Discrete Wavelet Transforms and Artificial Neural Networks for Recognition of Isolated Spoken Words

Sonia Sunny
Dept. of Computer Science
CUSAT
Kochi-682022, India

David Peter S
School of Engineering
CUSAT
Kochi-682022, India

K Poulose Jacob
Dept. of Computer Science
CUSAT
Kochi-682022, India

ABSTRACT

Speech recognition is a fascinating application of Digital Signal Processing and has many real-world applications. In this paper, a speech recognition system is developed for isolated spoken words using Discrete Wavelet Transforms (DWT) and Artificial Neural Networks (ANN). Speech signals are one-dimensional and are random in nature. Isolated words from Malayalam, one of the four major Dravidian languages of southern India are chosen for recognition. Daubechies wavelets are employed here. A multi-layer neural network trained with back propagation training algorithm is used for classification purpose. The proposed method is implemented for 50 speakers uttering 20 isolated words each. The experimental results show good recognition accuracy and the efficiency of combining these two techniques.

General Terms

Speech Recognition, Feature Extraction, Pattern Recognition, Recognition accuracy.

Keywords

Discrete Wavelet Transforms, Artificial Neural Networks, Speech Database, Classification, Daubechies Wavelets.

1. INTRODUCTION

Speech signals are one of the most important means of communication among the human beings. Speech recognition has tremendous growth over the last five decades due to the advances in signal processing, algorithms, new architectures and hardware [1]. Speech processing and recognition are intensive areas of research due to the wide variety of applications. Speech recognition is involved in our daily life activities like mobile applications, weather forecasting, agriculture, healthcare, video games etc. [2]. The performance of a speech recognition system is measurable and the most widely used method for measuring the performance is calculating the recognition accuracy. Speech recognition is a complicated task. This is due to the fact that speech signals vary a lot because of noise, speaker variations and differences between the training and testing environments, such as the microphones used [3].

Since speech signals are non stationary in nature, many parameters affect the speech recognition process. A lot of research work has gone into speech recognition. But there is requirement of much more research and development in this field. Speech recognition system usually involves some kind of classification or recognition based upon speech features. The speech features are usually obtained via time-frequency representations [4]. In this work, the speech recognition system is divided into 3 modules. The first module deals with the creation of the spoken words database. In the second module, the features from these speech signals are extracted. The classification process is done in the third module. Among these three stages, feature extraction part plays an important role because good features improve the recognition accuracy.

This paper is organized as follows. Section 2 explains the isolated spoken words database. In the subsequent section, the theory of feature extraction is reviewed followed by the concepts of discrete wavelet transforms used during this stage. The classification stage using artificial neural networks is discussed in section 4. Section 5 presents the detailed analysis of the experiments done and the results obtained. Last section contains the conclusions and future work.

2. ISOLATED WORDS DATABASE

A database is created for Malayalam language using 50 speakers. Each speaker utters 20 words. We have used twenty male speakers and thirty female speakers for creating the database. 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 these 20 isolated spoken words under the same configuration. Our database consists of a total of 1000 utterances of the spoken words. The spoken words are preprocessed, numbered and stored in the appropriate classes in the database. The spoken words and their International Phonetic Alphabet (IPA) format are shown in Table 1.

Table 1. Words 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:mΛrə/
പാവ	Paava	/pa:və/

ഗീതം	Geetham	/gi:θʌm/
പത്രം	Pathram	/pʌθrəm/
ദയ	Daya	/ðʌjə/
ചിന്ത	Chintha	/tʃinθʌ/
കടൽ	Kadal	/kʌdʌl/
ഓണം	Onam	/əunʌm/
ചിരി	Chiri	/tʃiri/
വീട്	Veedu	/vi:də/
കുട്ടി	Kutti	/kuţi/
മരം	Maram	/mʌrəm/
മയിൽ	Mayil	/mʌjil/
ലോകം	Lokam	/ləukʌm/
മൗനം	Mounam	/maunəm/
വെള്ളം	Vellam	/ve!!ʌm/
അമ്മ	Amma	/ʌmmʌ/

3. FEATURE EXTRACTION MODULE

Feature extraction is the process of extracting the relevant features from speech signals for further processing. During feature extraction, the short-time temporal or spectral parameters of speech are extracted and those features relevant for classification are retained. It is one of the important components of any type of recognition systems because the recognition accuracy depends on the features extracted. Researchers have experimented with many different types of methods for use in 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. Moreover, many of these methods accept signals stationary within a given time frame. So, it is difficult to analyze the localized events correctly.

3.1 Discrete Wavelet Transforms

A wavelet can be thought of as an extension of the classic Fourier transform to overcome the resolution problem. A wavelet transform works on a multi-scale basis instead of a single scale time or frequency basis. Wavelet transform decomposes a signal into a set of basic functions called wavelets. The Discrete Wavelet Transform (DWT) is any wavelet transform for which the wavelets are discretely sampled and is a special case of the wavelet transform that provides a compact representation of a signal in time and frequency that can be computed efficiently. It is a relatively recent and computationally efficient technique for extracting information about non-stationary signals like audio. The wavelet transform is a multi- resolutional, multi-scale analysis, which has been shown to be very well suited for speech processing. The extracted wavelet coefficients provide a compact representation that shows the energy distribution of the signal in time and frequency. Pattern recognition rate is improved by this method. The purpose of using the DWT is to benefit from its localization property [5] in the time and frequency domains. The Discrete Wavelet Transform is defined by the following equation.

$$W(j, K) = \Sigma_j\Sigma_k X\ (k)\ 2^{-j/2}\Psi\ (2^{-j}n-k) \qquad (1)$$

Where Ψ (t) is the basic analyzing function called the mother wavelet. The functions with different region of support that are used in the transformation process are derived from the mother wavelet. DWT is used to obtain a time-scale representation of the signal by means of digital filtering techniques. The original signal passes through two complementary filters, namely low-pass and high-pass filters. In speech signals, low frequency components known as the approximation coefficients h[n] are of greater importance than high frequency signals known as the detail coefficients g[n] as the low frequency components characterize a signal more than its high frequency components [6]. The wavelet decomposition tree is shown in figure 1.

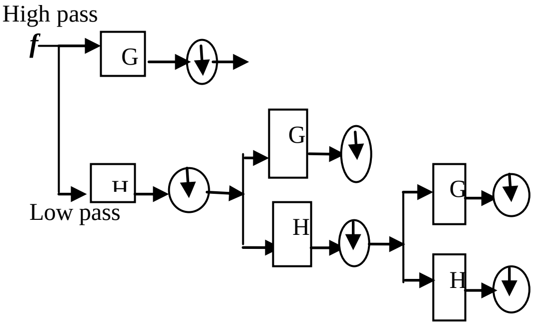


Figure 1. Wavelet decomposition tree

The low frequencies sequence of the first level forms as an input to the second stage. The discrete time domain signal is subjected to successive low pass filtering and high pass filtering to obtain DWT [5]. This algorithm is called the Mallat algorithm [7]. At each decomposition level, the half band filters produce signals spanning only half the frequency band. The filtering and decimation process is continued until the desired level is reached. 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 [8]. The DWT of the original signal is then obtained by concatenating all the coefficients starting from the last level of decomposition. Though it is possible to decompose the high frequency and low frequency components as in the case of wavelet packet decomposition, decomposing only the low frequency components gives better recognition accuracy [9]. The successive high pass and low pass filtering of the signal can be obtained by the following equations.

$$Y_{high}[k]= \Sigma_n x[n]g[2k-n] \qquad (2)$$

$$Y_{low}[k]= \Sigma_n x[n]h[2k-n] \qquad (3)$$

Where Y_{high} (detail coefficients) and Y_{low} (approximation coefficients) are the outputs of the high pass and low pass filters obtained by sub sampling by 2 [10]. Here the time resolution is halved, but since the output has half the frequency band of the input, the frequency resolution has been doubled. The filter analysis block diagram is given in figure 2.

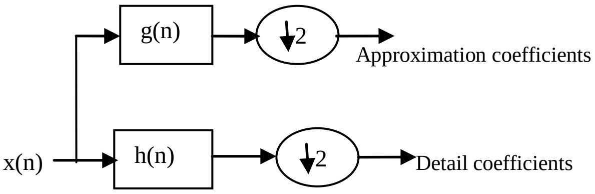


Figure 2. Filter Analysis block diagram

DWT is also related to a multi-resolution framework. Now, DWT is more popular in the field of Digital Signal Processing (DSP) due to its multiresolution capability [11]. Also it has the property of constant Q, which is one of the demands of many signal processing applications, especially in the processing of the speech signals as human's hearing system is constant Q perceptional [12].

4. SPEECH CLASSIFICATION MODULE

Speech recognition is basically a pattern recognition problem. Pattern recognition deals with mathematical and technical aspects of classifying different objects. 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. During classification stage, decision is taken based on all the similarity measures after trained using information relating to known patterns and the similarity measured from the pattern. In this study also, neural networks are used as the classifier. Neural networks can perform pattern recognition; handle incomplete data and variability well.

4.1 Artificial Neural Networks

Neural networks are an artificial intelligence method for modeling complex non-linear functions. Neural networks can be viewed as massively parallel computing systems consisting of an extremely large number of simple processors called nodes with many interconnections. In the neural network mode, the nodes are artificial neurons and directed edges with weights are connections between neuron outputs and neuron inputs. Inspired by the human brain, neural network models attempt to use some organizational principles such as learning, generalization, adaptivity, fault tolerance etc. [13]. During the learning process, network architecture and connection weights are updated for proper classification. The main advantage of using neural networks is that they have the ability to learn complex nonlinear input-output relationships by using training procedures and adapting themselves to the data. Algorithms based on neural networks are well suitable for addressing speech recognition tasks. If x_1, x_2, x_3, ……x_n are the inputs and w_1, w_2, w_3…w_n are the corresponding weights, then the total input to the next neuron or the output neuron I is calculated by the summation function

$$I = w_1x_1 + w_2x_2 + \ldots\ldots.+ w_nx_n = \sum_{I=1}^{n} w_ix_i \qquad (4)$$

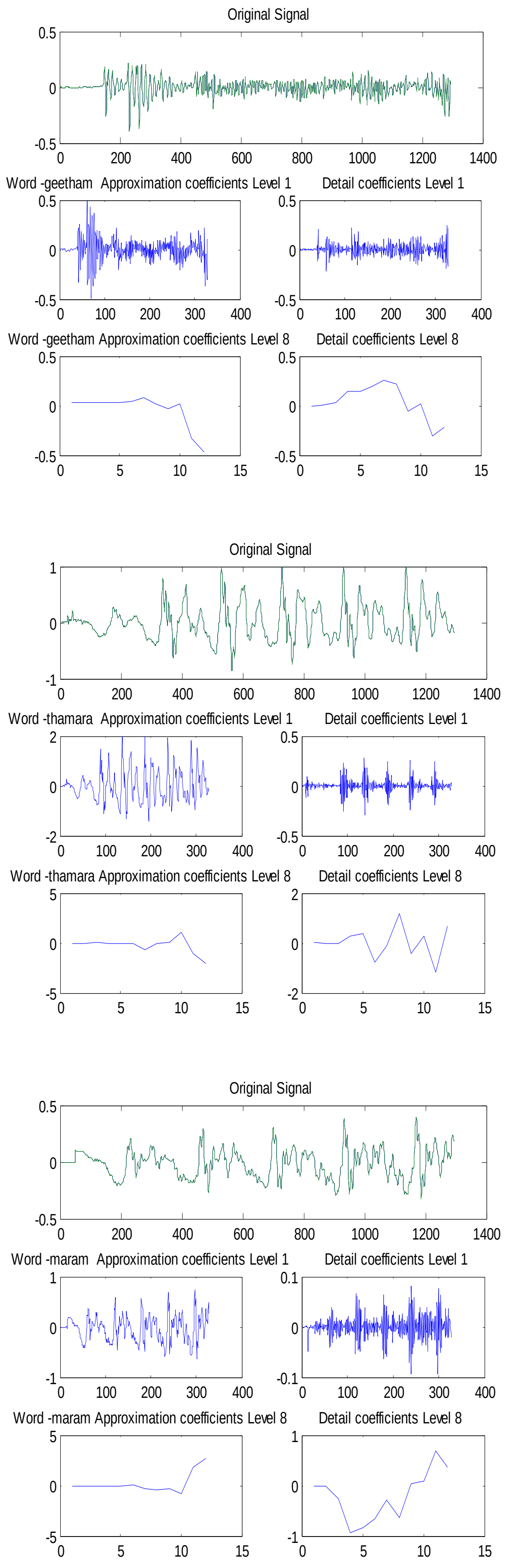
The result of the summation function, which is the weighted sum, is transformed to a working output through an algorithmic process called the activation function or the transfer function.

The feed-forward network is the most commonly used type of neural network used in the area of pattern classification, which includes multilayer perceptron. 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 in this case 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 error back propagation correction algorithm. After prolonged 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 recognition accuracy depends on the feature vectors obtained, training samples selected and the ability of the classifier to learn from these samples. 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 [12].

5. EXPERIMENTAL RESULTS

There are different types of wavelet families [5] such as Daubechies, Symmlet, Coiflet etc. Selection of the suitable wavelet and the number of decomposition levels play an important role in obtaining good recognition accuracy in speech recognition. Among the various wavelet bases, the most popular wavelets that represent foundations of Digital Signal Processing called the Daubechies wavelets are used because of its orthogonality property and efficient filter implementation [14]. In this paper, we are using db4 type of mother wavelet feature extraction purpose. The speech samples in the database are successively decomposed into approximation and detailed coefficients. The approximation coefficients from eighth level are used to create the feature vectors for each spoken word and the number of approximation coefficients obtained at the eighth level is twelve.

The feature vectors obtained using DWT are given as the input to the ANN classifier. Here we have divided the database into three. 70% of the data is used for training, 15% for validation and 15% for testing. MLP architecture is used for the classification scenario. Using this network, the classifier could successfully recognize the spoken words. After testing, the corresponding accuracy of the isolated spoken words is obtained. The results obtained clearly shows the efficiency of neural networks in classifying the extracted coefficients. Results obtained using DWT and ANN is given below. The original signal and various decomposition level coefficient values of 5 Malayalam words geetham, thamara, maram, vellam and amma are shown in figure 3 and the performance analysis based on error percentage is given in table 2.



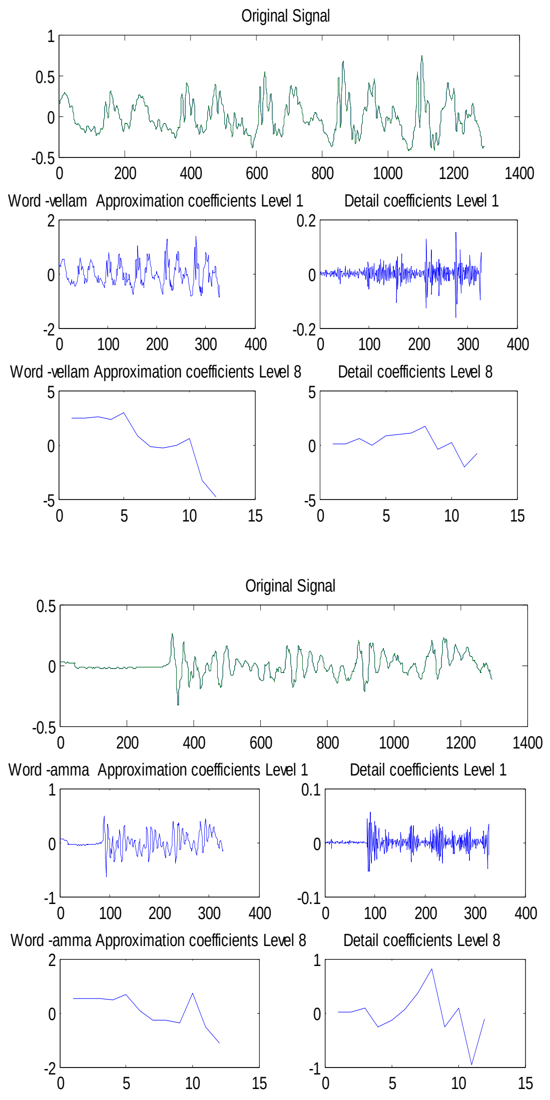


Figure 3. Different decomposition levels of spoken words

Table 2. Performance Analysis based on Error Percentage

	Mean Squared Error (MSE)	Percent error (%E)
Training	**5.29349 e-3**	**11.57**
Validation	**4.0099 e-3**	**9.33**
Testing	**4.67599 e-3**	**10.00**

By using Discrete Wavelet Transforms as the feature extraction tool and Artificial Neural Networks as classifier, the overall recognition accuracy obtained is 90%.

6. CONCLUSION AND FUTURE WORK

In this paper, an Automatic Speech Recognition system is designed for isolated spoken words in Malayalam using discrete wavelet transforms and artificial neural networks. A better performance of identification with high recognition accuracy of 90% is obtained from this study. The computational complexity and feature vector size is successfully reduced to a great extent by using discrete wavelet transforms. Thus a wavelet transform is an elegant tool for the analysis of non-stationary signals like speech. The experiment results show that this hybrid architecture using discrete wavelet transforms and neural networks could effectively extract the features from the speech signal for automatic speech recognition. In this experiment, we have used a limited number of samples. Recognition rate can be increased by increasing the number of samples. The neural network classifier which is used in the experiment provides good accuracies. 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 as an extension of this study.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

[1] Lawrence R., 1997, Applications of Speech Recognition in the Area of Telecommunications, Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding.

[2] Kuldeep Kumar, R. K. Aggarwal, 2011, Hindi Speech Recognition System Using Htk, International Journal of Computing and Business Research, Volume 2 Issue 2.

[3] Evandro B. Gouva, Pedro J. Moreno, Bhiksha Raj, Thomas M. Sullivan, Richard M. Stern, 1996, Adaptation and Compensation: Approaches to Microphone and Speaker Independence in Automatic Speech Recognition, Proc. DARPA Speech Recognition Workshop.

[4] Jiang Hai, Er Meng Joo, 2003, Improved Linear Predictive Coding Method for Speech Recognition, ICICS-PCM, Singapur.

[5] S. Mallat, A, 1999, Wavelet Tour of Signal Processing (second edition),Academic Press.

[6] S. Kadambe, P. Srinivasan, 1994, Application of Adaptive Wavelets for Speech, Optical Engineering Vol 33(7).

[7] S .G. Mallat, 1989, A Theory for Multiresolution Signal Decomposition: The Wavelet Representation, IEEE Transactions on Pattern Analysis And Machine Intelligence, Vol.11.

[8] Elif Derya Ubeyil, 2009, Combined Neural Network model Employing wavelet Coefficients for ECG Signals Classification, Digital signal Processing, Vol 19.

[9] Sonia Sunny, David Peter S., K. Poulose Jacob, 2011, Wavelet Packet Decomposition and Artificial Neural Networks based Recognition of Spoken Digits, International journal of machine intelligence, Vol.3, issue 4.

[10] M. Vetterli, C. Herley, 1992, Wavelets and Filter Banks: Theory and Design, IEEE Transactions on Signal Processing,Vol.40.

[11] N. S. Nehe and R. S. Holambe, 2008, New Feature Extraction Methods Using DWT and LPC for Isolated Word Recognition, Proceedings of TENCON 2008 - 2008 IEEE Region 10 Conference,

[12] Y. Hao, X. Zhu, 2000, A new feature in speech recognition based on wavelet transform, Proc. IEEE 5th Inter. Conf. on Signal Processing, vol 3.

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

[14] Hu Dingyin, Li Wei, Chen Xi, 2011, Feature extraction of motor imagery EEG signals based on wavelet packet decomposition, Proceedings of the 2011 IEEEIICME International Conference on Complex Medical Engineering.