ISSN: 2321-7782 (Online)

Volume 3, Issue 1, January 2015

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Three Sigma Limits: A Statistical Method for Improving Recognition Accuracy of Speech Signals

Sonia Sunny[1]
Dept. of Computer Science
Prajyoti Niketan College
Thrissur, India

David Peter S[2]
Dept. of Computer Science
CUSAT
Kochi, India

K. Poulose Jacob[3]
Dept. of Computer Science
CUSAT
Kochi, India

Abstract: Speech is the most natural means of interaction between human beings and is used to communicate our thoughts and ideas. In this work, a speech recognition system is developed for recognizing speaker independent spoken digits in Malayalam. The spoken signals from 200 speakers uttering 10 digits each are sampled directly from the microphone. These signals are first pre-processed using wavelet denoising technique. The features from the signals are extracted using Discrete Wavelet Transforms (DWT). Pattern classification is performed using Artificial Neural Networks (ANN). This produced a recognition accuracy of 91%. This paper employs a statistical thresholding technique using Three Sigma Limits to bring the feature vectors within the specified range in order to improve the recognition rate during classification. Application of this technique produced an accuracy of 94.7%. The results obtained clearly shows that this proposed post processing method yields better results for the proper recognition of spoken digits.

Keywords: Speech Recognition; Soft Thresholding; Feature Extraction; Discrete Wavelet Transforms; Classification; Three Sigma Limits; Artificial Neural Networks.

I. Introduction

Speech is a complex signal which is non linear in nature. They are produced as a result of several transformations occurring at different levels. Speech processing and speech recognition are intensive areas of research with wide range of applications [1]. There has been lot of research in the area of speech recognition for the last few decades. 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 samples. Since speech is the primary means of communication between people, research in Automatic Speech Recognition (ASR) and speech synthesis by machine has attracted a great deal of attention over the past five decades [2].

Designing a speech recognition system involves several independent modules. While designing a speech recognition system, several things are to be taken into consideration like creating a good database, defining the speech classes, signal pre-processing methods selected, feature extraction techniques adopted, post processing methods used, speech classifiers used and the performance evaluation methods used [3]. The performance of an ASR depends on these techniques and is measured in terms of recognition accuracy. There has been lot of research in the area of speech recognition for different languages like English, Chinese, Arabic, Turkish, Bengali, Hindi, Tamil etc. But only few works have been reported in Malayalam. So developing an efficient speech recognition system which has more ability to recognize speech is of great importance and is an important and challenging area of research.

There are 5 modules in the speech recognition system developed in this research work. First module is the creation of the words database. Pre-processing techniques are used to tune the speech signals by removing the noise from these during the second stage. In the third module, the speech signals are converted to a set of parameters called feature vectors. Post processing techniques are applied to the feature vector set obtained to reduce the dimensions and to tune the features for appropriate classification. These features are then classified using pattern classification techniques to classify them into proper classes in the fifth module.

Rest of the paper is organized as follows. Section 2 explains the digits database created in Malayalam. In section 3, the pre-processing technique using soft thresholding is described. The feature extraction module using DWT is illustrated in section 4. The new proposed method used for post processing is explained in section 5. Section 6 describes the pattern classification using ANN. Section 7 presents the detailed analysis of the experiments done and the results obtained. Conclusions are given in the last section.

II. Digits Database

In Malayalam, since there are no standard databases available, a spoken digits database is created using 200 speakers of age between 6 and 70 uttering 10 Malayalam digits. We have used 75 male speakers, 75 female speakers and 50 children for creating the database with a total of 2000 utterances of the digits. Male and female speech differ in pitch, frequency, phonetics and many other factors due to the difference in physiological as well as psychological factors. 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 ten Malayalam digits under the same configuration. Digits in Malayalam, Digits in numeric format, their IPA format and the corresponding English translation are shown in Table 1.

TABLE 1: NUMBERS STORED IN THE DATABASE IN MALAYALAM, THEIR DIGIT FORMAT, IPA FORMAT AND ENGLISH TRANSLATION

Digits in Malayalam	Digits	IPA Format	English Translation
പൂജ്യം	0	/pu:d□ʒjam/	Zero
ഒന്ന്	1	/on□n□ə/	One
രണ്ട്	2	/ɾ□aɳɖə/	Two
മൂന്ന്	3	/mu:n□n□ə/	Three
നാല്	4	/ n□a:lə/	Four
അഞ്ച്	5	/and□ʒə/	Five
ആറ്	6	/a:rə/	Six
ഏഴ്	7	/e:ɻə/	Seven
എട്ട്	8	/eʈʈə/	Eight
ഒൻപത്	9	/onpad□ə/	Nine

III. Preprocessing using Wavelet Denoising

Speech signals are often affected by noises from background and this causes degradation in the speech signals. So, these signals are tuned so that the noise present in it is removed before extracting the features. There are a number of techniques available for speech enhancement. Since we are using wavelets for feature extraction, wavelet denoising algorithms are used for reducing the noise in the signal. The two popular thresholding functions used in wavelet denoising method are the hard and the soft thresholding functions [4]. In both the methods, a threshold value is selected. In hard thresholding, if the absolute value of an element is lower than the threshold, then these values are set to zero. Soft thresholding is an extension of hard thresholding. Here, the elements whose absolute values are lower than the threshold are first set to zero and then the nonzero elements are shrinked towards 0. Hard and soft thresholding can be expressed as

$$X_{Hard} = \begin{cases} X & if \quad |X|>\tau \\ 0 & if \quad |X|\le\tau \end{cases} \qquad (1)$$

$$X_{Soft} = \begin{cases} sign(X)\ (|X|-|\tau|) & if \quad |X|>\tau \\ 0 & if \quad |X|\le\tau \end{cases} \qquad (2)$$

Where X represents the wavelet coefficients and ι is the threshold value. In this work, soft thresholding technique is used. There are different standard threshold values available and we have used the universal threshold derived by Donoho and Johnstone [5] for the white Gaussian noise under a mean square error criterion which is defined as

$$\iota = \sigma\sqrt{2\log(N)} \qquad (3)$$

where σ is the standard deviation and N is the length of the signal. Standard deviation σ can be calculated as σ = MAD/0.6745, where MAD is the median of the absolute value of the wavelet coefficients. The outline of the algorithm used for denoising mainly consists of 3 steps.

- Apply wavelet transform to the noisy signal to produce the noisy wavelet coefficients up to 8 levels.
- Detail wavelet coefficients are then shrinked using soft thresholding technique by selecting an appropriate threshold limit.
- The inverse DWT of the threshold wavelet coefficients is computed which produces the denoised signal.

IV. Feature Extraction

Feature Extraction is a major part of the speech recognition system since it plays an important role to separate one speech from other and this has been an important area of research for many years. Selection of the feature extraction technique plays an important role in the recognition accuracy, which is the main criterion for a good speech recognition system. Here, DWT is used for extracting features.

A. Discrete Wavelet Transforms

DWT is a relatively recent and computationally efficient technique for extracting information from non-stationary signals like audio. 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 [6]. DWT uses digital filtering techniques to obtain a time-scale representation of the signals. DWT is defined by

$$W(j,K) = \sum_{j}\sum_{k} X(k)2^{-j/2}\psi(2^{-j}n-k) \qquad (4)$$

Where Ψ (t) is the basic analyzing function called the mother wavelet. In DWT, the original signal passes through a low-pass filter and a high-pass filter and emerges as two signals, called approximation coefficients and detail coefficients [7]. In speech signals, low frequency components h[n] are of greater importance than high frequency components g[n] as the low frequency components characterize a signal more than its high frequency components [8]. The successive high pass and low pass filtering of the signal is given by

$$Y_{low}[k] = \sum_{n} x[n]h[2k-n] \qquad (5)$$

$$Y_{high}[k] = \sum_{n} x[n]g[2k-n] \qquad (6)$$

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 process is continued until the desired level is reached according to Mallat algorithm [9].

V. Post Processing using Three Sigma Limits

Thresholding techniques are used to limit the set of values of the features below a threshold value or to limit the values of the features within a certain range. This range is defined differently depending on the central value we take. But the actual data values may include values outside this predefined range. In this work, thresholding technique based on statistical distribution method namely Three Sigma Limits has been used. Instead of selecting one value for the threshold limit, two limits are used - Upper Specification Limit (USL) and Lower Specification Limit (LSL) [10]. These are used to limit the values of the feature set to a uniform format so that the recognition rate can be improved. This is a statistical calculation that is used to refer to data within three standard deviations from a mean. Usually 3 sigma limits are used to set the upper and lower control limits in statistical quality control charts. Here σ represents standard deviation in statistical analysis and μ denotes the mean which are the fundamental building blocks in Statistics. Standard deviation is a measure of how flat a data distribution is. High value of sigma indicates that the data is more dispersed from the norm. The algorithm for post processing using Three Sigma Limits is given below.

1. For each feature do the following

   1.1 Find the mean μ of that feature obtained after feature extraction using the equation

$$\mu = \frac{\Sigma Xi}{n}$$

   1.2 Calculate the standard deviation σ of the feature using the equation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

2. For each observation X in a feature, do the following.

$$\text{If } \mu - 3\sigma < X < \mu + 3\sigma \text{ then } X = X$$

Else if

$$X > \mu + 3\sigma \text{ or } X < \mu - 3\sigma \text{ then } X = \mu$$

VI. Speech Classification

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 [11]. ANNs are well suited for speech recognition due to their fault tolerance and non-linear property.

A. Neural Networks Classifier

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 [12]. Neural Networks have become a very important method for pattern recognition because of their ability to deal with uncertain, fuzzy, or insufficient data. 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 [13]. 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 [14]. After extensive training, the network will eventually establish the input-output relationships through the adjusted weights on the network [15].

VII. 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 [16]. The speech samples in the database are successively decomposed into approximation and detailed coefficients. In this work, better results are obtained at level 8 during decomposition. The original signal and the 8th level decomposition coefficients of spoken digit Poojyam (Zero) using DWT is given in figure 1.

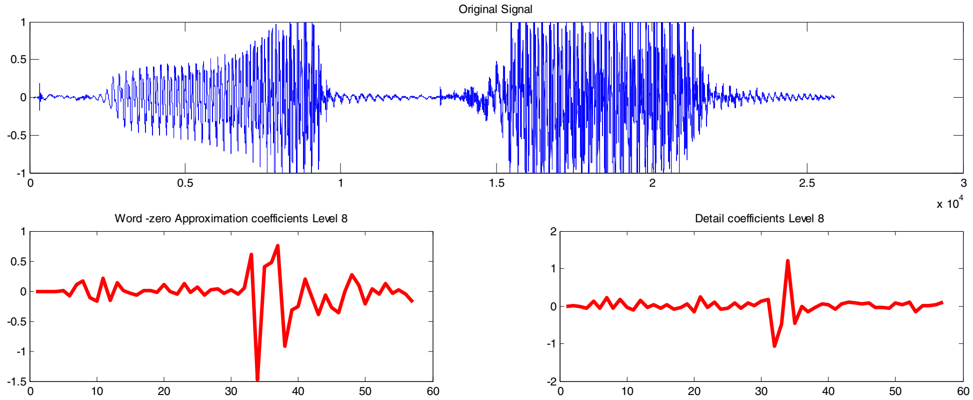

Fig. 1 Decomposition of digit poojyam at 8th level using DWT

The feature vectors thus generated are given to the MLP architecture, which uses one input layer, one hidden layer and one output layer.

A. Results obtained without post processing using Three Sigma Limits

The feature vectors obtained after feature extraction are given directly to an ANN for classification. This produced an accuracy of 91%. The confusion matrix obtained using this classification is given in fig 2 below.

Class	0	1	2	3	4	5	6	7	8	9
0	**187**	2	3	1	0	4	1	2	0	0
1	1	**190**	0	2	3	0	2	0	1	1
2	4	3	**165**	6	2	5	7	4	2	2
3	3	1	2	**174**	6	1	2	3	2	6
4	0	1	0	2	**195**	0	1	0	0	1
5	3	0	4	2	5	**181**	2	1	2	0
6	4	0	1	0	2	3	**186**	2	1	1
7	7	2	0	1	3	1	1	**180**	3	2
8	0	2	1	0	3	0	0	1	**192**	1
9	2	4	8	5	1	4	2	3	1	**170**

Fig. 2 Confusion matrix for digits database without using Three Sigma Limits

B. *Results obtained after post processing using Three Sigma Limits*

Here, the feature vectors obtained are brought within the three sigma limits. The feature values that are outside the Three Sigma Limits are substituted by the Mean. In this work, the mean is calculated as the sum of all data values of a feature divided by the number of observations in that particular feature. After post processing, the feature vectors are classified using ANN and an overall recognition accuracy of 94.7% is obtained. The confusion matrix obtained using this method is given in fig 3.

Class	0	1	2	3	4	5	6	7	8	9
0	**190**	2	3	1	0	2	0	2	0	0
1	1	**192**	0	2	1	0	2	0	1	1
2	3	1	**177**	4	2	3	2	4	2	2
3	3	1	2	**185**	4	0	2	0	2	1
4	0	1	0	1	**195**	0	1	1	0	1
5	3	0	2	2	1	**190**	0	1	1	0
6	0	0	1	0	0	1	**196**	0	1	1
7	1	2	0	1	3	1	1	**188**	1	2
8	0	2	1	0	3	0	0	1	**193**	0
9	2	3	1	0	1	0	2	3	0	**188**

Fig. 3 Confusion matrix for digits database using Three Sigma Limits

C. *Comparison of Results*

Table 2 given below shows the comparison of results obtained using both the techniques. The results clearly shows that post processing technique using Three Sigma Limits outperform the results obtained without using any post processing method.

TABLE 2: COMPARISON OF RESULTS

No. of Speakers	Total Samples	Without using three sigma limits		Using three sigma limits	
		Correctly classified	Recognition Accuracy %	Correctly classified	Recognition Accuracy %
200	2000	1820	91	1894	94.7

VIII. Conclusion

In this work, a speech recognition system is designed for spoken digits in Malayalam. This paper shows the importance of a fine and tuned set of features for the proper recognition of speech samples. For this purpose, a statistical thresholding technique based on Three Sigma Limits is applied to the speech samples for bringing the feature vectors within a range. A comparative study of the results obtained with and without post processing of the feature vectors is performed here. These methods are combined with neural networks for classification purpose. The performance of both these are tested and evaluated. The accuracy rate obtained by using post processing technique is found to be more than that of the other. Moreover, a wavelet transform is found to be an elegant tool for the analysis of non-stationary signals like speech. The experiment results show the necessity of proper feature vectors for the correct classification of the speech signals. 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. Joseph P Campbell, JR, "Speaker Recognition: A Tutorial", Proceedings of the IEEE, Vol. 85, No. 9, 1997.
2. L. Rabiner, B. H. Juang, "Fundamentals of Speech Recognition", Prentice-Hall, Englewood Cliffs, NJ, 1993.
3. J. H. M. Daniel Jurafsky, "Speech and Language Processing, An introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition", Prentice Hall, Upper Saddle River, New Jersey 07458, 2000.
4. Yasser Ghanbari, Mohammad Reza Karami, "A new Approach for Speech Enhancement based on the Adaptive Thresholding of the Wavelet Packets", Speech Communication, Vol. 48 (8), pp. 927–940, 2006.
5. D.L. Donoho., "De-noising by Soft Thresholding", IEEE transactions on Information Theory, vol. 41, No. 3, pp. 613-627, 1995.
6. Elif Derya Ubeyil.,"Combined Neural Network model Employing Wavelet Coefficients for ECG Signals Classification", Digital Signal Processing, Vol 19, pp. 297-308, 2009.

7. S. Chan Woo, C.Peng Lin, R. Osman., "Development of a Speaker RecognitionSystem using Wavelets and Artificial Neural Networks", Proc. of Int. Symposium on Intelligent Multimedia, Video and Speech processing, pp. 413-416, 2001.

8. S. Kadambe, P. Srinivasan., "Application of Adaptive Wavelets for Speech, Optical Engineering" , Vol 33(7), pp. 2204-2211, 1994.

9. S .G. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation", IEEE Transactions on Pattern Analysis And Machine Intelligence, Vol.11, 674-693, 1989.

10. J. Wild, and G. A. F. Seber, "Chance Encounters: A First Course in Data Analysis and Inference", 1st ed. USA: Wiley, 1999.

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

12. K. Economou and D. Lymberopoulos, "A new perspective in learning pattern generation for teaching neural networks", Neural Networks, (1999), Volume 12, Issue 4-5, pp. 767-775.

13. Eiji Mizutani and James W. Demmel, "On structure-exploiting trustregion regularized nonlinear least squares algorithms for neural-network learning", Neural Networks, (2003), Volume 16, Issue 5-6, pp. 745-753.

14. Wouter Gevaert, Georgi Tsenov, and Valeri Mladenov, "Neural Networks used for Speech Recognition," Journal of Automatic Control, vol. 20, pp. 1-7, 2010.

15. Ajith Abraham, "Artificial Neural Networks," in Handbook of Measuring System Design, vol. 1, Wiley, 2005, ch. 129, pp. 901-908.

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.