Development of a Speech Recognition System for Speaker Independent Isolated Malayalam Words

Sonia Sunny[1]
Department of Computer Science,
Cochin University of Science and Technology
Kochi-682022, India

David Peter S[2]
School of Engineering,
Cochin University of Science and Technology
Kochi-682022, India

K Poulose Jacob[3]
Department of Computer Science,
Cochin University of Science and Technology
Kochi-682022, India

Abstract— **In this paper, a speech recognition system is developed for recognizing speaker-independent, isolated words. Speech recognition is a fascinating application of Digital Signal Processing and is a pattern classification task wherein an input pattern is classified as a sequence of stored patterns that have previously been learned. Isolated words in Malayalam, which belong to one of the four Dravidian languages of Southern India, are used to create the database. Feature extraction in the time-frequency domain is performed using Wavelet Packet Decomposition (WPD). Artificial Neural Networks (ANNs) are used for training, testing and pattern recognition. Wavelets are very much suitable for processing non stationary signals like speech because of its multi-resolution characteristics and efficient time frequency localizations. Algorithms based on neural networks are well suitable for addressing speech recognition tasks. Recognition accuracy of 87.5% is obtained using this hybrid architecture of WPD and ANN.**

Keywords- Speech recognition; Feature extraction; Wavelet packet decomposition; Classification; Artificial neural networks.

I. INTRODUCTION

Speech recognition is still an intensive field of research in the area of digital signal processing due to its versatile applications. Despite of the advances made in this area, machines cannot match the performance of human beings in terms of accuracy and speed especially in the case of speaker independent speech recognition. Since speech is the primary means of communication between people, research in automatic speech recognition and speech synthesis by machine has attracted a great deal of attention over the past five decades [1].

Recent technological advances have made much progress in the recognition of complex speech patterns. But much more research and development is needed in this field. Speech recognition system mostly performs two fundamental operations: signal modeling and pattern matching [2]. During signal modeling, speech signal is converted into a set of parameters using an operation called feature extraction. Pattern matching is the task of finding parameter set from memory which closely matches the parameter set obtained from the input speech signal also known as classification. Among these stages, feature extraction is a key, because better feature is good for improving recognition rate. Recognition accuracy is an important measure for calculating the performance of a speech recognition system.

The paper is organized as follows. In section 2, the theory of feature extraction is reviewed followed by the concepts of wavelet packet decomposition employed during this stage. Section 3 describes the pattern classification stage using artificial neural networks. The isolated spoken words database is explained in section 4. Section 5 presents the detailed analysis of the experiments done and the results obtained. Conclusions are given in the last section.

II. FEATURE EXTRACTION TECHNIQUE

During feature extraction, the input signals are transformed into a set of features. The extracted features set will contain only the relevant information from the input data in order to perform the desired task of classification. The technique selected for feature extraction has great importance since good features increase the speech recognition rate. Different types of feature extraction methods are available and many researches have been conducted in the field of speech recognition. Most of the speech-based studies are based on Fourier Transforms (FTs), Short Time Fourier Transforms (STFTs), Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding (LPCs), and prosodic parameters. Literature on various studies reveals that in case of the above said parameters, the feature vector dimensions and computational complexity are higher to a greater extent. Many of the above mentioned techniques consider a signal to be stationary. When compared to Fourier spectrum, the wavelet coefficients can capture more time and frequency localized information from the speech signal [3]. Wavelets can decompose complex information such as speech, music, images etc. into elementary forms at different positions and scales.

A. The Wavelet Packet Decomposition

Wavelets are a relatively recent and computationally efficient technique used for extracting information about non-stationary signals like audio and WPD uses finite durative wavelets instead of periodical sinusoidal waves. WPD allows time domain information to be incorporated with frequency domain information using multiple window durations. Long windows are used when high resolution information is needed and short windows for extracting low resolution information. It allows simultaneous use of long-time interval for low-frequency information and short-time interval for high-frequency information [4]. When compared to wavelet transform, it is a more detailed method since the signal is passed through more filter [5].

In WPD, the original signal passes through two complementary filters, namely low-pass and high-pass filters, and emerges as two signals called approximation coefficients and detail coefficients [6]. Wavelet packet decomposition is based on wavelet transform and decomposes a signal with the same widths in all frequency bands [7]. In the next level, both the low frequency sub-bands and high frequency sub-bands are decomposed into lower and higher frequency parts. The decomposition procedure is repeated until the desired level of decomposition is reached. Wavelet packet decomposition can provide a multi-level time-frequency decomposition of signals and more precise information of the signal. By using wavelets, the size of the feature vector is less compared to other methods and so the computational complexity can also be successfully reduced. It also provides good time and frequency localizations. The WPD decomposition tree is shown in figure 1.

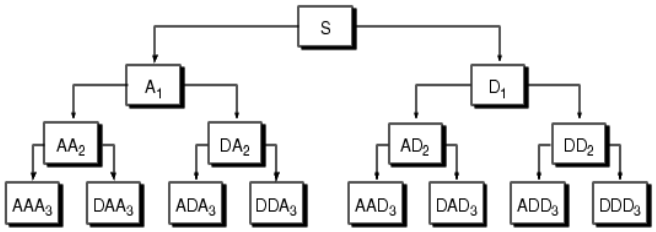


Figure 1. WPD decomposition Tree

III. PATTERN CLASSIFICATION

Classification is one of the most commonly used decision making task among human beings. Speech recognition is basically a pattern recognition problem. During classification stage, the input data is trained using information relating to known patterns and then they are tested using the test data set. Pattern recognition is becoming increasingly important in the age of automation and information handling and retrieval. Since neural networks are good at pattern recognition, many early researchers applied neural networks for speech pattern recognition. In this study also, we have used neural networks as the classifier. Neural networks can perform pattern recognition; handle incomplete data and variability well. The increasing acceptability of neural network models to solve pattern recognition problems has been mainly due to its low dependence on domain-specific knowledge relative to model-based and rule-based approaches and due to the availability of efficient learning algorithms for users to implement [8].

A. Artificial Neural Networks

ANNs have been applied to an increasing number of real-world problems of considerable complexity. An Artificial Neural Network is an information processing paradigm consisting of a number of simple processing units or nodes called neurons that is inspired by the way biological nervous systems, such as the brain, process

information. It can store experimental knowledge and make it available for use. Each neuron accepts a weighted set of inputs and produces an output [9]. Neural Networks have become a very important method for pattern recognition because of their ability to deal with uncertain, fuzzy, or insufficient data. Algorithms based on Neural Networks are well suitable for addressing speech recognition tasks. Inspired by the human brain, neural network models attempt to use some organizational principles such as learning, generalization, adaptivity, fault tolerance etc. [10].

In this work, we use architecture of the Multi Layer Perceptron (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. In most networks, the principle of learning a network is based on minimizing the gradient of error [11,12]. After extensive training, the network will eventually establish the input-output relationships through the adjusted weights on the network. After training the network, it is tested with the dataset used for testing. The structure of an MLP network is given in figure 2.

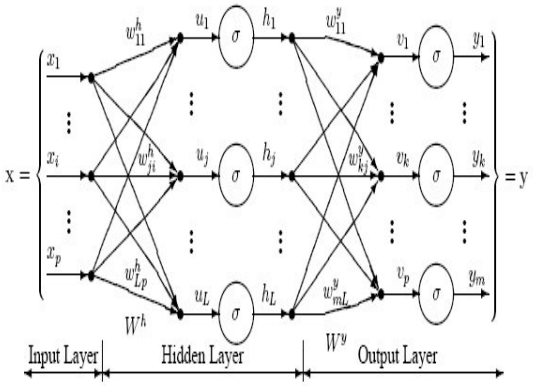


Figure 2. Structure of an MLP network

IV. DATABASE USED

Twenty commonly used isolated words from Malayalam language is chosen to create the database. Fifty speakers are selected to record the words. Each speaker utters 20 words. We have used twenty male speakers and thirty female speakers for creating the database. The speech samples are taken from speakers of age between 20 and 30. Thus the database consists of a total of 1000 utterances of the spoken words. 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). The same configuration and conditions are utilized for the recognition of these 20 isolated spoken words. The spoken words are preprocessed, numbered and stored in the appropriate classes in the database. The spoken words, words in English and their International Phonetic Alphabet (IPA) format are shown in Table 1.

TABLE 1. ISOLATED WORDS STORED IN THE DATABASE AND THEIR IPA FORMAT

Words in Malayalam	Words in English	IPA format
കേരളം	Keralam	/kēra!am/
വിദ്യ	Vidya	/vidjə/
പൂവ്	Poovu	/pu:və/
താമര	Thamara	/θa:mʌrə/
പാവ	Paava	/pa:və/

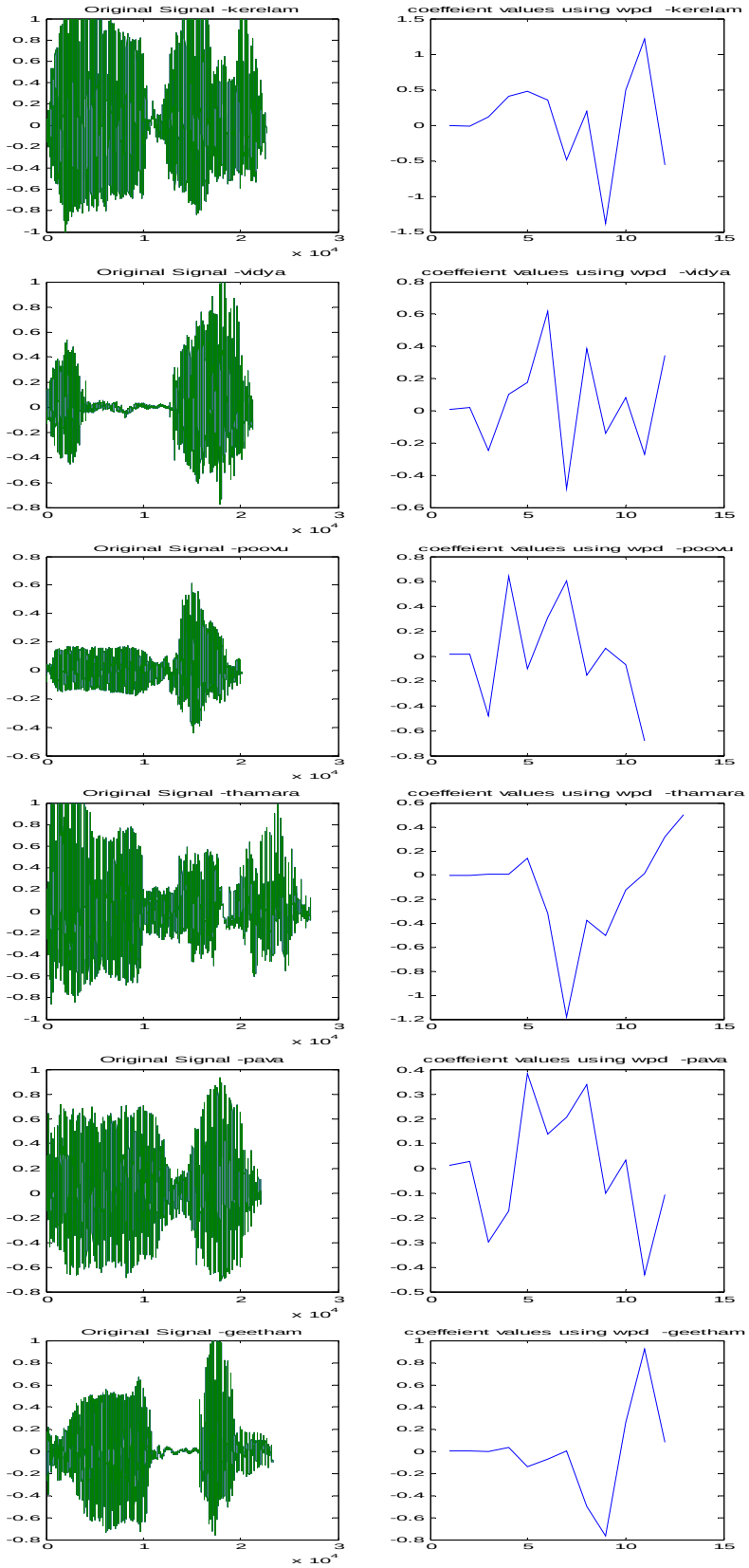
ഗീതം	**Geetham**	**/gi:θΛm/**
പത്രം	**Pathram**	**/pΛθrəm/**
ദയ	**Daya**	**/ðΛjə/**
ചിന്ത	**Chintha**	**/tʃinθΛ/**
കടൽ	**Kadal**	**/kΛdΛl/**
ഓണം	**Onam**	**/əunΛm/**
ചിരി	**Chiri**	**/tʃiri/**
വീട്	**Veedu**	**/vi:də/**
കുട്ടി	**Kutti**	**/kuṭi/**
മരം	**Maram**	**/mΛrəm/**
മയിൽ	**Mayil**	**/mΛjil/**
ലോകം	**Lokam**	**/ləukΛm/**
മൗനം	**Mounam**	**/maunəm/**
വെള്ളം	**Vellam**	**/ve!!Λm/**
അമ്മ	**Amma**	**/ΛmmΛ/**

V. EXPERIMENTAL RESULTS

In the first part of the experiment, feature vectors are extracted from the speech signals using wavelet packet decomposition. Since there are a number of wavelet families available, selection of the suitable wavelet and the number of decomposition levels play an important role in obtaining good recognition accuracy in speech recognition using the WPD. Among the various wavelet bases, the most popular wavelets that represent foundations of Digital Signal Processing called the Daubechies wavelets are used here because of its orthogonality property and efficient filter implementation [13]. Recent research has pointed out that the Daubechies order-4 (DB4) wavelet is an appropriate basis for speech recognition [14]. Their frequency responses have maximum flatness at frequencies 0 and π and so here also db4 type of mother wavelet is used for feature extraction purpose. The speech samples in the database are successively decomposed into approximation and detailed coefficients. The feature vector coefficients obtained using WPD at 12^{th} level is taken for classification. The number of feature vectors obtained is 11.

During the second phase, the feature vectors obtained from wavelet packet decomposition are given as the input to the artificial neural network classifier. Here we have divided the database into three. Out of the 1000 speech samples, 700 samples are used for training, 150 samples for validation and 150 samples for testing. MLP architecture is used for the classification scenario. Using this network, the classifier could successfully recognize the spoken words. The results obtained clearly shows the efficiency of Neural Networks in classifying the extracted coefficients.

Results obtained using WPD and ANN is given below. The original signal and the coefficient values after decomposition of 12 isolated words are shown in figure 3.


Original Signal -kerelam
coeffeient values using wpd -kerelam
Original Signal -vidya
coeffeient values using wpd -vidya
Original Signal -poovu
coeffeient values using wpd -poovu
Original Signal -thamara
coeffeient values using wpd -thamara
Original Signal -pava
coeffeient values using wpd -pava
Original Signal -geetham
coeffeient values using wpd -geetham

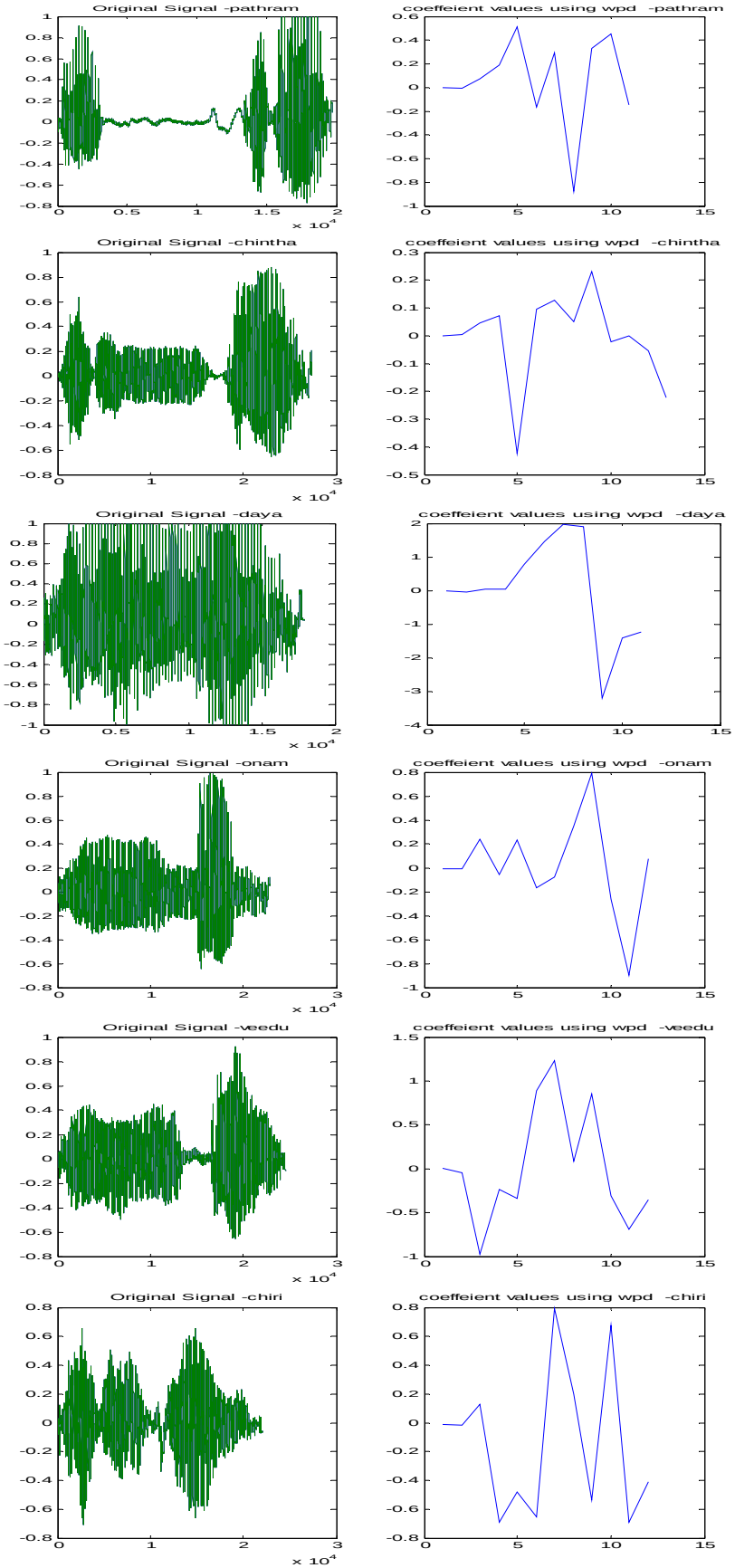


Figure 3. Decomposition of different words

An overall recognition accuracy of 87.5% is obtained by using the combination of Wavelet packet decomposition and Artificial neural networks.

VI. CONCLUSION AND FUTURE WORK

The research in the domain of audio and speech recognition in various languages is still going on and is a challenging task. In this paper, we have used a hybrid system consisting of wavelet packet decomposition and artificial neural networks for recognizing speaker independent isolated spoken words in Malayalam. The computational complexity and feature vector size is successfully reduced to a great extent by using wavelet packet decomposition and an overall recognition accuracy of 87.5% is obtained from this work. In this experiment, we have used a limited number of samples. The vocabulary size can be increased to obtain more recognition accuracy. Neural network classifiers are well suited for speech recognition and it provides good accuracies. As an extension of this work, alternate classifiers like Support Vector Machines, Genetic algorithms, Fuzzy set approaches etc. can also be used and a comparative study of these can be performed.

ACKNOWLEDGMENT

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

REFERENCES

[1] Rabiner L., Juang B. H., Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ., 1993.

[2] Picone J.W., "Signal Modelling Technique in Speech Recognition", Proc. of the IEEE, Vol. 81, No.9, pp.1215-1247, 1993.

[3] Suping Li, "Speech Denoising Based on Improved Discrete Wavelet Packet Decomposition", Proceedings of the International Conference on Network Computing and Information Security, pp. 415-419, 2011.

[4] Ting W., Guo-zheng Y., Banghua Y., Hong S., " EEG Feature Extraction based on Wavelet Packet Decomposition for Brain Computer Interface", Measurement, 41(6), 618-625, 2008.

[5] Sonia Sunny, David Peter S, K Poulose Jacob, "Discrete Wavelet Transforms and Artificial Neural Networks for Recognition of Isolated Spoken Words", . International Journal of Computer Applications, 38(9) , pp. 9-13, January 2012.

[6] Chan Woo S., Peng Lin C., Osman R., "Development of a Speaker Recognition System using Wavelets and Artificial Neural Networks", Proceedings of International Symposium on Intelligent Multimedia, Video and Speech processing, 413-416, 2001.

[7] Fecit Science and Technology Production Research Center, "Wavelet Analysis and Application by MATLAB6.5 [M]", Electronics Industrial Press, Beijing, 2003.

[8] Y. Hao, X. Zhu, "A New Feature in Speech Recognition based on Wavelet Transform", Proc. IEEE 5th Inter. Conf. on Signal Processing, vol 3, 2000.

[9] Freeman J. A, Skapura D. M.," Neural Networks Algorithm, Application and Programming Techniques", Pearson Education, 2006.

[10] Economou K., Lymberopoulos D., "A New Perspective in Learning Pattern Generation for Teaching Neural Networks", Volume 12, Issue 4-5, pp. 767-775, 1999.

[11] Eiji Mizutani, James W. Demmel,"On Structure-exploiting Trust Region Regularized Nonlinear Least Squares Algorithms for Neural-Network Learning", Neural Networks, Volume 16, Issue 5-6, pp. 745-753, 2003.

[12] Anil K. Jain, Robert P.W. Duin, Jianchang Mao, "Statistical Pattern Recognition: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence", Vol. 22., pp.4-37, 2000

[13] Hu Dingyin, Li Wei, Chen Xi , "Feature Extraction of Motor Imagery EEG Signals based on Wavelet Packet Decomposition", Proceedings of the 2011 IEEE International Conference on Complex Medical Engineering , 694-697 , 2011.

[14] Sonia Sunny, David Peter S, K Poulose Jacob, " Optimal Daubechies Wavelets for Recognizing Isolated Spoken Words with Artificial Neural Networks Classifier", International Journal of Wisdom Based Computing, Vol. 2(1), pp. 35-41, 2012.