A Wavelet Based Recognition System for Malayalam Vowels using Artificial Neural Networks

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

Abstract: *This work explores the use of a discrete wavelet transform, a feature extractor mechanism for speech recognition. Speech recognition is a fascinating application of digital signal processing offering unparalleled opportunities. The real-world applications deploying speech recognition and its implications can be varied across various fields. Speech recognition can automate many tasks that previously required hands-on human interaction. Accurate vowel recognition forms the backbone of most successful speech recognition systems. The vowel set of Malayalam, one of the South Indian languages, is used to create the database. A hybrid approach with discrete wavelet transforms and neural networks are used to form a system with improved performance. Daubechies wavelet is employed in this experiment. Features are extracted by using Discrete Wavelet Transforms (DWT). Training, testing and pattern recognition are performed using Artificial Neural Networks (ANN). The results show excellent overall recognition accuracy above 95%. The high accuracy obtained shows promising potentials of discrete wavelet transforms and neural networks in speech recognition.*

Key words: Speech recognition, Discrete Wavelet Transforms (DWT), Neural networks, Malayalam vowels

Received: 3 March 2010, Revised 29 March 2010, Accepted 2 April 2010

1. Introduction

Speech is the vocalized form of human communication. It is based upon the syntactic combination of lexicals and names that are drawn from very large vocabularies. Human speech is parameterized over many variables such as amplitude, pitch, and phonetic emphasis that vary from speaker to speaker. The ultimate aim of research on Automatic Speech Recognition (ASR) is to make machines understand and convert the spoken sounds and words to text. Speech recognition is one of the intensive areas of research [1]. Speech recognition is widely gaining attention because it allows natural interaction between computer and human beings without the use of keyboard. Many parameters affect the accuracy of the speech recognition system. Speech consists of acoustic pressure waves created by the voluntary movements of anatomical structures in the human speech production system. These waveforms are broadly classified into voiced and unvoiced speech. Voiced sounds, (vowels, for example), are caused by the periodical air puffs generated by larynx and passed through the vocal tract. This paper is focused on vowels.

2. Architecture of the system

Conventional speech recognizers can be easily separated into different modules. Here we have divided the speech recognition process into two stages. The front-end processing is the feature extraction stage wherein short time temporal or spectral parameters of speech signals are extracted. Minimizing the number of parameters in the speech model increases the speed of speech processing. The second one 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. Among these two stages, feature extraction is a key, because better feature is good for improving recognition rate. In this experiment, we have chosen the Discrete Wavelet Transforms as the feature extraction tool. For classification, we have used Artificial Neural Networks. Using a combination of these two methods can increase the recognition accuracy.

The paper is organized as follows. In the following section, we give a brief review of the feature extraction stage followed by the concepts of discrete wavelet transforms. The classification stage and the description of artificial neural networks are explained in the following section. The subsequent section presents the vowel database. Section 4 presents the summary of experiments done. Next section gives the results obtained and the last section contains conclusions.

2.1 The Feature Extraction Stage

Feature extraction method plays a vital role in speech recognition process. This is the initial signal processing front end that converts speech signal into a more compact and convenient mode called feature vectors. The extracted feature vectors contain only the relevant information about the given utterance that is essential for its correct recognition. The information irrelevant for the classification is suppressed. Only these parameters are used for further processing of the input signals.

Researchers have experimented with many different types of features for use in speech recognition. Most of the speech-based studies are based on 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. The computational complexity can be successfully reduced using wavelet transforms, since the size of the feature vector is very less compared to other methods. Now, more and more studies are being done on wavelets.

The Wavelet (WT) can be thought of as an extension of the classic Fourier transform, except that, instead of working on a single scale (time or frequency), it works on multi-scale basis. The wavelet transform has become a popular tool for image and speech processing [2][3][4]. It has been used in speech processing for analysis and pitch detection with much success. The wavelet transform has the best features of narrow band and wide band analysis within one transform without assuming a stationary signal. It has better location characteristic and the resolution in time-frequency domain can be changed. These characteristics of wavelet make it become an effective way for analyzing non-stationary signal. In this experiment, we have used discrete wavelet transforms.

2.1.1 Discrete Wavelet Transform

Discrete Wavelet Transform (DWT) is a relatively recent and computationally efficient technique for extracting information about non-stationary signals like audio. It uses finite durative wavelets instead of periodical sinusoidal waves. The wavelet transform is a multi- resolutional, multi-scale analysis, which has been shown to be very well suited for speech processing because of its similarity to how the human ear processes sound [5]. The Discrete Wavelet Transform (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. The extracted wavelet coefficients provide a compact representation that shows the energy distribution of the signal in time and frequency. Wavelet transform decomposes a signal into a set of basic functions called wavelets. Pattern recognition rate is improved by this method. It was shown that the wavelet analysis was one of the promising methodologies for the pattern recognition noisy signal. The purpose of using the DWT is to benefit from its localization property [6] in the time and frequency domains.

The Discrete Wavelet Transform is defined by the following equation [7, 8].

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

Where $\Psi(t)$ 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 wavelet transform is applied to an input signal at different levels. This is often known as the analysis stage. After passing the signal through the first level filters, the corresponding wavelet coefficients are generated. The coefficients are represented by two sequences. The first sequence corresponds to the low frequencies of the signal, while the second sequence represents the high frequency components. Similarly, the wavelet coefficients of the second level are computed. In speech signals, low frequency components are of greater importance than high frequency signals as the low frequency components characterize a signal more than its high frequency components [9]. So, the low frequencies sequence of the first level is taken as an input to the second stage of the wavelet structure. The discrete time domain signal is subjected to successive low pass filtering and high pass filtering to obtain DWT [10]. This algorithm is called the Mallat algorithm [11]. At each level, the decomposition of the input signal having two kinds of outputs forms the low frequency components, the approximations and high frequency components. At each decomposition level, the half band filters produce signals spanning only half the frequency band. The uncertainty in frequency is reduced by half and thus the frequency resolution is doubled. With this approach, at high frequencies, the time resolution becomes arbitrarily good while the frequency resolution becomes arbitrarily good at low frequencies. The filtering and decimation process is continued until the desired level is reached.

The DWT of the original signal is then obtained by concatenating all the coefficients starting from the last level of decomposition. 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 [12]. DWT is also related to a multi-resolution framework.

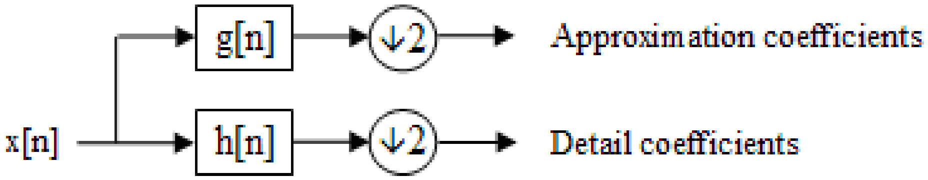


Figure 1. Wavelet decomposition

2.2. The Classification stage

Pattern recognition is becoming increasingly important in the age of automation and information handling and retrieval. The classification stage makes its determination based on all the similarity measures after having been trained using information relating to known patterns and the similarity measured from the pattern. Speech recognition is basically a pattern recognition problem. 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 a neural network as the classifier. Neural networks have been shown to perform pattern recognition, handle incomplete data and variability very well [13].

2.2.1 Artificial Neural Networks

A Neural Network is a massively parallel-distributed processor made up of simple processing units, which has a natural propensity for storing experimental knowledge and making 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 responds with 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.

Let $x_1, x_2, x_3, \ldots\ldots x_n$ be the inputs and $w_1, w_2, w_3 \ldots w_n$ be the corresponding weights. The total input to the next neuron or the output neuron I is calculated by the summation function [15] as

$$I = w_1x_1 + w_2x_2 + \ldots\ldots.. + w_nx_n$$

$$= \sum_{I=1}^{n} w_ix_i \qquad (4)$$

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 [16, 17]. The output layer units often have linear activations, so that output activations equal net function values. The structure of an artificial neuron with summation function and activation function is shown in figure 2.

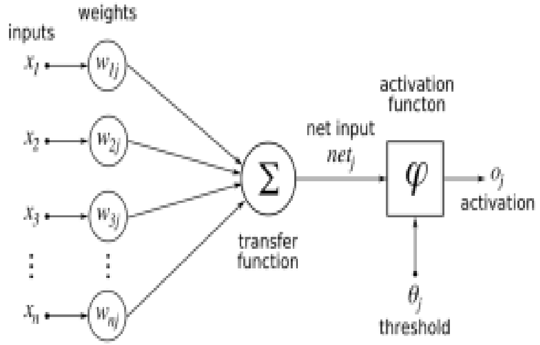


Figure 2. Structure of an artificial neuron with weights and activation function

In this experiment, a Multi Layer Perceptron (MLP) network architecture is used for training and testing. The MLP network consists of an input layer, one or more hidden layers, and an output layer [18]. Each layer consists of multiple neurons. In this work, we use architecture of the MLP network, which is the feed forward network with back propagation training algorithm (FFBP). 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. The structure of an MLP network is given in figure 3.

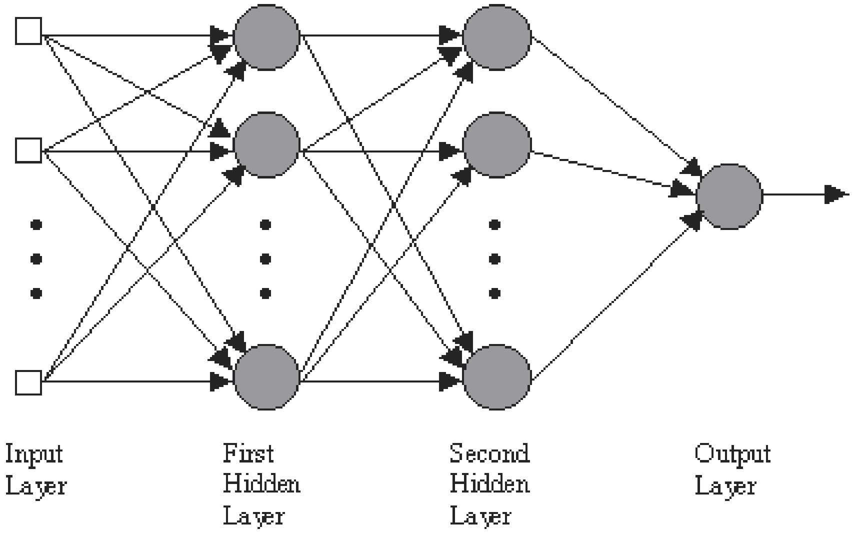
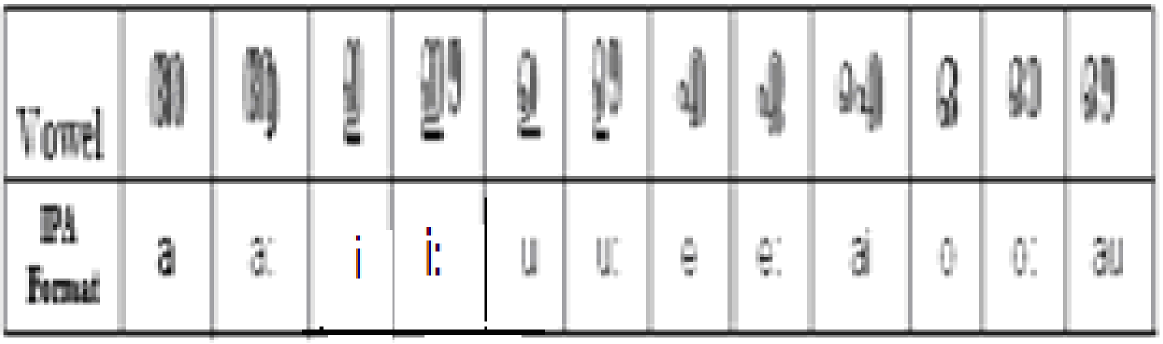


Figure 3. Structure of an MLP network.

3. Vowels database for Malayalam

For this experiment, an isolated vowel database is created for Malayalam language using five speakers. We have used three male speakers and two female speakers for creating the database. The samples stored in the vowels database are recorded by using a high quality studio-recording microphone at a sampling rate of 8 KHz (4 KHz band limited). Twelve vowels are used in this experiment. Our database consists of a total of 60 utterances of the vowels. The vowels are preprocessed, numbered and stored in the appropriate classes in the database. The vowels and their International Phonetic Alphabet (IPA) format are shown in Table 1.



Vowel	അ	ആ	ഇ	ഈ	ഉ	ഊ	എ	ഏ	ഐ	ഒ	ഓ	ഔ
IPA Format	a	a:	i	i:	u	u:	e	e:	ai	o	o:	au

Table 1. Vowels stored in the database and their IPA format

4. Experiment

Daubechies 4 (db4) type of mother wavelet is used for feature extraction purpose. Daubechies wavelets are the most popular wavelets that represent foundations of wavelet signal processing. These are also called Maxflat wavelets as their frequency responses have maximum flatness at frequencies 0 and ð. The speech samples in the database are successively decomposed into approximation and detailed coefficients. Less frequency components from seventh level is used to create the feature vectors for each Malayalam vowel. The feature vectors are taken from the approximation coefficients since it gives more information than detailed coefficients.

The developed feature vectors are given to an artificial neural network for parameter classification. We have divided the database into two. One set for training and the other set for testing purpose. 80% of the database volume is used for training and 20% is used for testing respectively. MLP architecture is used for the classification scenario. The hidden layers and units are chosen in accordance with recognition accuracies.

5. Results

Using the Multi Layer Perceptron network, the classifier can recognize the vowels successfully after training. After testing, the corresponding accuracy of each vowel is obtained. 100% recognition accuracy is obtained for three vowels and most of the other vowels also showed good recognition accuracy. An overall recognition accuracy of 95.75% is obtained from this experiment. The obtained recognition accuracies can be represented using a confusion matrix which shows the recognition accuracy and the confusion in percentage. The confusion matrix obtained for the above experiment is given in table 2. The obtained accuracy and confusion for each vowel is shown in the graph given below.

6. Conclusion

In this paper, wavelet transforms have been found to be an efficient tool for extracting information from speech signals. In this study, we have used an integrated model using discrete wavelet transforms and the neural network classifier model. When compared to other feature extraction techniques used in various researches done earlier, Discrete Wavelet Transform gives substantial improvement in the recognition rate. Also neural networks with back propagation algorithm produces better results than other approaches. This hybrid architecture could effectively extract the features from the speech signal for automatic speech recognition. A better performance of identification with very high recognition accuracy is obtained from this study. The computational complexity and feature vector size is successfully reduced to a great extent by using Discrete Wavelet Transforms. This is due to the fact that wavelet transforms can zoom into time discontinuities and those orthogonal bases localized in time and frequency. Thus a wavelet transform is an elegant tool for the analysis of non-stationary signals like speech. It is also observed that neural network is an efficient tool which can be successfully employed along with wavelet transforms in order to obtain excellent recognition accuracy.

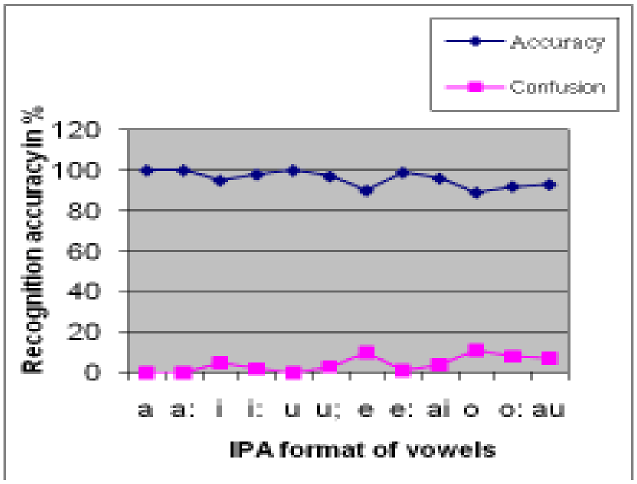


Figure 4. Graph showing recognition accuracy and confusion of vowels

	അ	ആ	ഇ	ഈ	ഉ	ഊ	എ	ഏ	ഐ	ഒ	ഓ	ഔ
അ	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
ആ	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
ഇ	0%	0%	95%	5%	0%	0%	0%	0%	0%	0%	0%	0%
ഈ	0%	0%	2%	98%	0%	0%	0%	0%	0%	0%	0%	0%
ഉ	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%
ഊ	0%	0%	0%	0%	0%	97%	0%	0%	0%	0%	3%	0%
എ	0%	0%	0%	0%	0%	0%	90%	10%	0%	0%	0%	0%
ഏ	0%	0%	0%	0%	0%	0%	1%	99%	0%	0%	0%	0%
ഐ	0%	0%	0%	0%	0%	0%	4%	0%	96%	0%	0%	0%
ഒ	0%	0%	0%	0%	0%	0%	0%	0%	0%	89%	11%	0%
ഓ	0%	0%	0%	0%	0%	8%	0%	0%	0%	0%	92%	0%
ഔ	0%	0%	0%	0%	0%	0%	0%	0%	0%	7%	0%	93%

Table 2. Confusion matrix showing the % recognition of the vowels

References

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

[2] Grossman, A., Kronland-Martinet, R., Morlet, J (1989). Reading and Understanding Continuous Wavelet Transforms Wavelets, Time- Frequency Methods and Phase Space, Springer-Verlag Berlin, p. 2-20.

[3] Martinet, R. K. (1988). The Wavelet Transform for Analysis, Synthesis, and Processing of Speech & Music Sounds, *Computer Music Journal*, 12 (4) 11-20.

[4] Rioul, O., Vetterli, M (1991). Wavelets and Signal Processing, *IEEE Signal Processing*, 1, 14-38.

[5] Vaidyanathan, P. P. (1993). Multirate Systems and Filter Banks, Prentice Hall, Englewood liffs, NJ.

[6] Mallat, S. (1999). A wavelet Tour of Signal Processing, Academic Press, San Diego.

[7]. Martinet, R. K., Morlet, J., Grossman, A (1987). Analysis of Sound Patterns through Wavelet Transform, *International Journal of Pattern Recognition and Artificial Intelligence,* 1 (2) 237-301.

[8] Tzanetakis, G., Essl, G., Cook, P. (2001). Audio Analysis using the Discrete Wavelet Transform, *In:* Proc. WSES Int. Conf. Acoustics and Music: Theory and Applications (AMTA 2001) Skiathos, Greece.

[9] Kadambe, S., Srinivasan, P. (1994). Application of Adaptive Wavelets for Speech, *Optical Engineering* 33 (7) 2204-2211.

[10] Kharate, G.K., Ghatol, A.A., Rege, P.P (2007). Selection of Mother Wavelet for Image Compression on Basis of Image, IEEE - ICSCN 2007, India. p.281-285.

[11] Mallat, S .G. (1989). A Theory for Multiresolution Signal Decomposition: The Wavelet Representation, IEEE *Transactions on Pattern Analysis And Machine Intel*ligence, 11, p.674-693.

[12] Vetterli, M., Herley, C (1992). Wavelets and Filter Banks: Theory and Design, *IEEE Transactions on Signal Processing*, 40, p. 2207- 2232.

[13] Lippmann, R. P. (1989). Review of Neural Networks for Speech Recognition , *Neural Computing* 1, p.1-38.

[14] Smith, L. (1996). An Introduction to Neural Networks, Center for Cognitive and Computational Neuroscience.

[15] Haykin, S. (1999). Neural Network a Comprehensive Foundation, Prentice Hall Upper Saddle River, New Jersey.

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

[17] Lippmann , R. P. (1987). An Introduction to Computing with Neural Nets, *IEEE Trans. On Acoustics, Speech and Signal Processing magazine*, 35 (4) 2-22.

[18] Bishop, C.M. (1999). Neural Networks for Pattern Recognition, London. Oxford.