March 2012

[9] Roy, S. "Design of Adaptive Neuro-Fuzzy Inference System for predicting surface Roughness in Turning Operations". Journal of Scientific and Industrial Research. Vol. 64 pp. 653-659. September 2005

[10] Khoshsaadat, A. ,et al. "Design of a Controller with ANFIS Architecture Attendant Learning Ability for SSSC-Based Damping Controller Applied in Single Machine Infinite Bus System".Iranian Journal of Electrical & Electronic Engineering, Vol. 10, No. 3, Sep. 2014

[11] Buket D. Barkana et al., "Normal and Abnormal Non-Speech Audio Event Detection Using MFCC and PR-Based Features", Advanced Materials Research, Vol. 601, pp 200-208 December 2012