# PERFORMANCE OF DIFFERENT CLASSIFIERS IN SPEECH RECOGNITION

**Sonia Suuny[1], David Peter S[2], K. Poulose Jacob[3]**

*[1] Associate Professor, [2, 3]Professor,*
*[1]Dept. of Computer Science, Prajyoti Niketan College, Kerala, India,*
*[2, 3] Professor., Dept. of Computer Science, Cochin University of Science & Technology, Kerala, India,*
***sonia.deepak@yahoo.co.in, davidpeter@cusat.ac.in, kpj@cusat.ac.in***

## Abstract

*Speech is the most natural means of communication among human beings and speech processing and recognition are intensive areas of research for the last five decades. Since speech recognition is a pattern recognition problem, classification is an important part of any speech recognition system. In this work, a speech recognition system is developed for recognizing speaker independent spoken digits in Malayalam. Voice signals are sampled directly from the microphone. The proposed method is implemented for 1000 speakers uttering 10 digits each. Since the speech signals are affected by background noise, the signals are tuned by removing the noise from it using wavelet denoising method based on Soft Thresholding. Here, the features from the signals are extracted using Discrete Wavelet Transforms (DWT) because they are well suitable for processing non-stationary signals like speech. This is due to their multi-resolutional, multi-scale analysis characteristics. Speech recognition is a multiclass classification problem. So, the feature vector set obtained are classified using three classifiers namely, Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Naive Bayes classifiers which are capable of handling multiclasses. During classification stage, the input feature vector data is trained using information relating to known patterns and then they are tested using the test data set. The performances of all these classifiers are evaluated based on recognition accuracy. All the three methods produced good recognition accuracy. DWT and ANN produced a recognition accuracy of 89%, SVM and DWT combination produced an accuracy of 86.6% and Naive Bayes and DWT combination produced an accuracy of 83.5%. ANN is found to be better among the three methods.*

***Index Terms:*** *Speech Recognition, Soft Thresholding, Discrete Wavelet Transforms, Artificial Neural Networks, Support Vector Machines and Naive Bayes Classifier.*

----------***----------

## 1. INTRODUCTION

Automatic Speech recognition (ASR) is one of the intensive areas of research since it helps people to communicate in a more natural and effective way[1]. Speech recognition systems can be characterized by many parameters. The commonly used method to measure the performance of a speech recognition system is the recognition accuracy. Many parameters affect the accuracy of the speech recognition system. The accuracy and acceptance of speech recognition has improved a lot in the last few years. Though automatic speech recognition systems have improved a lot these years and are now extensively used, their accuracy continues to lag behind human performance, particularly in adverse conditions [2]. In spite of the advances in ASR, there is still a considerable gap between human and machine performance [3]. Speech is a multi-component signal with time varying frequency and amplitude. Due to this variability, transitions may occur at different times in different frequency bands.

Automatic recognition of spoken digits is one of the challenging tasks in the field of speech recognition [4]. A spoken digit recognition process is needed in many applications that need numbers as input such as automated banking system, airline reservations, voice dialing telephone, automatic data entry, command and control etc [5]. ASR is basically a pattern recognition problem which also involves a number of technologies and research areas like Signal Processing, Natural Language Processing, Statistics etc [6]. Speech signals are non stationary in nature. Speech recognition is a complex task due to the differences in gender, emotional state, accent, pronunciation, articulation, nasality, pitch, volume, and speed variability in people speak [7]. Presence of background noise and other types of disturbances also affect the performance of a speech recognition system. Speaker independence is difficult to achieve because these models recognize the speech patterns of a large group of people. The paper is organized as follows. Section 2 gives a brief description of the problem definition. The methodology used in the design is explained in section 3. The creation of the digits database is illustrated in section 4. Section 5 describes the method used for preprocessing. Section 6 elaborates the feature extraction technique used in this work followed by the classification techniques in section 7. Section 8 presents a

detailed analysis of the experiments done and the results obtained. Conclusion is given in the last section.

## 2. PROBLEM DEFINITION

The main objective of a speech recognition system is to recognize speech with maximum accuracy. Here, digits in Malayalam, one of the four major Dravidian languages of southern India and the official language of the people of Kerala is chosen for recognition. Research in Malayalam is in its infancy stage and only very few works have been reported in Malayalam. The performance of the overall speech recognition system depends on the techniques used for pre-processing, feature extraction technique selected and the classifiers used. In this work, wavelet denoising method based on soft thresholding is used for pre-processing the captured signals [8]. Wavelet based feature extraction method namely DWT is selected for feature extraction due to the good time and frequency resolution properties of the wavelets. For classification, three classifiers are selected namely ANN, SVM and Naive Bayes classifier. The performance of these classification methods in classifying the digits is evaluated along with DWT.

## 3. ARCHITECTURE OF THE SYSTEM

In this work, the speech recognition system is divided into 4 modules. First the spoken digits database is created since there is no built-in standard database available in Malayalam. The speech samples in the database are often corrupted by additive noises like background noise. So these speech signals are pre-processed using wavelet denoising techniques to suppress the noise in it. For processing speech, the speech signal has to be represented in the form of parameters. So, after denoising the signals, they are presented to feature extraction stage where the features are extracted using a wavelet based feature extraction technique called Discrete Wavelet Transforms. The extracted features are then classified into appropriate classes during classification phase. Here, three classifiers namely ANN, SVM and Naive Bayes classifier are used and the performance of these techniques are evaluated based on recognition accuracy. The four stages of this work are explained below.

## 4. CREATION OF THE SPEECH DATABASE

In Malayalam, since there is no standard database available, a spoken digits database is created using 200 speakers of age between 6 and 70 uttering 10 Malayalam digits. We have used 80 male speakers, 80 female speakers and 40 children for creating the database. Male and female speech differ in pitch, frequency, phonetics and many other factors due to the difference in physiological as well as psychological factors. The samples stored in the database are recorded by using a high quality studio-recording microphone at a sampling rate of 8 KHz (4 KHz band limited). Recognition has been made on the ten Malayalam digits from 0 to 9 under the same configuration. Our database consists of a total of 2000 utterances of the digits. The spoken digits are preprocessed, numbered and stored in the appropriate classes in the database. The number digit, spoken digits in Malayalam, their International Phonetic Alphabet (IPA) format and English translation are shown in Table 1.

**Table -1:** Malayalam digits database

| Number digit | Digits in Malayalam | IPA format | English Translation |
|---|---|---|---|
| 0 | പൂജ്യം | /puː/d͡ʒya/ /ṁ/ | Zero |
| 1 | ഒന്ന് | /o/ /nn/ | One |
| 2 | രണ്ട് | /ɾa/ /ɳʈ/ | Two |
| 3 | മൂന്ന് | /muː/ /nn/ | Three |
| 4 | നാല് | /n̪aː/ /l/ | Four |
| 5 | അഞ്ച് | /a/ /ñc/ | Five |
| 6 | ആറ് | /aː/ /r/ | Six |
| 7 | ഏഴ് | /eː/ /ɻ/ | Seven |
| 8 | എട്ട് | /e/ /ʈʈ/ | Eight |
| 9 | ഒൻപത് | /o//na//pa/ /t̪/ | Nine |

## 5. PRE-PROCESSIG SIGNALS

Noises from background cause degradation in the speech signals. So, these signals are tuned so that the noise present in it is removed before extracting the features. There are a number of techniques available for speech enhancement. Since we are using wavelets for feature extraction, wavelet denoising algorithms are used for reducing the noise in the signal. The two popular thresholding functions used in wavelet denoising method are the hard and the soft thresholding functions [9]. In both the methods, a threshold value is selected. In hard thresholding, if the absolute value of an element is lower than the threshold, then these values are set to zero. Soft thresholding is an extension of hard thresholding. Here, the elements whose absolute values are lower than the threshold are first set to zero and then the nonzero elements are shrinked towards 0. Hard and soft thresholding can be expressed as

$$X_{Hard} = \begin{cases} X & if \quad |X| > \tau \\ 0 & if \quad |X| \le \tau \end{cases} \quad (1)$$

$$X_{Soft} = \begin{cases} sign(X)\ (|X| - |\tau|) & if \quad |X| > \tau \\ 0 & if \quad |X| \le \tau \end{cases} \quad (2)$$

Where X represents the wavelet coefficients and ι is the threshold value. In this work, soft thresholding technique is used. There are different standard threshold values. In this work, we have used the universal threshold derived by Donoho and Johnstone [10] for the white Gaussian noise under a mean square error criterion which is defined as

$$\iota = \sigma\sqrt{2\log(N)} \quad (3)$$

Where σ is the standard deviation and N is the length of the signal. Standard deviation σ can be calculated as σ = MAD/0.6745, where MAD is the median of the absolute value of the wavelet coefficients. The outline of the algorithm used for denoising mainly consists of 3 steps.

- Apply wavelet transform to the noisy signal to produce the noisy wavelet coefficients up to 8 levels.
- Detail wavelet coefficients are then shrinked using soft thresholding technique by selecting an appropriate threshold limit.
- The inverse discrete wavelet transform of the threshold wavelet coefficients is computed which produces the denoised signal.

## 6. FEATURE EXTRACTION

Feature Extraction is a major part of the speech recognition system since it plays an important role to separate one speech from other and this has been an important area of research for many years. Selection of the feature extraction technique plays an important role in the recognition accuracy, which is the main criterion for a good speech recognition system. Here, DWT is used for extracting features and a brief description of this is given below.

### 6.1 Discrete Wavelet Transforms

DWT is a relatively recent and computationally efficient technique for extracting information from non-stationary signals like audio. The main advantage of the wavelet transforms is that it has a varying window size, being broad at low frequencies and narrow at high frequencies, thus leading to an optimal time–frequency resolution in all frequency ranges [11]. DWT uses digital filtering techniques to obtain a time-scale representation of the signals. DWT is defined by

$$W(j,K) = \sum_j \sum_k X(k) 2^{-j/2} \psi(2^{-j} n - k) \quad (4)$$

Where Ψ (t) is the basic analyzing function called the mother wavelet. In DWT, the original signal passes through a low-pass filter and a high-pass filter and emerges as two signals, called approximation coefficients and detail coefficients [12]. In speech signals, low frequency components h[n] are of greater importance than high frequency components g[n] as the low frequency components characterize a signal more than its high frequency components [13]. The successive high pass and low pass filtering of the signal is given by

$$Y_{low}[k] = \sum_n x[n] h[2k - n] \quad (5)$$

$$Y_{high}[k] = \sum_n x[n] g[2k - n] \quad (6)$$

Where Yhigh (detail coefficients) and Ylow (approximation coefficients) are the outputs of the high pass and low pass filters obtained by sub sampling by 2. The filtering process is continued until the desired level is reached according to Mallat algorithm [14]. The discrete wavelet decomposition tree is shown in figure 1.

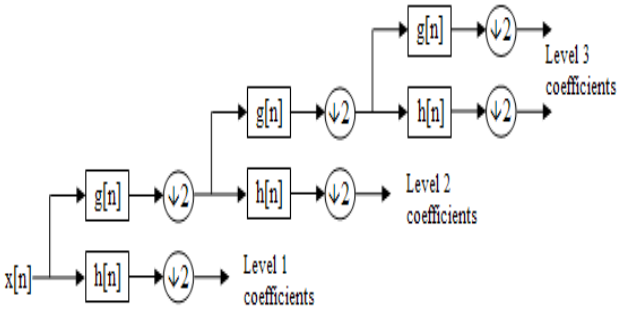


**Fig -1:** DWT Decomposition over 3 levels

## 7. CLASSIFICATION

Classification is another important part of a speech recognition system since the patterns are classified into different classes during this stage. Pattern recognition is becoming increasingly important in the age of automation and information handling and retrieval. During classification stage, decisions are made based on the similarity measures from training patterns using information relating to known patterns. Then they are tested using the unknown patterns. Since there are a number of different classes in a speech recognition problem, a multiclass classification technique is needed.

In almost all classification methods, the data is separated into training and test sets. Each instance in the training set contains a target value which represents the corresponding class and a

set of attributes. The test data do not contain a target value. The objective of the classifier is to produce a model from the training data which predicts the target values of the test data. A brief description of the various classifiers used in this work like ANN, SVM and Naive Bayes classifier are given below.

### 7.1 Artificial Neural Networks

Nowadays, ANNs are utilized in many applications due to their parallel distributed processing, distributed memories, error stability, and pattern learning and distinguishing ability. ANN is an information processing paradigm consisting of a number of simple processing units or nodes called neurons. Each neuron accepts a weighted set of inputs and produces an output [15]. Algorithms based on ANN are well suitable for addressing speech recognition tasks. Inspired by the human brain, neural network models use a number of characteristics such as learning, generalization, adaptively, fault tolerance etc. [16].

In this work, we are using the architecture of the MLP network, which consists of an input layer, one or more hidden layers, and an output layer. The algorithm used is the back propagation training algorithm. In this type of network, the input is presented to the network and moves through the weights and nonlinear activation functions towards the output layer, and the error is corrected in a backward direction using the well-known error back propagation correction algorithm. After extensive training, the network eventually establishes the input-output relationships through the adjusted weights on the network. After training the network, it is tested with the dataset used for testing.

### 7.2 Support Vector Machines

SVM is a very useful technique used for classification. It is a classifier which performs classification methods by constructing hyper planes in a multidimensional space that separates different class labels based on statistical learning theory [17][18]. Though SVM is inherently a binary nonlinear classifier, we can extend it to multiclass classification since ASR is a multiclass problem. There are two major strategies for multiclass classification namely One-against-All [17] and One-against-One or pair wise classification [19]. The conventional way is to decompose the M-class problem into a series of two-class problems and construct several binary classifiers. In this work, we have used One-against-One method in which there is one binary SVM for each pair of classes to separate members of one class from members of the other. This method allows us to train all the system, with a maximum number of different samples for each class, with a limited computer memory [20].

### 7.3 Naive Bayes Classifier

Since speech recognition is a multiclass classification problem and Naive Bayes classifiers can handle multiclass classification problems, it is also used here for classifying the digits. Naive Bayes classifier is based on the Bayesian theory which is a simple and effective probability classification method. This is a supervised classification technique. For each class value it estimates that a given instance belongs to that class [16]. The feature items in one class are assumed to be independent of other attribute values called class conditional independence [17]. Naive Bayes classifier needs only small amount of training set to estimate the parameters for classification. The classifier is stated as

$$P(A|B) = P(B|A) * P(A)/P(B) \quad (7)$$

Where P(A) is the prior probability of marginal probability of A, P(A|B) is the conditional probability of A, given B called the posterior probability, P(B|A) is the conditional probability of B given A and P(B) is the prior or marginal probability of B which acts as a normalizing constant. The probability value of the winning class dominates over that of the others [18].

## 8. EXPERIMENTS

Using DWT, the signals are decomposed into approximation and detail coefficients. The low frequency components are decomposed upto 10th level. The number of features obtained for a digit at 10th level is 16. The decomposition of digits zero to nine using DWT at 10th level is shown below.

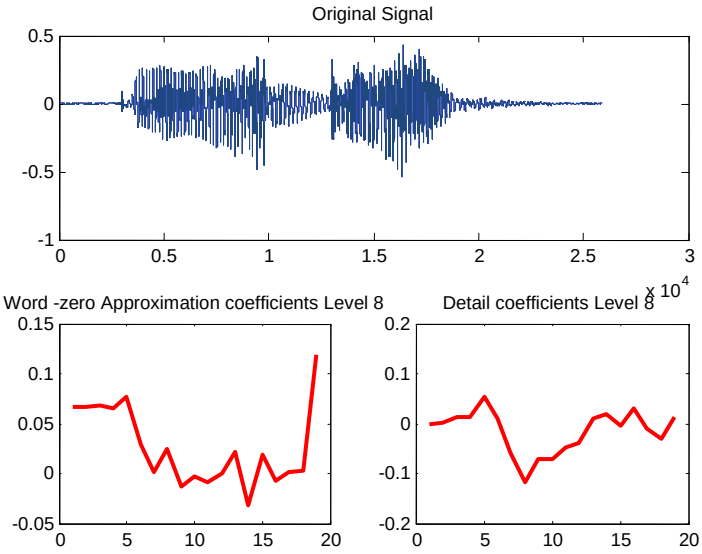


**Fig -2:** Decomposition of digit zero at 10th level

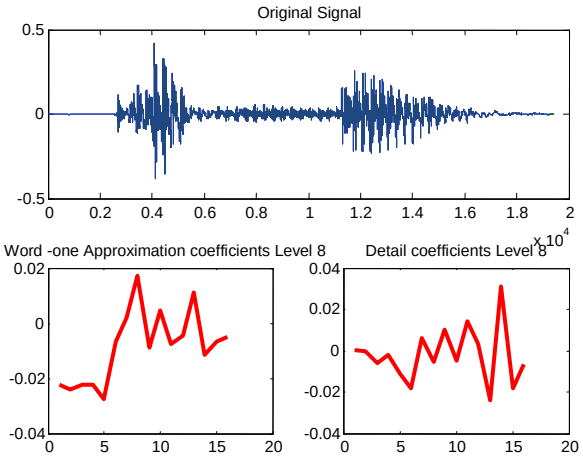


**Fig- 3:** Decomposition of digit one at 10th level

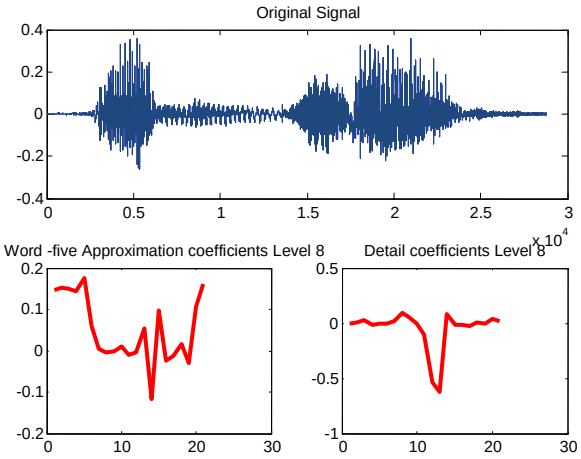


Fig- 6: Decomposition of digit five at 10th level

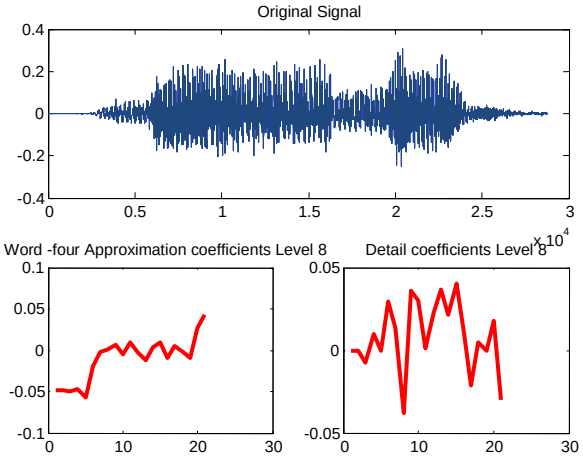


**Fig- 4:** Decomposition of digit two at 10th level

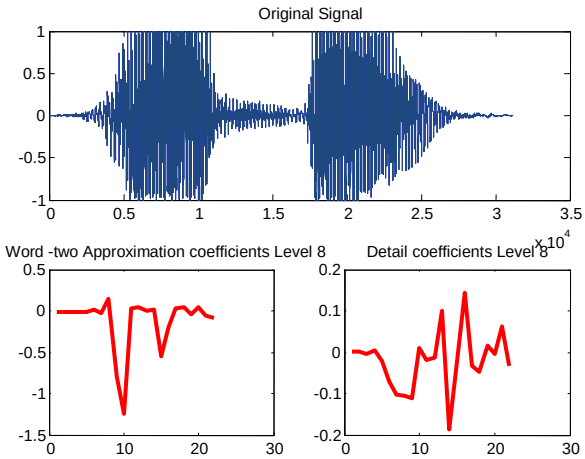


**Fig- 7:** Decomposition of digit four at 10th level

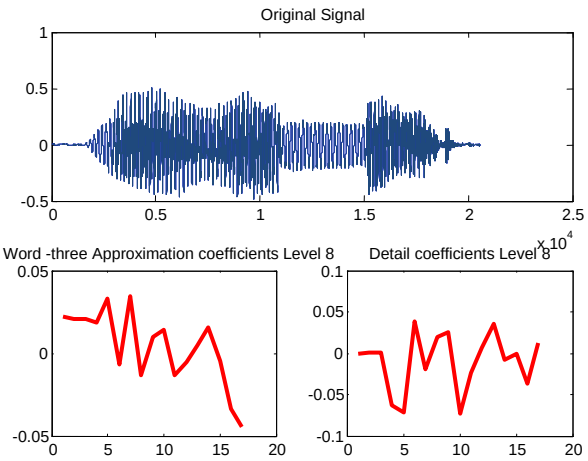


**Fig- 5:** Decomposition of digit three at 10th level

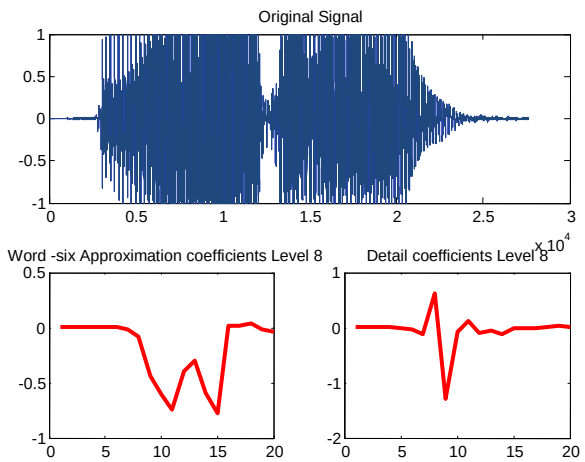


**Fig- 8:** Decomposition of digit six at 10th level

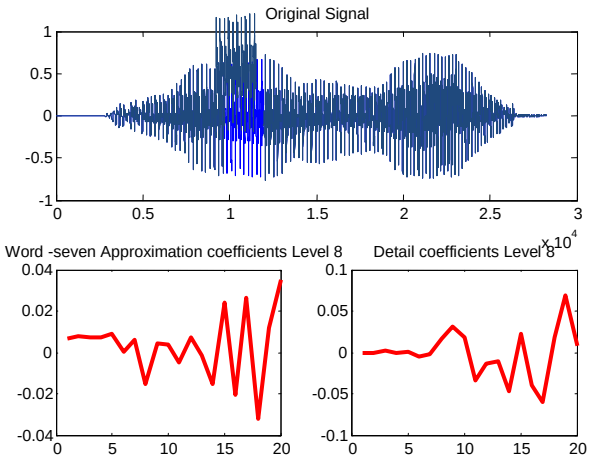


**Fig- 9:** Decomposition of digit seven at 10th level

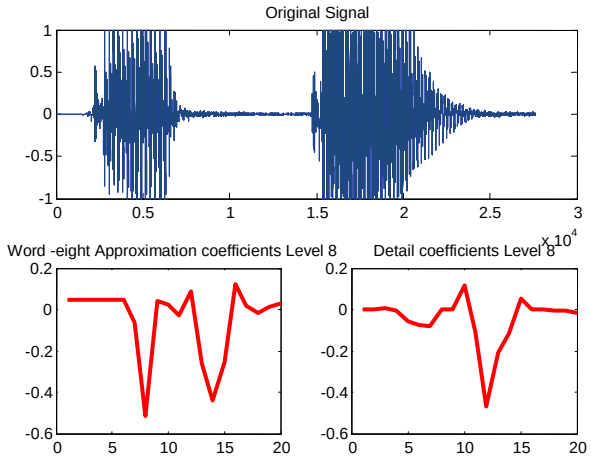


**Fig- 10:** Decomposition of digit eight at 10th level

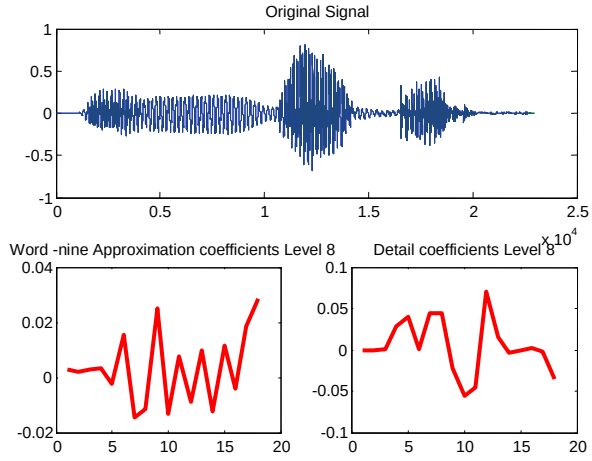


**Fig- 11:** Decomposition of digit nine at 10th level

For classification using ANN, SVM and Naive Bayes classifier, the feature set data is divided into training, validation and test set. 70% data is used for training, 15% for validation and 15% for testing. That is, 7000 data is used for training, 1500 data for validation and 1500 for testing.

### 8.1 Performance Evaluation of Classifiers

For classification using ANN, one input layer, one hidden layer and one output layer is used. Back propagation algorithm is used for error correction.

For classification using SVM, since one-against-one SVM classification is used, we construct a binary SVM for each pair of classes. The number of binary SVMs built is defined as C* (C-1)/2, where C is the number of classes. So a total of 10*(10-1)/2 = 45 binary SVMs are built. The table given below shows the comparison of various classifiers.

**Table -2:** Performance Evaluation of Classifiers

| Classifier | Recognition Accuracy % |
|---|---|
| ANN | 89 |
| SVM | 86.6 |
| Naive Bayes | 83.5 |

From the results it is clear that all these classifiers are good in recognizing speech. But better results are obtained using ANN.

## CONCLUSIONS

In this work, a speech recognition system is designed for speaker independent spoken digits in Malayalam. The speech samples in the database are applied to wavelet denoising using soft thresholding for removing noise. DWT is used for feature extraction. Three classifiers namely ANN, SVM and Naïve Bayes classifier are used for classifying the digits into proper classes. A comparative study of these three classifiers is performed. The performance of all these techniques are tested and evaluated. All techniques are found to be efficient in recognizing speech. The accuracy rate obtained by using DWT and ANN combination is 89% which is found to be better than that of the other two methods. The experiment results show that this hybrid architecture using discrete wavelet transforms and neural networks could effectively extract the features from the speech signal and classify it for efficient recognition of digits. For future work, other feature extraction techniques like Linear Predictive Coding (LPC), Mel Frequency Cepstral Coefficients (MFCC), Wavelet packet Decomposition (WPD) can also be used and classified using these classifiers.

## ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to all who have contributed throughout the course of this work.

## REFERENCES:

[1]. L. Rabiner, B. H. Juang, "Fundamentals of Speech Recognition", Prentice-Hall, Englewood Cliffs, NJ, 1993.
[2]. Kenneth Thomas Schutte, "Parts-based Models and Local Features for Automatic Speech recognition", PhD Thesis in Electrical Engineering and Computer Science, 2009.
[3]. Scharenborg, "Reaching Over the Gap: A Review of Efforts to Link Human and Automatic Speech Recognition Research", Speech Communication, 49:336–347, 2007.
[4]. Y. Ajami Alotaibi, Investigating Spoken Arabic Digits in Speech Recognition Setting, Information Sciences, Vol 173, pp. 115-139, 2005.
[5]. C. Kurian, K. Balakrishnan, Speech Recognition of Malayalam Numbers, World Congress on Nature and Biologically Inspired Computing, pp. 1475-1479, 2009.
[6]. B. Gold, N. Morgan, Speech and Audio Signal Processing, New York, John Wiley and Sons, 2002.
[7].recognition.http://www.learnartificialneuralnetworks.com/ speechrecognition.html
[8]. Daniel M. Rasetshwane, J. Robert Boston, Ching-Chung Li, Identification of Speech Transients Using Variable Frame Rate Analysis and Wavelet Packets, Proc.of the 28th IEEE EMBS Annual International Conference, USA, 2006.
[9]. Yasser Ghanbari, Mohammad Reza Karami, A new Approach for Speech Enhancement based on the Adaptive Thresholding of the Wavelet Packets, Speech Communication, Vol. 48 (8), pp. 927–940, 2006.
[10]. D.L. Donoho., De-noising by Soft Thresholding, IEEE transactions on Information Theory, vol. 41, No. 3, pp. 613-627, 1995.
[11]. Elif Derya Ubeyil.,Combined Neural Network model Employing Wavelet Coefficients for ECG Signals Classification, Digital Signal Processing, Vol 19, pp. 297-308, 2009.
[12]. S. Chan Woo, C.Peng Lin, R. Osman., Development of a Speaker RecognitionSystem using Wavelets and Artificial Neural Networks, Proc. of Int. Symposium on Intelligent Multimedia, Video and Speech processing, pp. 413-416, 2001.
[13]. S. Kadambe, P. Srinivasan., Application of Adaptive Wavelets for Speech, Optical Engineering , Vol 33(7), pp. 2204-2211, 1994.
[14]. S .G. Mallat., A Theory for Multiresolution Signal Decomposition: The Wavelet Representation, IEEE Transactions on Pattern Analysis And Machine Intelligence, Vol.11, 674-693, 1989.
[15]. Freeman J. A, Skapura D. M., 2006. Neural Networks Algorithm, Application and Programming Techniques, Pearson Education.
[16]. Economou K., Lymberopoulos D., 1999. A New Perspective in Learning Pattern Generation for Teaching Neural Networks, Volume 12, Issue 4-5, 767-775.
[17]. V.N. Vapnik., Statistical Learning Theory, J. Wiley, N.Y., 1998.
[18]. N. Cristianini, J. Shawe-Taylor., An introduction to Support Vector Machines, Cambridge University Press, Cambridge, U.K., 2000.
[19]. Ulrich H.-G. Kreßel., Pairwise Classification and Support Vector Machines, Advances inKernel Methods Support Vector Machine Learning, Cambridge, MA, MIT press, pp. 255-268, 1999.
[20]. C.W. Hsu, C.J. Lin, A Comparison of Methods for Multi-class Support Vector Machines. IEEE Transactions on Neural Networks, 13(2), pp. 415–425, 2002.
[21]. Li Dan, Liu Lihua, Zhang Zhaoxin, Research of Text Categorization on WEKA, proc. Of Third International Conference on Intelligent System Design and Engineering Applications, pp. 1129-1131, 2013
[22]. J. Han and M. Kamber, Data Mining Concepts and Techniques, Morgan Kauffman Publishers, 2000.
[23]. Laszlo Toth, Andras Kocsor, Janos Csirik, On Naive Bayes In Speech Recognition, Int. J. Appl. Math. Comput. Sci., Vol. 15, No. 2, pp. 287–294, 2005.

## BIOGRAPHIES:

Sonia Sunny is working as Associate professor in the department of Computer Science, Prajyoti Niketan College, Pudukad, Thrissur, Kerala State, India. Currently she is doing research in the area of Speech Recognition at Cochin University of Science and Technology, Cochin, Kerala State, India. In 1995, she received her M.Sc in Computer Science from Bharathidasan University, Thiruchirappalli and M.Phil in Computer Science from Bharathiar University, Coimbatore in 2008. Her research interest includes Speech Processing, Artificial Intelligence and Pattern Recognition.

David Peter S is presently working as professor in Computer Science at School of Engineering, Cochin University of Science and Technology, Cochin, Kerala, India. He did his Post-graduation in Computer Science at IIT Madras and PhD at Cochin University. He has presented research papers in several International Conferences and has published many articles in International Journals. His areas of interest includes Artificial Intelligence, Natural Language Processing, Information Systems, Engineering, etc.

K. Poulose Jacob, Professor of Computer Science at Cochin University of Science and Technology (CUSAT) since 1994, is currently Director of the School of Computer Science Studies. He holds additional charge as the honorary Director of CUSAT Planning & Development. He has presented research papers in several

International Conferences in Europe, USA, UK, Australia and other countries. He has delivered invited talks at several national and international events. Dr. Jacob is a Professional member of the ACM (Association for Computing Machinery) and a Life Member of the Computer Society of India. Till now twelve candidates have obtained PhD degrees in Computer Science & Engineering under his supervision. He has been PhD Theses examiner for several Universities. He has more than 75 research publications to his credit. His research interests are in Information Systems Engineering, Intelligent Architectures and Networks.