2012 International Conference on Advances in Computing and Communications

Feature Extraction Methods based on Linear Predictive Coding and Wavelet Packet Decomposition for Recognizing Spoken Words in Malayalam

Sonia Sunny[1], David Peter S[2], K Poulose Jacob[3]
[1] Dept. of Computer Science, [2] School of Engineering, [3] Dept. of Computer Science
Cochin University of Science and Technology
Kochi-682022, India
E-mail: sonia.deepak@yahoo.co.in, davidpeter@cusat.ac.in, kpj@cusat.ac.in

Abstract— **Speech signals are one of the most important means of communication among the human beings. In this paper, a comparative study of two feature extraction techniques are carried out for recognizing speaker independent spoken isolated words. First one is a hybrid approach with Linear Predictive Coding (LPC) and Artificial Neural Networks (ANN) and the second method uses a combination of Wavelet Packet Decomposition (WPD) and Artificial Neural Networks. Voice signals are sampled directly from the microphone and then they are processed using these two techniques for extracting the features. Words from Malayalam, one of the four major Dravidian languages of southern India are chosen for recognition. Training, testing and pattern recognition are performed using Artificial Neural Networks. Back propagation method is used to train the ANN. The proposed method is implemented for 50 speakers uttering 20 isolated words each. Both the methods produce good recognition accuracy. But Wavelet Packet Decomposition is found to be more suitable for recognizing speech because of its multi-resolution characteristics and efficient time frequency localizations.**

Keywords- Speech Recognition; Feature Extraction; Linear Predictive Coding; Wavelet Packet Decomposition; Neural Networks.

I. INTRODUCTION

Human speech is a complex signal containing a multitude of parameters. Speech processing is a diverse field with many applications in which speech recognition is an important one [1]. Speech recognition has tremendous growth over the last five decades due to the advances in signal processing, algorithms, new architectures and hardware [2]. Since the human vocal tract and articulators are biological organs with nonlinear properties, these are affected by factors ranging from gender, emotional state etc. So there is wide difference in terms of their accent, pronunciation, articulation, roughness, nasality, pitch, volume, and speed. Moreover, speech patterns are also distorted by background noise and echoes, as well as electrical characteristics. Thus a speech recognition problem is a very complex task [3]. Due to the wide variety of applications like mobile applications, weather forecasting, agriculture, healthcare, automatic translation, robotics, video games, transcription etc, lot of research is being done in the area of speech processing and recognition [4].

Though different speech processing techniques and feature extraction methods are available, much more research and development is needed in this field especially in the area of speaker independent speech recognition. Selection of the feature extraction technique plays an important role in the recognition accuracy, since it is the main criteria for a good speech recognition system. Among the various feature extraction methods, two techniques namely linear predictive coding and wavelet packet decomposition are used here and a comparative study of these is performed here. Classification and recognition are performed using Multi Layer Perceptron (MLP) architecture.

II. METHODOLOGY

Signal modeling and pattern matching are two fundamental operations in a speech recognition system [5]. Feature extraction is an important part of signal modeling where speech signal is converted into a set of parameters called feature vectors. Pattern matching or classification is the task of finding parameter set from memory which closely matches the parameter set obtained from the input speech signal. In this work, we have divided the speech recognition process into three stages. The first stage is the creation of the database. Next one is the feature extraction stage wherein short time temporal or spectral parameters of speech signals are extracted. The third module is the classification stage wherein the derived parameters are compared with stored reference parameters and decisions are made based on some kind of a minimum distortion rule.

The outline of this study is as follows. The isolated spoken words database is explained in section 3. In section 4, the theory of various feature extraction techniques namely linear predictive coding and wavelet packet decomposition are reviewed. The pattern classification stage using artificial neural networks is described in section 5. Section 6 presents the detailed analysis of the experiments done and the results obtained. Conclusions are given in the last section.

III. WORDS DATABASE IN MALAYALAM

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

978-0-7695-4723-7/12 $26.00 © 2012 IEEE
DOI 10.1109/ICACC.2012.7

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 pre-processed, 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. WORDS 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: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/**

IV. FEATURE EXTRACTION TECHNIQUES USED

Feature extraction plays an important role in a recognition system because the recognition accuracy depends on the features extracted. In this section, the feature extraction techniques like LPC and WPD are explained.

A. *Linear Predictive Coding*

LPC is one of the most powerful speech analysis techniques used in audio and speech signal processing. It is a useful method for encoding good quality speech at a low bit rate and it extracts speech parameters like pitch formants and spectra. It provides a good model of the speech signal and LPC front-end processing has been used in a large number of speech recognition accuracies [6]. It is a powerful technique used for feature extraction purpose of speech signals since it characterizes the vocal tract well. The LPC analysis estimates the vocal tract resonance from a signal's waveform, removing their effects from the speech signal in order to get the source signal. The most important aspect of LPC is the linear predictive filter which allows the value of the next sample to be determined by a linear combination of previous samples [7]. At a particular time, k, the speech sample s(k) is represented as a linear sum of the n previous samples. This can be represented by the equation

$$S(k) = ak\text{-}1\ s(k\text{-}1) + ak\text{-}2\ s(k\text{-}2) + \ldots\ldots + ak\text{-}n\ s(k\text{-}n) \quad (1)$$

Where S (k) is the value of the signal at time (k). The coefficients aki are called the Linear Predictive Coding Coefficients. The coefficients can be analyzed to provide insight to the nature of the signal. Another important feature of LPC is that it minimizes the sum of the squared differences between the original speech and estimated speech signal over a finite duration. It produces a unique set of predictor coefficients which are normally estimated with every frame which is of usually 20 ms to 50 ms long. The predictor coefficients are represented by ak. Another important parameter is the gain (G). The transfer function of the time-varying digital filter is given by equation 2.

$$H(z) = \frac{G}{1 - \sum a_k z^{-k}} \quad (2)$$

Summation is computed starting at k = 1 up to p. Here LPC 10 is used. This means that only the first 10 coefficient are transmitted to the LPC synthesizer.

B. *Wavelet Packet Decomposition*

Wavelets are mathematical functions that cut up data into different frequency components. The wavelet transform can capture more time and frequency localized information than a Fourier Transform. Wavelet packet transform is computed using a time domain filtering along with a sub signal representation obtained from frequency components within each subband [8]. Wavelet transforms were introduced to address the problems associated with non-stationary signals like speech because of its multi- resolutional, multi-scale analysis characteristics and time frequency localizations [9].

Wavelets produce optimal time–frequency resolution in all frequency ranges because of its varying window size, being broad at low frequencies and narrow at high frequencies [10].

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 in WPD [11]. Wavelet packet decomposition is based on wavelet transform and decomposes a signal with the same widths in all frequency bands [12]. 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. 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. The WPD decomposition tree is shown in figure 1.

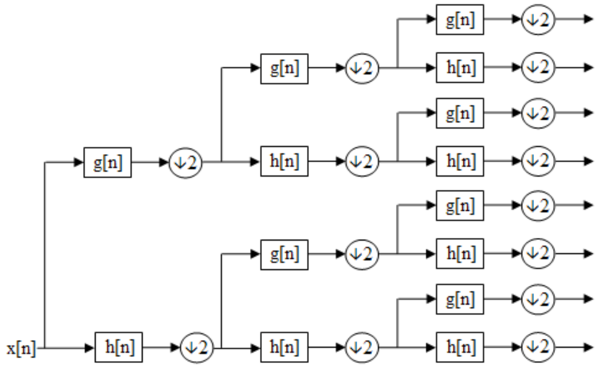

Fig. 1. WPD Decomposition Tree

V. CLASSIFICATION

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. 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. 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 [13].

A. Artificial Neural Networks

Nowadays, ANNs are utilized in wide ranges for their parallel distributed processing, distributed memories, error stability, and pattern learning 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 [14]. Inspired by the human brain, neural network models attempt to use some organizational principles such as learning, generalization, adaptivity, fault tolerance etc. [15].

In this work, we use 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 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.

VI. EXPERIMENTS AND RESULTS

In the LPC method, the input signals are broken into blocks or frames. In this system, voice signal is sampled using sampling frequency of 8 kHz. The 8000 samples in each second of speech signal are broken into 260 sample segments. That is, each frame represents 30.7 milliseconds of the speech signal. This frame size gives good results. The signal is passed through a low pass filter with bandwidth 1 KHz to split up the signal into voiced and unvoiced sound. Here the LPC order taken is 10. So the first ten LPC coefficients are used here.

In the case of WPD, a variety of wavelets are available for signal analysis. Moreover, 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 since it gives better results [16]. The speech samples in the database are successively decomposed and the feature vectors from level 12 are taken.

MLP architecture is used for the classification scenario in both the cases. The feature vectors obtained from WPD and LPC are given as input to the ANN classifier. Here we have divided the database into three. Out of the 1000 samples**,** 700 samples are used for training, 150 samples for validation and 150 samples for testing. The network uses one input layer, one hidden layer and an output layer. Using this network, the classifier could successfully recognize the spoken words.

The original signal and the twelfth level decomposition coefficients of four words keralam, thamara, poovu, and pathram obtained using WPD are shown in fig. 2, 3,4 and 5.

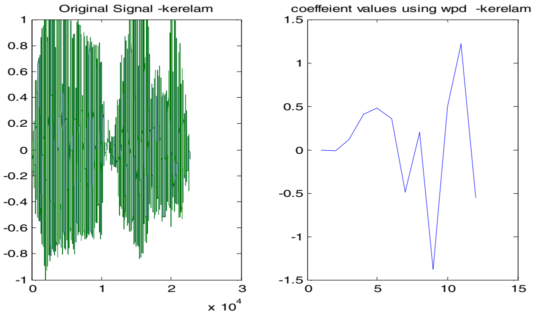

Fig. 2 Decomposition of word keralam using WPD

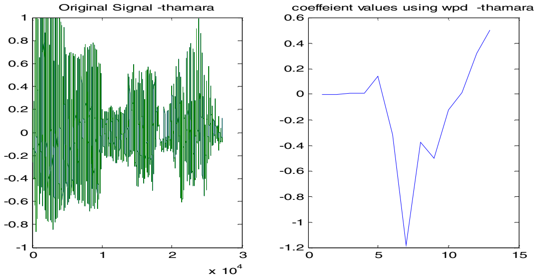

Fig. 3 Decomposition of word thamara using WPD

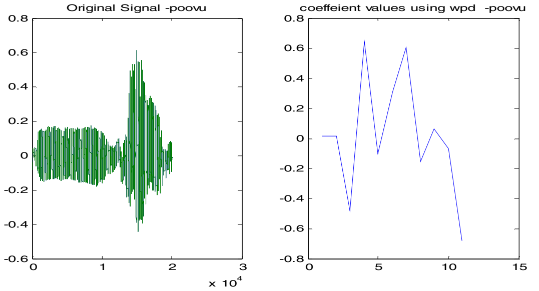

Fig. 4. Decomposition of word poovu using WPD

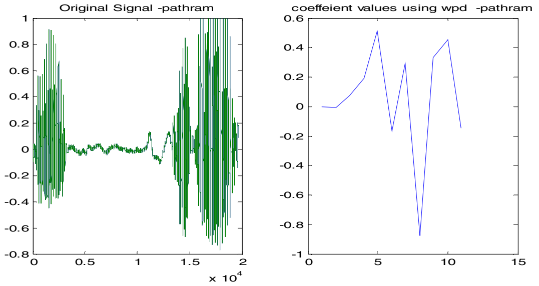

Fig. 5. Decomposition of word pathram using WPD

The overall recognition accuracies obtained using LPC and WPD are shown in table 2.

TABLE 2 COMPARISON OF RESULTS

Feature Extraction Method	Recognition Accuracy (%)
LPC	81.20
WPD	87.50

From the results obtained, it is clear that both the feature extraction methods give good recognition accuracies and are suited for speech recognition. But recognition rate is more in the case of WPD.

VII. CONCLUSIONS AND FUTURE WORK

This work gives a comparative assessment of two major feature extraction techniques such as linear predictive coding and wavelet packet decomposition for isolated spoken words in Malayalam. These methods are combined with neural networks for classification purpose. The performance of both these techniques are tested and evaluated. Both techniques are found to be efficient in recognizing speech. The accuracy rate obtained by using wavelet based technique is found to be more than LPC based method. The experiment results show that this hybrid architecture using wavelet packet decomposition and neural networks could effectively extract the features from the speech signal for automatic speech recognition. 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**.** Increasing the number of samples may also result in improving the recognition accuracy.

REFERENCES

[1] Joseph P Campbell, JR., "Speaker Recognition: A Tutorial" Proceedings of the IEEE, Vol. 85, No. 9, 1997.

[2] Lawrence R., "Applications of Speech Recognition in the Area of Telecommunications", Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 501-510, 1997.

[3] recognition.http://www.learnartificialneuralnetworks.com/speechrecognition.html

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

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

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

[7] Jeremy Bradbury, "Linear Predictive Coding", 2000.

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

[9] S. Mallat, "A wavelet Tour of Signal Processing", Academic Press, San Diego, 1999.

[10] Elif Derya Ubeyil, "Combined Neural Network Model Employing Wavelet Coefficients for ECG Signals Classification", Digital signal Processing, Vol 19, pp.297-308, 2009.

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

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

[13] 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.

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

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

[16] 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.