A Robust Environmental Sound Recognition System using BPNN and RBFNN

T.Sivaprakasam #1, P.Dhanalakshmi *2

# Assistant Professor, Depart. of Computer science and Engg., Annamalai University, India.
1 tsivaprakasam@gmail.com

* Associate Professor, Depart. of Computer science and Engg., Annamalai University, India.
2 abidhana01@gmail.com

***Abstract*— In a reverberant environment, the performance of acoustic event recognition system can be bolstered by choosing appropriate feature descriptors and classifier techniques. Neural networks are by far providing stunning classification results when compared to other classifiers. This paper analyses two different neural networks and their precision when they both stumble upon same targets in similar environment. The analysis is done on back propagation neural network (BPNN) and radial basis function neural network (RBFNN) with same dataset and then a conclusion is formed on the basis of their performance and efficiency. The experiments on various categories illustrate that the results of recognition for BPNN are significant and effective.**

Keyword- Feature Extraction, Pattern Classification.

(Spectral crest, Spectral decrease, Spectral slope, Spectral skewness, Spectral flatness, Back propagation neural network (BPNN), Radial basis function neural network (RBFNN)).

I INTRODUCTION

Audio events classification/detection is receiving a growing interest by the scientific community. It is especially the case in the context of audio retrieval and indexing applications but also in the context of multimedia event detection applications where audio can be used as a complementary source of information. In surveillance or homeland security (security of public places such as bank, subway, airport...) most of the systems are only based on visual signals to detect anomalous situations. In some of these situations audio conveys more significant information than video.

Environmental sounds provide many contextual cues that enable us to recognize important aspects of our surroundings. Our goal is then to use these acoustic cues as complementary information to automatically detect and analyze real-world situations from predefined classes of sounds in that environment.

II RELATED WORK

Environmental sound recognition concerns with the identification of sounds that do not originate from speech or music. The range of environmental sounds is extremely wide. Hence, most investigations concentrate on restricted domains. A popular research field is audio recognition in broadcasted video. In [4], the authors recognize the scene content of TV programs (e.g. weather reports, advertisements, basket ball and football games) by analyzing the audio track of the video. They extract pitch, volume distribution, frequency centroid and bandwidth to characterize TV programs. Classification is performed by a separate neural network for each class. A well investigated problem is to highlight detection in sports videos. The authors of [5] retrieve crucial scenes in soccer games by analyzing play-breaks. Whistles, that often refer to play breaks in sports, are detected using Spectral Energy within an appropriate frequency band. Another indicator for highlights is the audience. Excitement is quantified by Loudness, Silence and Pitch. Another area of interest is surveillance and intruder detection. The authors of [6] detect intruders in a room by monitoring variations in a room-specific transfer function.

III OUTLINE OF THE WORK

A. Audio event analysis

A general consideration says that similar event types produce similar sound. The techniques for feature extraction and classification that are reviewed in this paper are not centralized about the classification and identification of similar events that are occurring in diverse time and in various distances from the recording device, but instead this the goal of these techniques is to detect and classify different events that belong to a particular class. An acoustic event detection system consist three main components: pre-processing, Feature extraction and Classification.

B. *Acoustic features*

Feature representation of audio signal can also be considered as dimensionality reduction technique. Since environmental sounds may differ significantly from speech, we additionally consider features that address the possibility of high non-stationarily of sounds. Acoustic signals have very few representations and we must find the most important coefficients or characteristic, which contain information that will be used to discriminate among input classes [7].

Similar acoustic events occurring in comparable conditions would cause a similar acoustic signature that could be used for recognition. Characteristic patterns may be extracted from the Fourier description of the signature and used for recognition.

Frequency-Domain Features

Spectral characteristic of signatures vary significantly among target classes. In the spectral domain, acoustic signal waveforms generated by events appears as narrowband harmonic components. The information provided by these components, is used to construct the characteristic features for a particular event. Five spectral domain feature generation methods are utilized in this work.

a) Spectral Skewness

Spectral skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable that in this context is the spectrum of the signal. For a sample of N values forming a frame, the skewness is:

$$Skewness = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{N}\cdot\sum_{n=0}^{N-1}(x(n)-\overline{x})^3}{\left(\frac{1}{N}\cdot\sum_{n=0}^{N-1}(x(n)-\overline{x})^2\right)^{3/2}} \qquad (1)$$

In this equation (1) where $\overline{x}$ represents the mean of the magnitudes, m_3 is the sample third central moment, and m_2 is the sample variance.

b) Spectral Decrease

Spectral decrease represents the amount of decreasing of the spectral amplitude. This formulation comes from perceptual studies and it is supposed to be more correlated to human perception. The formula is:

$$Decrease = \frac{1}{\sum_{n=1}^{N-1} x(n)} \cdot \sum_{n=1}^{N-1} \frac{x(n)-x(0)}{N-1} \qquad (2)$$

In this equation (2) where x(n) represents the weighted frequency value or magnitude of bin number n.

c) Spectral Slope

Spectral slope also represents the amount of spectral energy decrease but as a function of frequency. It assumes that the amplitude spectrum follows a linear model:

$$A(k) = mk + b \qquad (3)$$

The slope *m* is computed by linear regression.

$$m = \frac{\frac{K}{2}\sum_{k=0}^{\frac{K}{2}-1} kA(k) - \sum_{k=0}^{\frac{K}{2}-1} k \sum_{k=0}^{\frac{K}{2}-1} A(k)}{\frac{K}{2}\sum_{k=0}^{\frac{K}{2}-1} k^2 - \left(\sum_{k=0}^{\frac{K}{2}-1} k\right)} \qquad (4)$$

In this equation (3), (4) where K is the total number of frequency values, A(k) is the spectral amplitude with frequency index k.

d) Spectral Crest

Spectral crest factor indicate how flat or "peaky" the power spectral density is in a given sub band. The spectral crest Factor or peak-to-average ratio is a measurement of a waveform, calculated from the peak amplitude of the waveform divided by the mean value of the waveform.

$$Crest\ Factor = \frac{|x_n(i)|_{peak}}{\sqrt{\frac{\sum_{i=1}^{N} x n^2(i)}{N}}} \qquad (5)$$

In this equation (5) where N is the frame length, and $x_n[i]$ represents spectral amplitude of the *i* th sample in the nth frame.

e) Spectral Flatness

Spectral flatness is a measure of distribution of spectral power in an audio spectrum. The spectral flatness is calculated by dividing the geometric mean of the power spectrum by the arithmetic mean of the power spectrum. The spectral flatness used is measured across the whole band. The formula is:

$$Flatness = \frac{\sqrt[N]{\prod_{n=0}^{N-1} x(n)}}{\frac{\sum_{n=0}^{N-1} x(n)}{N}} \qquad (6)$$

In this equation (6) where $x(n)$ represents the magnitude of bin number n of the power spectrum.

C) Neural network

Neural networks have emerged as an important tool for classification. The recent vast research activities in neural classification have established that neural networks are a promising alternative to various conventional classification methods.

Neural network is a bio-inspired network made from neurons and can solve the problems that are hard to be modeled analytically. The model is rough, as the human brain has very parallel computational device, achieving great power due to connectivity of large number of simple neurons. Neural networks have boundless applications in fields like signal processing, image and video processing, weather forecasting, stock market predictions, genetics, bioinformatics, power systems, defense systems, etc [8].

D) Modeling Techniques for audio classification

a) Back propagation neural network (BPNN)

Back propagation Neural Network is a network that based on Back propagation learning technique and that works on the principle of supervised learning [8]. In general it is called the Feed Forward Back propagation neural network. With regard to architecture it is basically Multi-layer Perception [8]. The Back propagation neural network is the gemstone that enchanted and mesmerized researchers and showed the true power of neural networks. It opened research doors with endless opportunities in various fields of engineering, sciences and statistics while being computationally economical. But on the darker side the BPNN has also been called the ‘black box’ as it has a fixed algorithmic operation and that’s about it, there is no fixed topology (number of nodes and neurons used) for it and exhibits different results with different data subsets of the same dataset that is it is very hard to sway it to global minima. Regardless of all these factors, overall the BPNN is quite accurate and easy to manipulate with respect to other neural networks [9].Figure 1 shows a basic BPNN comprising of an input, hidden and output layer.

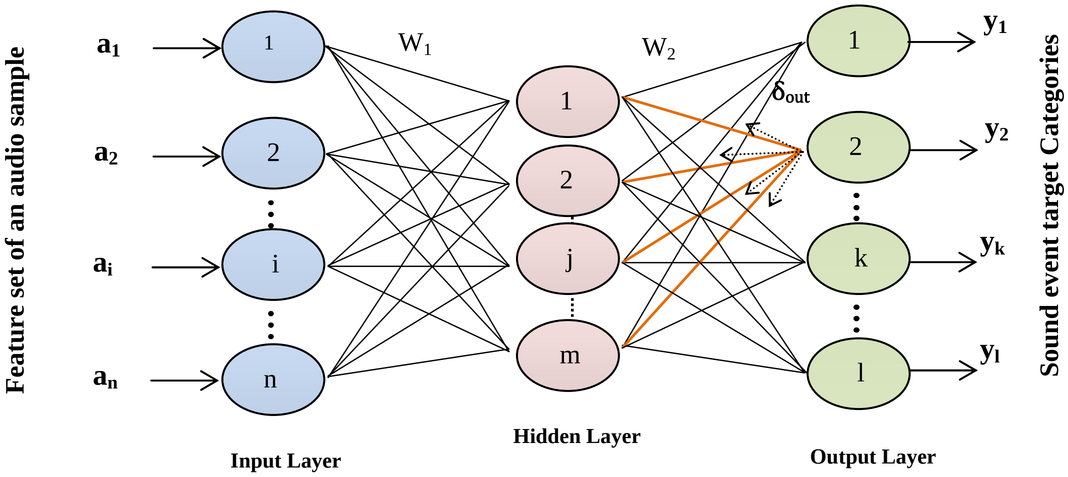

Fig. 1: BPNN Architecture

Where a_1, a_2 and a_n are the inputs applied on the input layer, middle one is the hidden layer and latter is the output layer.

Consider the output layer neuron2 where ‘δout’ is the error signal that is generated when the output ‘y2’ is compared to the target output of the training dataset comprising of the ideal classification result [8].

The error signal moves from the output layer to the hidden layer changing the weights W_2 to adjust to the correct result once this error is minimized close to zero the weights are fixed meaning the network is trained and can be tested. Similarly every weight in the network is updated.

b) Radial basis function neural network (RBFNN)

Among these artificial neural networks, the RBF network forms a special architecture with several distinctive features. A typical RBF neural network classifier has three layers, namely input, hidden, and output layer. The input layer of the network is made of source nodes that connect the coordinates of the input vector to the nodes in the second layer. The second layer, the only hidden layer in the network, includes processing units called the hidden basis function units which are located on the centers of well chosen clusters. Each hidden layer node adopts a radial activated function, and output nodes implement a weighted sum of hidden unit outputs [3].

The output layer is linear, and it produces the predicted class labels based on there sponse of the hidden units. The structure of multi-input and multi-output (MIMO) RBF neural network is represented by Figure 2.The parameters of an RBF type neural network consist of the centers μ_m and the spread $\boldsymbol{\sigma}_m$ of the basis functions at the hidden layer nodes and the synaptic weights W_{ij} of the output layer nodes. The RBF centers are also points in the input space. It would be ideal to have them at each distinct point on the input space, but for any realistic problem, only a few input points from all available points are selected using clustering.

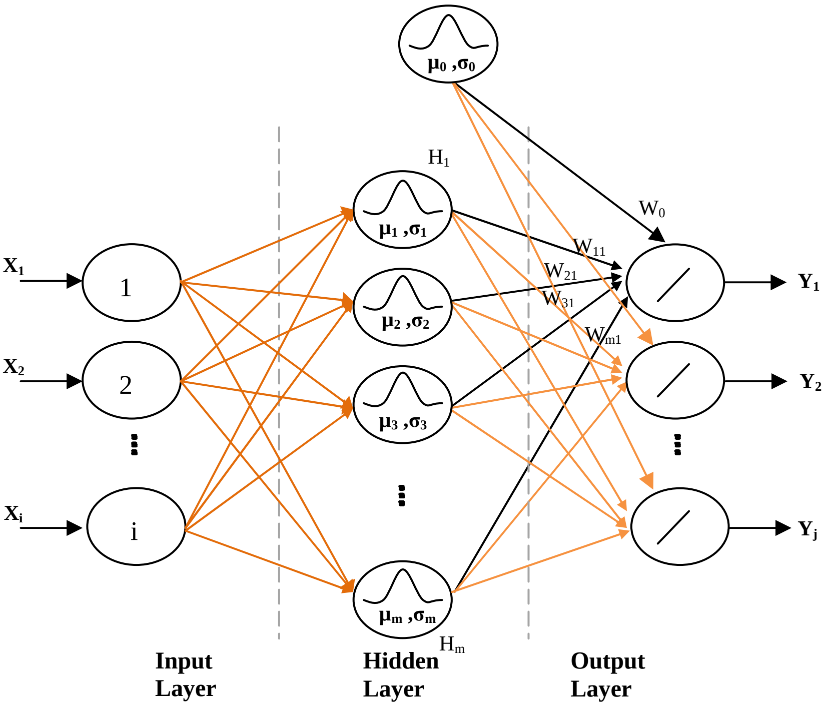

Fig. 2: RBFNN Architecture

For an input vector X_i, the j th hidden node produces a activation function h_j is given by

$$h_{j=}\exp\left\{\frac{-\left\|X_i-\mu_j\right\|^2}{2\sigma_j^2}\right\} \qquad (7)$$

In this equation (7) where $-\left\|X_i-\mu_j\right\|^2$ is the distance between the point representing the input Xi and the center of the *j* th hidden node as measured by some norm. In this study, the Euclidean norm is used. The output of the network at the *k*th output node is given by

$$y_{ik} = \sum_{j=1}^{L} h_j w_{kj} \qquad (8)$$

In this equation (8) where h_j is Gaussian function and w_{kj} is the weight between the hidden and output layer.

The performance of the RBF network depends highly on the number and initial locations of the hidden units. Generally, the positions of the hidden units are initialized using unsupervised clustering algorithms such as k-means or Expectation Maximization or supervised clustering algorithms such as the ones introduced in [10,11]. In this study, we initialized the hidden unit centers using the k-means clustering. The μ_i and $\boldsymbol{\sigma}_i$ are calculated by using suitable clustering algorithm. Here the k-means clustering algorithm is employed to determine the centers. The algorithm is composed of the following steps:

1. Randomly initialize the samples to k means (clusters) $\mu_1 \ldots \mu_k$.
2. Classify n samples according to nearest μ_k.
3. Re-compute μ_k.
4. Repeat steps 2 and 3 until no change in μ_k.

The number of activation functions in the network and their spread influence the smoothness of the mapping.

E) Experimental Analysis and Results

In this section we discuss the steps taken to conduct the experiments. Experimental setups and steps will be present in this section.

a) Database

The database for the experiments contains 1000 samples which are taken from AURORA database. The recordings are categorized into general classes according to common characteristics of the scenes (220 kitchen noises, 180 living room noises, 210 laundry sounds, 230 meeting sounds, 160 office sounds) and events (Pan boiling, steel plate ,music player, paper scrap ,washing machine, flush,overlapped speech, footsteps, typewriter, dust bin, etc.). The categorization of the scenes was somewhat ambiguous; some of the recordings are associated with more than one higher-level class. The recordings are manually labelled and are separated into 2-second, 3-second and 5-second fragments. Every sound signal was stored with some properties that are also the initial conditions and criteria for the well-functioning of the algorithm. The sample database is split into training sets and test sets. We randomly select 80% sounds of each class for the training set. The remaining 20% sounds form the test set. Thus we have taken different proportion of samples based on class dependency in each category as shown in table 1.

Table 1:
ACOUSTIC DATABASE DESCRIPTOR

Context	Total amount of database
Kitchen	22%
Living Room	18%
Laundry	21%
Meeting	23%
Office	16%

b) Experimental Evaluation

This subsection performs an analysis on the performance of the proposed method. Figure 3, shows the block diagram of this sound recognition system. The Environmental scene classification system comprises of three main modules: 1) Pre-processing, 2) Feature extraction and 3) Final classifier.

1) Pre-processing

To extract the features from the acoustic signal, the signal must be pre-processed and divided into successive windows or analysis frames. Throughout this work, a sampling rate of 16 kHz, 16 bit monophonic, pulse code modulation (PCM) format in wave audio is adopted. Environmental audio signal which is recorded using multi-microphone setting is pre-processed before extracting features. This involves normalization of audio waveform, pre-emphasis and windowing of the frame. The process of pre-emphasis provides high frequency emphasis and windowing reduces the effect of discontinuity at the ends of each frame of audio. The training data is segmented into fixed-length and overlapping frames (in our experiments we used 20 ms frames with 10 ms overlapping). When neighboring frames are overlapped, the spectral characteristics of audio content can be taken into consideration in the training process.

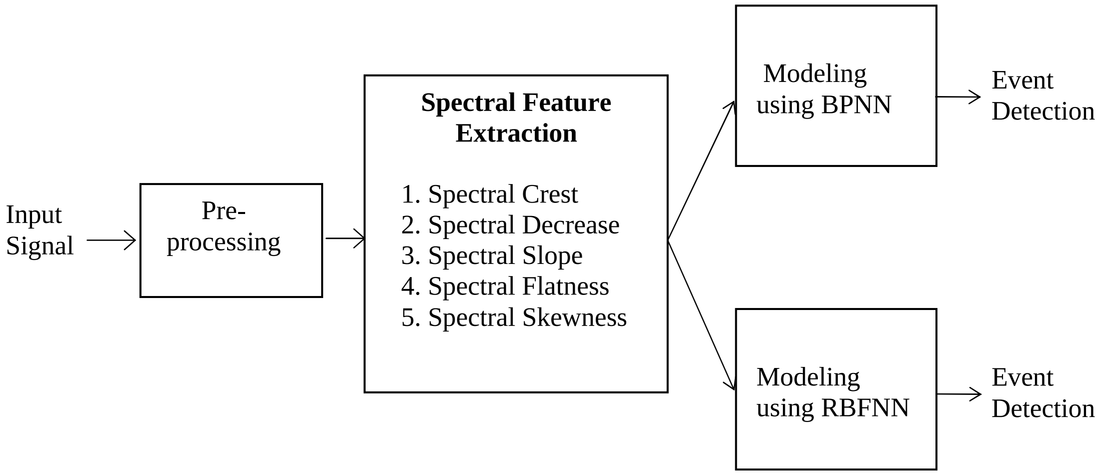

Fig.3: Block diagram for sound recognition system

2) *Feature extraction*

In this subsection we evaluated the quality of a large number of features from various fields of audio retrieval. The goal has been the identification of an optimal feature set for the retrieval of environmental sounds. For this purpose, we have performed a quantitative data analysis in order to identify independent features. Data analysis reveals redundancies and dependencies between features. Information obtained by data analysis supports the selection of feature combinations. Statistical data analysis of the optimal feature set reveals that the spectral Descriptor is independent from the other features and highly beneficial for retrieval.

3) *Modeling using BPNN*

We use a back propagation neural network classifier to discriminate various events. Classification parameters are calculated using BPNN learning. The training process analyzes audio training data to find an optimal way to classify audio frames into their respective classes. The training data should be sufficient to be statistically significant. The BPNN learning algorithm is applied to produce the classification parameters according to calculated features. The derived classification parameters are used to classify the context of the audio data. The classification results for the proposed feature set are shown in Figure 4 for various sample durations. From the results, we observe that the overall classification accuracy is high for 3-second samples when compared to other duration.

Table 2:
Recognition matrix for 15 hidden neurons

	2 second	**3 second**	**5 second**
Recognition Accuracy for BPNN	79%	91.7%	86%

To determine the performance of BPNN, we examine the results by varying the number of hidden layers. Using the same settings as the rest of the experiments, we examined hidden neurons of 5, 10, 15 and 20 and used the same number of neurons for each environment type. The overall recognition rates are plotted in figure 5.We see that the classification performance peaks around fifteen and the performance slowly degrades as the number of neurons varies. The highest recognition rate for each class across the number of mixtures was obtained with 15 neurons.

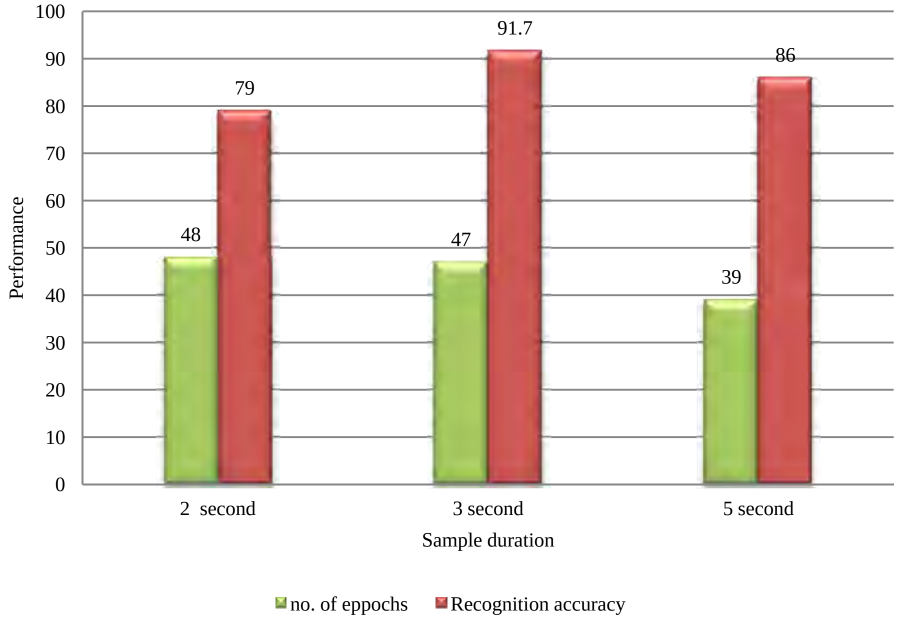

Fig.4: Recognition chart for different sound samples for 15 hidden neurons

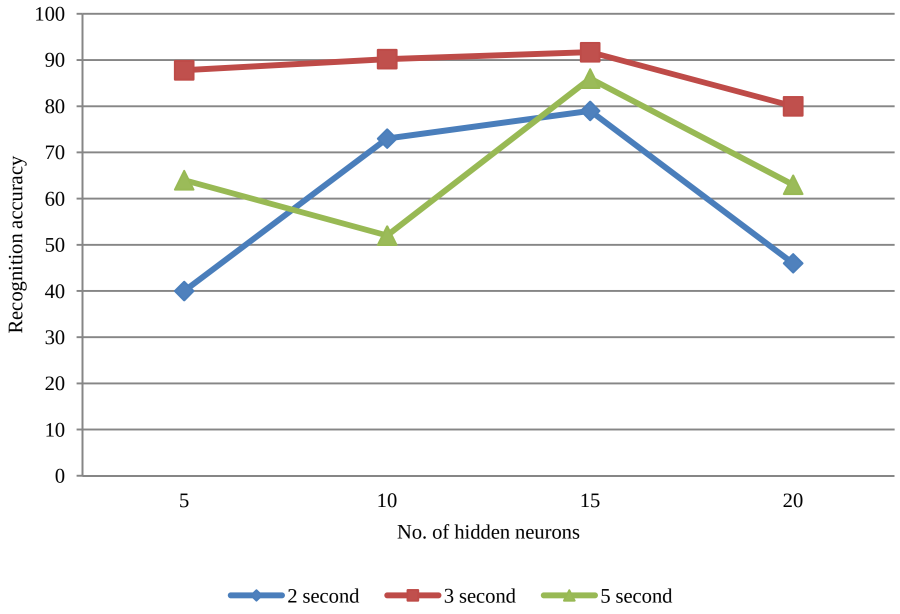

Fig.5: Recognition chart for different hidden neurons

4) Modeling using RBFNN

The RBFN is trained by adaptively updating the free parameters, i.e. center and width of the basis function, and the weight between the hidden and output neurons of the network. To select an optimal RBFN model, the number of neurons in the hidden layer was varied from 2 to 30, and the learning rate was varied between 0.05 and 0.5. The initial basis function centers were chosen randomly from the input space, and the initial weight values were chosen randomly between ±0.9. Normalized datasets were used for the training, testing, and validation of the RBFN model. The best network was found to be one having 26 basis functions with a learning rate of 0.9 and 0.05 for center and weight respectively. The prediction errors of the validation patterns are larger because these patterns are outside the training space.

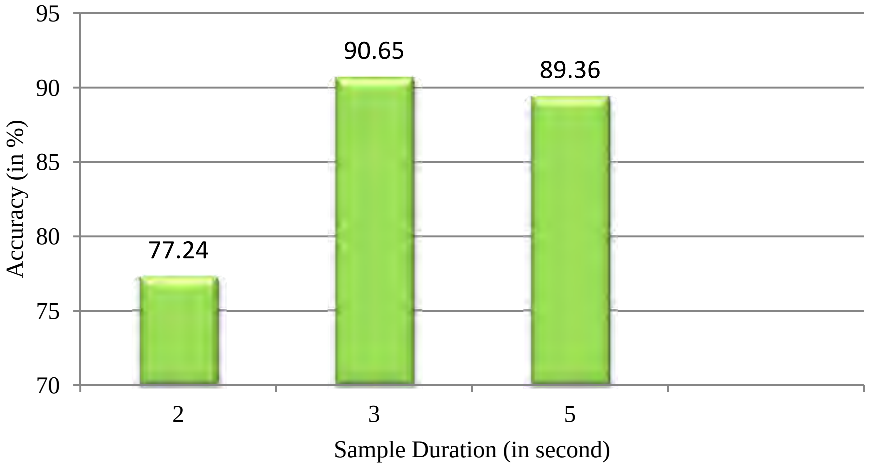

Fig:6 Recognition Accuracy for different sound sample durations

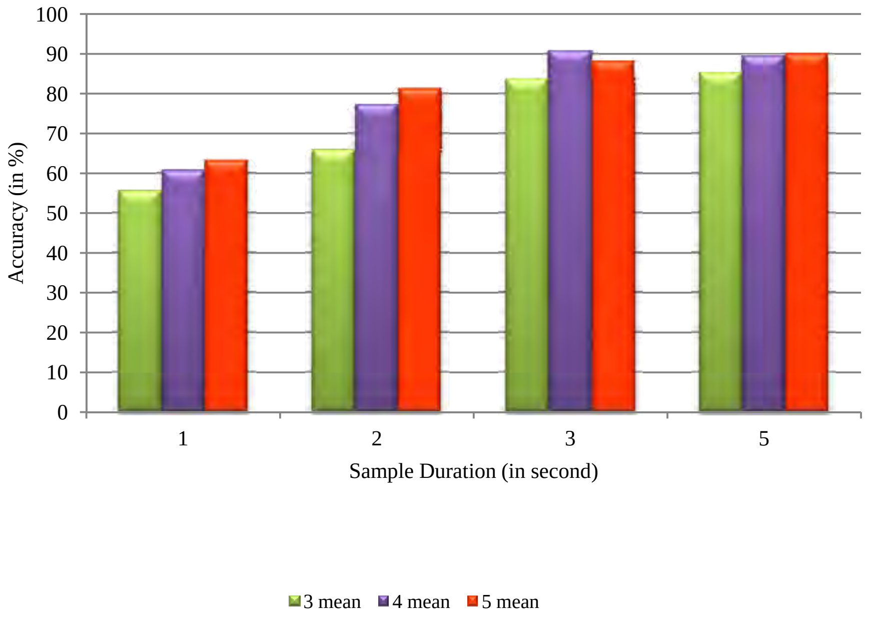

Fig:7 Comparison graph for various means

There performance are measured for various sample durations for RBFNN network structure with mean=4.

Table 3:
Recognition matrix for means k=4.

	2 second	**3 second**	**5 second**
Recognition Accuracy for RBFNN	77.24%	90.65%	89.36%

IV CONCLUSION AND FUTURE WORK

The BPNN showed very promising results. Still there are some issues that must be pointed out such as the BPNN exhibits slow learning for big datasets, like this may be due to the iterative learning used by BPNN. The choice of topology that is the number of neurons used in various layers of BPNN drastically affects the prediction and classification capability of the NN. If too many neurons are used for BPNN training the NN can attain good learning performance but that does not ensure accurate and good generalization. Still it is very effective and unlike most of the neural networks which produce good results for some datasets and bad for others, BPNN performs fairly well on most of them.

RBFNN on the other hand is also accurate as it is non-iterative, highly parallel and takes less time during the training and testing phase.

Future work will be dedicated to the extension of the current system to different types of acoustic events that occur in abnormal situations such as shouts, cries or manifestation of fear.

REFERENCES

[1] Selina Chu, Shrikanth Narayanan and C.C. Jay kuo, "Environmental Sound Recognition With Time-Frequency Audio Features", IEEE Transactions on Audio, Speech and Language Processing, Vol. 17, No. 17, No. 6, August 2009.

[2] P.Dhanalakshmi, S.Palanivel and V.Ramalingam, " Classification of audio signal using SVM and RBFNN", International Journal of Expert Systems with Application (Elsevier), United Kingdom, vol.36, issue 3, part 2, pp. 6069-6075, April 2009.

[3] Burak Uzkent, Buket D.Barkana, and Hakan Cevikalp, "Non-Speech Environmental Sound Classification Using SVMs with a New set of Features", International Journal Of Innovative Computing, Information and Control, ISSN 1349-4198, Volume 8, Number 5(B), pp. 3511-3524, May 2012.

[4] Z.Liu,J. Huang,Y. Wang, and T.Chnan, "Audio feature extraction and analysis for scene classification", In IEEE Workshop on Multimedia Signal processing, vol. 20, pp. 343-348, 1997.

[5] D.Tjondronegoro, Y.Chen, and B.Pham, "The power of play break for automatic detection and browsing of self consumable sport video highlights", In Proceedings of the ACM Workshop on Multimedia Information Retrieval, pp. 267-274, 2004.

[6] Y.Choi, K.Kim, J.Jung, S.Chun, and K.Park, "Acoustic intruder detection system for home security", In IEEE Transactions on Consumer Electronics, vol. 51, pp. 130-138, 2005.

[7] Varun Kumar Kakar, Manisha Kandpal, "Techniques of Acoustic Feature Extraction for Detection and Classification of Ground Vehicles", International Journal of Emerging Technology and Advance Engineering, ISSN 2250-2459, Volume 3, Issue 2, February 2013.

[8] Ali Raza, M.Umair Khalid, Alizaidi, "Comparison between Back-propagation and General Regression Neural Networks for Underwater Mine detection", Proceeding of International Bhurban Conference on Applied science & Technology, Islamabad, Pakistan (pp. 180-184) January 10-13, 2011.

[9] Nathalie Japkowicz, Editor Douglas Fisher, "Supervised versus Unsupervised Binary-Learning by Feed-forward Neural Networks", May 2000.

[10] H.Cevikalp, D.Larlus and F.Jurie, " A supervised clustering algorithm for the initialization of RBF neural network classifiers", Proc of the 15th IEEE signal processing and communications Application Conference, Eskisehir, Turkey, 2007.

[11] B.Scholkopf, K.K.Sung, C.J.C.Burges, F.Giriosi, P.Niyogi, T.Poggio and V.Bapnik, "Comparing Support vector machines with Gaussian kernals to radial basis function classifiers", IEEE Transactions on signal processing, vol. 45, no. 10, pp.2758-2765, 1997.