A Comparative Study of Wavelet Based Feature Extraction Techniques in Recognizing Isolated Spoken Words

Sonia Sunny, David Peter S., and K Poulose Jacob
Cochin University of Science and Technology, Kochi, India
Email: sonia.deepak@yahoo.co.in; davidpeter@cusat.ac.in; kpj@cusat.ac.in

***Abstract*—Speech is a natural mode of communication for people and speech recognition is an intensive area of research due to its versatile applications. This paper presents a comparative study of various feature extraction methods based on wavelets for recognizing isolated spoken words. Isolated words from Malayalam, one of the four major Dravidian languages of southern India are chosen for recognition. This work includes two speech recognition methods. First one is a hybrid approach with Discrete Wavelet Transforms and Artificial Neural Networks and the second method uses a combination of Wavelet Packet Decomposition and Artificial Neural Networks. Features are extracted by using Discrete Wavelet Transforms (DWT) and Wavelet Packet Decomposition (WPD). Training, testing and pattern recognition are performed using Artificial Neural Networks (ANN). The proposed method is implemented for 50 speakers uttering 20 isolated words each. The experimental results obtained show the efficiency of these techniques in recognizing speech.**

***Index Terms*—speech recognition, feature extraction, discrete wavelet transforms, wavelet packet decomposition, classification, artificial neural networks.**

I. INTRODUCTION

Automatic speech recognition (ASR) is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of text and is a broad area of research. 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]. The human vocal tract and articulators are biological organs with nonlinear properties, which is affected by factors ranging from gender to upbringing to emotional state. As a result, there is variation in terms of their accent, pronunciation, articulation, roughness, nasality, pitch, volume, and speed. Speech patterns can be further distorted by background noise and echoes, as well as electrical characteristics [2]. All these sources of variability make speech recognition a very complex problem. The performance of a speech recognition system is measurable and the most widely used method for measuring the performance is calculating the recognition accuracy. Many parameters affect the accuracy of the speech recognition system. Wavelets and wavelet packet analysis are found to be very effective in speech processing and signal processing applications.

Speech processing and recognition are intensive areas of research due to the wide variety of applications like mobile applications, weather forecasting, agriculture, healthcare, automatic translation, robotics, video games, transcription etc [3]. Speech recognition system generally carries some kind of classification or recognition based upon speech features which are usually obtained via time-frequency representations [4]. There are different phases for speech recognition using a computer. In this work, the speech recognition process is divided into 3 modules namely, creating the speech database, feature extraction and classification. Among these stages, the front end processing part called the feature extraction is a key, because better feature is good for improving recognition rate. While designing a wavelet based speech recognition system, the main issue is to choose the optimal wavelets and the appropriate number of decomposition levels. Likewise, the number of hidden layers and the number of hidden neurons chosen plays an important role in obtaining good recognition accuracy using neural networks.

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

II. ISOLATED WORDS DATABASE FOR MALAYALAM

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 speech samples are taken from speakers of age between 20 and 30. 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. The 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 in Malayalam, their representation in English, their International Phonetic Alphabet (IPA) format and English translation are shown in Table 1.

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

Words in Malayalam	Words in English	IPA format	English Translation
കേരളം	Keralam	/kēra!am/	Kerala
വിദ്യ	Vidya	/vidjə/	Knowledge
പൂവ്	Poovu	/pu:və/	Flower
താമര	Thamara	/θa:mΛrə/	Lotus
പാവ	Paava	/pa:və/	Doll
ഗീതം	Geetham	/gi:θΛm/	Song
പത്രം	Pathram	/pΛθrəm/	News paper
ദയ	Daya	/ðΛjə/	Mercy
ചിന്ത	Chintha	/tʃinθΛ/	Thought
കടൽ	Kadal	/kΛdΛl/	Sea
ഓണം	Onam	/əunΛm/	Onam
ചിരി	Chiri	/tʃiri/	Smile
വീട്	Veedu	/vi:də/	House
കുട്ടി	Kutti	/kuţi/	Child
മരം	Maram	/mΛrəm/	Tree
മയിൽ	Mayil	/mΛjil/	Peacock
ലോകം	Lokam	/ləukΛm/	World
മൗനം	Mounam	/maunəm/	Silence
വെള്ളം	Vellam	/ve!!Λm/	Water
അമ്മ	Amma	/ΛmmΛ/	Mother

III. SPEECH FEATURE EXTRACTION

Transforming the input data into a set of features is called feature extraction. During feature extraction step, the original speech signal is converted into a sequence of feature vectors and unnecessary information from the signal are stripped and the properties of the signal which are important for the pattern recognition task are converted to a format that simplifies the distinction of the classes. The dimension of the data is reduced during feature extraction. So, feature extraction plays a vital role in speech recognition process. The feature vector sequences obtained are the inputs to the classification step of a speech recognition system.

The technique selected for feature extraction plays a vital role in the speech recognition rate. 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. By using wavelets, the size of the feature vector can be reduced when compared to other methods and thus the computational complexity also can be successfully reduced.

A. *Discrete Wavelet Transform*

Wavelet transforms were introduced to address the problems associated with non-stationary signals like speech. A Wavelet transform decomposes a signal into a set of basic functions called wavelets. DWT 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. They are well suitable for processing signals like speech because of their efficient time-frequency localization [5] and the multi-resolutional, multi-scale analysis characteristics of the wavelet representations.

The Discrete Wavelet Transform is defined by the following equation [5]

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

where Ψ *(t)* is the basic analyzing function called the mother wavelet. Other functions are derived from this mother wavelet by translation and dilation operations. In DWT, 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]. In speech signals, low frequency components known as the approximations h[n] are of greater importance than high frequency signals known as the details g[n] as the low frequency components characterize a signal more than its high frequency components [7]. The successive high pass and low pass filtering of the signal can be obtained by the following equations.

$$Y_{high}[k] = \sum_{n} x[n] g[2k-n] \tag{2}$$

$$Y_{low}[k] = \sum_{n} x[n] h[2k-n] \tag{3}$$

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. The filtering is continued until the desired level is reached according to Mallat algorithm [8]. 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 [9].

B. The Wavelet Packet Decomposition

The wavelet packet transform is a direct expansion of the discrete wavelet transform and is a more detailed method than discrete wavelet transform. Wavelet packet decomposition can also provide a multi-level time-frequency decomposition of signals. Here also, the signal is decomposed into low frequency components and high frequency components at each level. In WPD, the approximation and detail coefficients are decomposed to get new low resolution approximation and detail coefficients. The difference between DWT and WPD is that the discrete wavelet transform is applied to the low pass result only whereas the wavelet packet decomposition applies the transform step to both the low pass and the high pass result.

It allows simultaneous use of long-time interval for low-frequency information and short-time interval for high-frequency information [10]. In wavelet packet decomposition, each detail coefficient vector is also decomposed into two parts using the same approach as in approximation vector splitting. When a signal is decomposed into sub-bands using wavelet transform, approximation components contain the characteristics of a signal and high frequency components are related with noise and disturbance in a signal [11]. Though removing the high frequency contents retain the features of the signal, sometimes it may contain useful features of the signal. So both the high and low frequency components are decomposed in WPD.

IV. Speech Classification

With the recent advances in the computing technology, many pattern recognition tasks have become automated. Speech recognition is basically a pattern recognition problem. An important application of neural networks is pattern recognition. Since neural networks are good at pattern recognition, many early researchers applied neural networks for speech pattern recognition. In this study also, we are using neural networks as the classifier. Neural networks can perform pattern recognition; handle incomplete data and variability well [12]. Artificial Neural networks are well suited for speech recognition due to their fault tolerance and non-linear property. Their ability to learn by example makes them very flexible and powerful. The increasing popularity of neural network models to solve pattern recognition problems has been primarily due to their seemingly low dependence on domain-specific knowledge and due to the availability of efficient learning algorithms for practitioners to use [13].

A. Neural Networks Classifier

Artificial neural networks have been investigated for many years in the hope that speech recognition can be done similar to human beings. A Neural Network is a massively parallel-distributed processor made up of simple processing units. It can store experimental knowledge and make it available for use. Inspired by the structure of the brain, a neural network consists of a set of highly interconnected entities, called nodes designed to mimic its biological counterpart, the neurons. Each neuron accepts a weighted set of inputs and produces an output [14]. 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.

The architecture of the Multi Layer Perceptron (MLP) network, which consists of an input layer, one or more hidden layers, and an output layer, is used here. The algorithm used is the back propagation training algorithm which is a systematic method for training multi-layer neural networks. This is a multi-layer feed forward, supervised learning network based on gradient descent learning rule. 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 will eventually establish the input-output relationships through the adjusted weights on the network [15]. In most networks, the principle of learning a network is based on minimizing the gradient of error [16]. After training the network, it is tested with the dataset used for testing.

V. Experiments and Results

Since there are different mother wavelets of different wavelet families available, the choice of the wavelet family and the mother wavelet plays an important role in the recognition accuracy. The most popular wavelets that represent foundations of digital signal processing called the Daubechies wavelets are used here. Among the Daubechies family of wavelets, the db4 type of mother wavelet is used for feature extraction. Daubechies wavelets are found to perform better than the other wavelet families based on recognition accuracy [17]. The speech samples in the database are successively decomposed into approximation and detailed coefficients. Less frequency components from level 12 is used to create the feature vectors for each spoken word in the case of discrete wavelet transforms. The speech signal is decomposed up to 12 levels in the case of wavelet packet decomposition also. The same signals used for DWT are also used for WPD.

Speech recognition is a multi class classification problem. So the developed feature vectors from both the methods are given to an ANN since it can handle multi class parameter classification. 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. It uses one input layer, one hidden layer and one output layer. Using this network, the feature vector set obtained is trained first and then they are tested. From the results obtained, it is found that the MPL structure could successfully recognize the spoken words. After testing, the corresponding accuracy of each spoken word is obtained.

Results obtained using DWT and ANN are given below. The original signal and the 12th level approximation and detail coefficients of spoken words pava and daya are shown in Fig. 1 and Fig. 2.

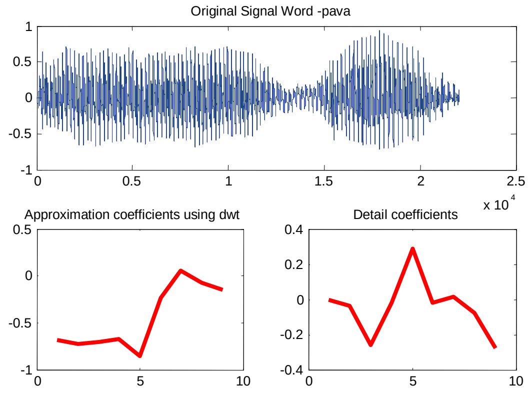

Figure 1. Decomposition of word pava using DWT

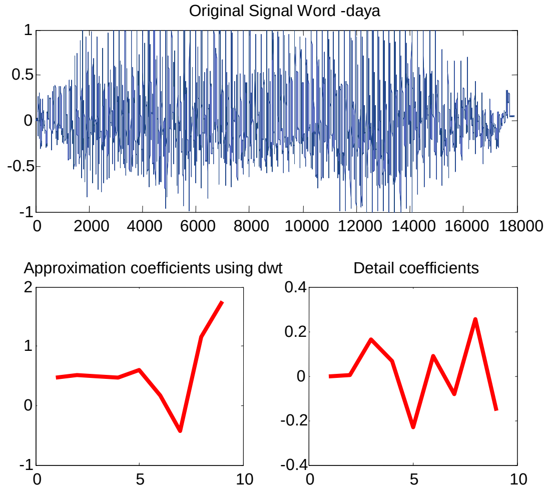

Figure 2. Decompsition of word daya using DWT

Results obtained using WPD and ANN are given below. The original signal and the 12th level decomposition coefficients of spoken words pava and daya are shown in Fig. 3 and Fig. 4.

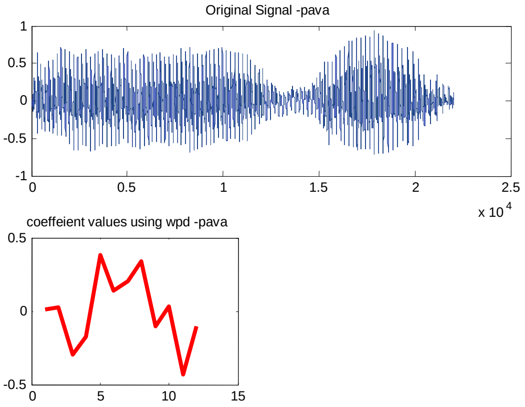

Figure 3. Decompsition of word pava using WPD

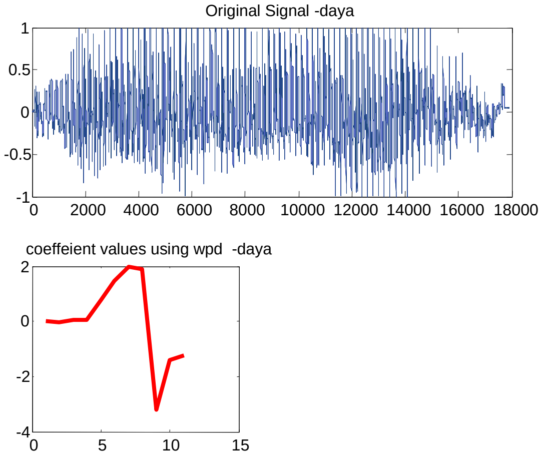

Figure 4. Decompsition of word daya using WPD

The overall recognition accuracies obtained using DWT and WPD is shown in table 2.

TABLE II. COMPARISON

Feature Extraction Method	Recognition Accuracy (%)
DWT	90.00
WPD	87.50

Using DWT and ANN, the overall recognition accuracy obtained is 90% and by using WPD and ANN, the overall recognition accuracy obtained is 87.5%. This shows that both these methods perform well. But DWT gives slightly better results than WPD.

VI. CONCLUSIONS AND FUTURE WORK

In this work, a speech recognition system is designed for solated spoken words in Malayalam. A comparative study of two major wavelet based feature extraction methods such as discrete wavelet transforms and wavelet packet decomposition are performed here. These methods are combined with neural networks for classification purpose. The performance of both these techniques are tested and evaluated. Both the techniques are found to be efficient in recognizing speech. The accuracy rate obtained by the combination of DWT and neural networks is higher than that of the architecture using WPD and neural networks. The computational complexity and the feature vector size is successfully reduced to a great extent by using 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. For future work, the vocabulary size can be increased to obtain more recognition accuracy. Though the neural network classifier which is used in this 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.

REFERENCES

[1] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition,* 1 st ed. Englewood Cliffs, NJ: Prentice-Hall, 1993, ch. 1, pp.37-45.

[2] T. F. Quatieri, *Discrete-Time Speech Signal Processing Principles and Practice,* 1st ed. Upper Saddle River, NJ: Prentice-Hall, 2001, ch. 3.

[3] K. Kumar and R. K. Aggarwal. (May 2011). Hindi speech recognition system using Htk. *International Journal of Computing and Business Research*. [Online]. 2(2). Available: http://www.researchmanuscripts.com/PapersVol2N2/IJCBRVOL2N2P3.pdf

[4] J. Hai and E. M. Joo, "Improved linear predictive coding method for speech recognition," in *Proc. 4 th International Conf. Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia*, Singapore, 2003, pp. 1614-1618.

[5] S. Mallat, *A Wavelet Tour of Signal Processing,* 3rd ed. U.S.: Academic Press, 2008, ch. 4.

[6] S. Chan Woo, C. P. Lin, and R. Osman, "Development of a speaker recognition system using wavelets and artificial neural networks," *in Proc. International Symposium on Intelligent Multimedia, Video and Speech processing,* Hong Kong, 2001, pp. 413-416.

[7] S. Kadambe and P. Srinivasan, "Application of adaptive wavelets for speech", *Optical Engineering*, vol. 33, no.7, pp. 2204-2211, 1994.

[8] S .G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 674-693, July 1989.

[9] E. D. Ubeyil, "Combined neural network model employing wavelet coefficients for ECG signals classification," *Digital Signal Processing*, vol. 19, no. 2, pp. 297-308, March 2009.

[10] W. Ting, Y. G. Zheng, Y. Bang-hua, and S. Hong, "EEG feature extraction based on wavelet packet decomposition for brain computer interface," *Measurement*, vol. 41, no. 6, pp. 618-625, July 2008.

[11] BC Li and JS. Luo, *Wavelet Analysis and its Applications,* Beijing: Electronics Engineering Press, 2003

[12] S. N. Sinanandam, S. Sumathi and S. N Deepa, *Introduction to Neural Networks using MATLAB 6.0,* McGraw-Hill, 2006, ch.2, pp. 12.

[13] J. K. Basu, D. Bhattacharyya, and T. Kim. (April 2010). Use of artificial neural network in pattern recognition. *International Journal of Software Engineering and its Applications* [Online]. 4(2). pp. 23-33. Available: http://www.sersc.org/journals/IJSEIA/vol4_no2_2010/3.pdf

[14] J. A. Freeman and D. M. Skapura, *Neural Networks Algorithm Application and Programming Techniques*, Pearson Education, 2006, ch. 2.

[15] G .P .K. Economou, D. Lymberopoulos, K. Spyropoulos, and P. D. Goumas, "A new perspective in learning pattern generation for teaching neural networks," *Neural Networks*, vol. 12, no. 4-5, pp. 767-775, June 1999.

[16] E. Mizutani and J. W. Demmel, "On structure-exploiting trust region regularized nonlinear least squares algorithms for neural-network learning," *Neural Networks*, vol. 16, no. 5-6, pp. 745-753, 2003.

[17] S. Sunny, David Peter S, and K P. Jacob. Performance Analysis of Different Wavelet Families in Recognizing Speech, *International Journal of Engineering Trends and Technology* [Online] 4(4). pp.512-517, 2013

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.

Dr. 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.

Dr. 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.