Speaker Verification System using Wavelet Transform and Neural Network for short utterances

Krishna Sarma[1], Fidalizia Pyrtuh[2], Debarun Chakraborty[3]

[1]Department of Electronics and Communication Engineering, North-Eastern Hill University, Shillong, India [2]Department of Electronics and Communication Engineering, North-Eastern Hill University, Shillong, India [3]Department of Electronics and Communication Engineering, North-Eastern Hill University, Shillong, India

[1]krishnasarma95@gmaill.com [2]fidapyrtuh@gmail.com [3]dkdebarun666@gmail.com

***Abstract*— In this paper, wavelet transform technique and neural network is used for development of Speaker Verification System for short utterances. The sampled data undergo 4-level decomposition in wavelet decomposition technique. DCT (Discrete Cosine Transform) is performed on the dataset, to improve the features extraction process. This study includes Hilbert Transform, which shows the importance of magnitude and phase for speaker classification and their performance was shown. Hilbert Transform is explored, to analyze performance of phase for the data. The features are then, fed to feed-forward back propagation neural network for further classification. The proposed technique is evaluated on fixed phrase of the RedDots dataset and self-recorded numerical dataset. The proposed method performs effectively up to 95% recognition rate.**

Keywords*— *Speaker verification system, short utterances, Discrete Wavelet Transform (DWT), Discrete Cosine Transform (DCT), Feedforward Neural Network (FFNN), Hillbert Transform.

1. Introduction

Speech processing is an interdisciplinary subject which describes the information transfer from one person to another person via speech signal. Speaker recognition is a technique of recognizing a speaker from the given speech signal. The analysis method includes speech production, process the acoustic waveform and extract interesting acoustic features. The process of Speaker recognition generally involves classification or recognition based upon speech features. The features are generally obtained via frequency domain [1], time domain [2] and time-frequency representation [3]. It has been proved that Wavelet Analysis is a powerful technique for various problems of signal processing [4],[5],[6]. In studies, it is seen that the features of speech signal can be strengthened by the coefficients of wavelet transform and gives smaller set of features in final classifier which is more robust [7], [8], [9].

People are working on Artificial Neural Networks (ANN) for last many years in order to achieve human-like performance in automatic recognition of speech signal[10]. The architecture of ANNs is similar to the structure of biological neural networks [11]. In speech recognition problems Artificial Neural Networks are widely used because of its non linear property and error tolerance [12]. In this work, we are using wavelet transform form features extraction because it has the ability to deal with non-stationary signals and analyze the signals in different scales and achieve variable time frequency localization.
In this work Discrete Wavelet Transform is used to extract features from original speech signals and a neural network is developed, to recognize the speech signals.

The second Section of the paper contains the basics of the wavelet transform technique; third section reviews Neural Network. In fourth section, the proposed method is described. Fifth section contains experiments and results. The conclusion is presented in the last section.

2. Wavelet Transform

Wavelet transforms have been studied comprehensively in the recent times and widely utilized in various areas of science and engineering [13], [14], [15]. The fundamental idea behind wavelets is to analyze a given signal according to a scale [16]. The wavelet successively decomposes the given signal into a set of smaller signals, at multiple levels and analyses each piece of the signal at different frequencies with different resolutions. Morlet first considered wavelets as a family of functions constructed from translations and dilations of a single function called the 'mother wavelet' [13], [17].They are defined by

$$\psi_{c,d}(t) = \frac{1}{\sqrt{|c|}}\psi\left(\frac{t-b}{c}\right), c, d \in R, c \neq 0 \qquad (1)$$

The parameter *c* is called the scaling parameter or scale, it calculates the degree of compression of the wavelet and the parameter *d* is called the translation parameter which evaluates wavelets time location.

2A. Discrete Wavelet Transform

DWT of a signal is obtained by passing it through a series of filters. First the signals are passed through a high pass filter and down sampled it by a factor of 2; again the signal is passed through a low pass filter and down sampled by a factor of 2. The output of the HPF $H_1(z)$ gives the detail coefficients and output of LPF $H_0(z)$ gives the approximation coefficients. As maximum of speech energy gathered on low frequency band, therefore the signals coming out from the low pass filter again need to be split into sub bands. For this purpose two filters $H_1(z)$ and $H_0(z)$ as shown in Fig. 1 are used.

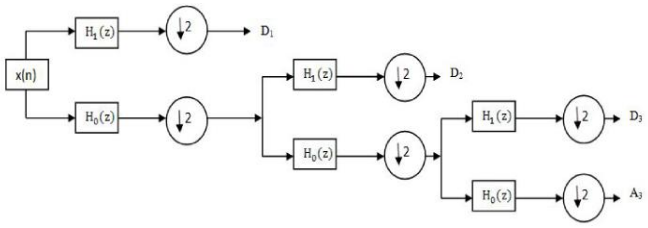

Fig. 1. *Discrete wavelet transform tree for three stages*

The wavelet co-efficients are given by the following expression [1]:

$$\alpha_1(k) = \sum_n x(n) g(2k - n) \tag{2}$$

$$\beta_1(k) = \sum_n x(n) h(2k - n) \tag{3}$$

Where, $\alpha_1(k)$ and $\beta_1(k)$ represents the detail and approximation coefficients respectively at level 1 and translation k. g is low pass filter and h is high pass filter.
Again after each level of decomposition the sampling frequency becomes half of the earlier. Sampling frequency of a signal after n^{th} level of decomposition is given by

$$F_n = f_s / 2^n \tag{4}$$

Where, f_s is the initial sampling frequency of the signal.

2B.Wavelet based cepstral

Cepstral analysis is a very important concept in speech processing for extracting features. Cepstrum are generally used for pitch detection and formant estimation. To increase the capability of the cepstral features so many researches have been conducted. One way provided by different researchers to use Wavelet Based Cepstral Analysis.
For a given speech frame Wavelet Cepstral Coefficients is calculated by the following steps:
Discrete wavelet Transform of the speech frame
Log energy spectrum calculation
Discrete Cosine Transform of the log energy spectrum.

To calculate the WCC for a given speech frame x[n] the signal is first decomposed using the DWT to obtain the wavelet co-efficients for each level as given in (1).

$$\alpha_s[k,2^j] = \frac{1}{\sqrt{2^j}} \sum_{n=0}^{N-1} x(n) \psi\left(\frac{n-k}{2^j}\right) \tag{5}$$

Here, $\alpha_s[k,2^j]$ is the wavelet coefficient for level j; 2^j is the scaling and k is translation parameter. After that the detail coefficient of each level is passed through the log energy calculation.

$$\alpha_p[k,2^j] = \frac{1}{\sqrt{2^j}} \log \sum_{n=0}^{N-1} \left| x(n) \psi\left(\frac{n-k}{2^j}\right) \right|^2 \tag{6}$$

Where, $\alpha_p[k,2^j]$ is the log energy of wavelet coefficients at a level j. At last the log energy of the wavelet coefficients is de-correlated using the DCT

$$\alpha[i,2^j] = \sqrt{\frac{2}{N}} \sum_{n=0}^{N} \alpha_p[k,2^j] \cos\left(\frac{\pi i}{N}(k - 0.5)\right) \tag{7}$$

Where, $\alpha[i,2^j]$ is the WCC of the j^{th} level of the DWT decomposition. To form all the WCC signal, each level of the WCC computed are concatenated into a single vector.

3. Neural Network

Artificial Neural Network is a tool used for data modeling in statistics [12]. It is not linear and is modeled from the structure of human brain. A technical neural network consists of simple processor unit, the neurons and directed weighted connections between those neurons [18].

The first layer is the input layer and it has input units which distribute the inputs to the subsequent layers. In the second layer that is hidden layer, each of the units sums the inputs and adds a threshold to it to produce a unit output. An artificial neurons is a function f_i of the input $x=(x_1,x_2,x_3.....x_n)$ weighted by a vector of connection weights $w_i=(w_{i,1},w_{i,2},w_{i,3}.....w_{i,n})$ combined with neuron bias b_i and associated to an activation function say ϕ. Output

$$y_i = f_i(x) = \varphi((w_i, x) + b_i) \tag{8}$$

3A.Pattern Recognition

An important application of neural network is pattern recognition. Pattern recognition can be implemented by using a feed-forward neural network that has been trained accordingly. During training, the network is trained to associate outputs with inputs patterns. When network is used, it identifies the input pattern and tries to output the associate output pattern.

4. Proposed Method

A speaker verification system on the basis of wavelet transform and neural network is developed in this paper. The block diagram is as shown in Fig. 3. The work contains three parts: (i) Data acquisition and data processing, (ii) feature extraction and (iii) classification as shown as Fig. 2.

MATLAB R2018b is used for training and testing purpose of the developed system.

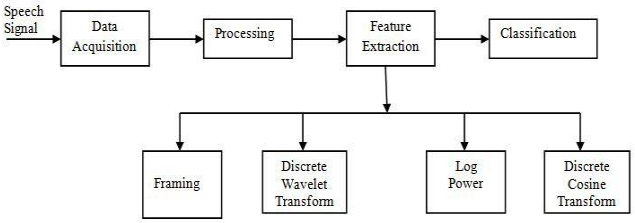

Fig. 2. *Block diagram of a Speaker Recognition System*

4A. Data Acquisition and Processing

The dataset consists of numbers from one to five spoken by five persons, recorded on a Laptop in Laboratory environment and RedDots dataset consists of nine different English sentences uttered by nine different English speakers (Female) at different interval of times. In RedDots data, the utterances are of short duration and variable phonetic contents [20]. The audio format is wave with a sampling rate of 8 kHz. Each individual uttered each number five times and each sentence nine times. The speech signal is frame in 20ms window length.

4B.Feature Extraction

For feature extraction Discrete wavelet transform (DWT) is performed on the dataset. Second order Daubechies wavelet (db2) at fourth level is used to obtain a good feature representation. The coefficients obtained after Wavelet Transform is not robust against the additive noise, so it should be normalized. Therefore we take log of the energy spectrum of the coefficients, which weakens the influence of low energy components like noise. Discrete Cosine Transform of the log energies is processed to get wavelet based cepstral coefficients. Discrete cosine transforms (DCT) based speech compression is used to reduce the size of the speech information [19]. It is used to speed up the system by removing the redundancy from the audio information. Hilbert transformation is then performed with the final data to represent it in complex form, which highlight the phase of the speech signal.

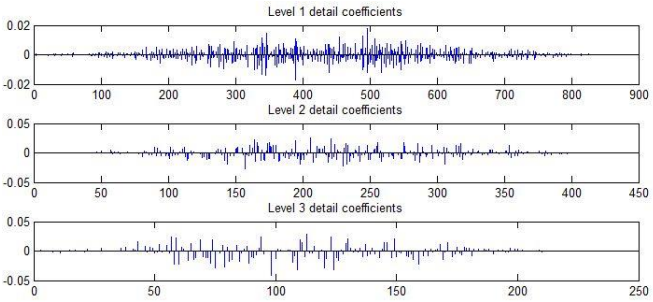

Fig. 3. *DWT coefficients*

4C. Classification

A feed-forward neural network is used here for classification. The architecture of the network is 2-layer architecture (input, hidden layer, output layer and output). Pattern recognition tool of MATLAB is used for training and testing of the neural network; it will recognize the pattern and classify it accordingly. In pattern recognition process, we want a neural network to classify inputs into a set of target categories.
A matrix of the cepstral coefficients are applied here as input and a target is designed accordingly. The neural network pattern recognition tool will select data, create and train a network and evaluate its performance using mean square error(mse) and confusion matrices.
MSE is given by

$$MSE = \frac{1}{a}\sum_{1}^{n}\left(y_b - y_c\right)^2 \qquad (9)$$

Where, y_b represents value of target, y_c represents actual valu of output and a is the number of target data.

The parameters used for Neural Network is shown in Table I.

Table I Parameters used for Neural Network

Architecture	Example
Network type	feed-forward backpropagation
Activation function	Sigmoid
Training algorithm	Scaled conjugate gradient backpropagation(trainscg)
Number of epochs	10000

Training algorithm trainscg is a network training function that updates weight and bias values according to the scaled conjugate gradient method. It takes a network with input data and target data and returns the network after training it and a training record.

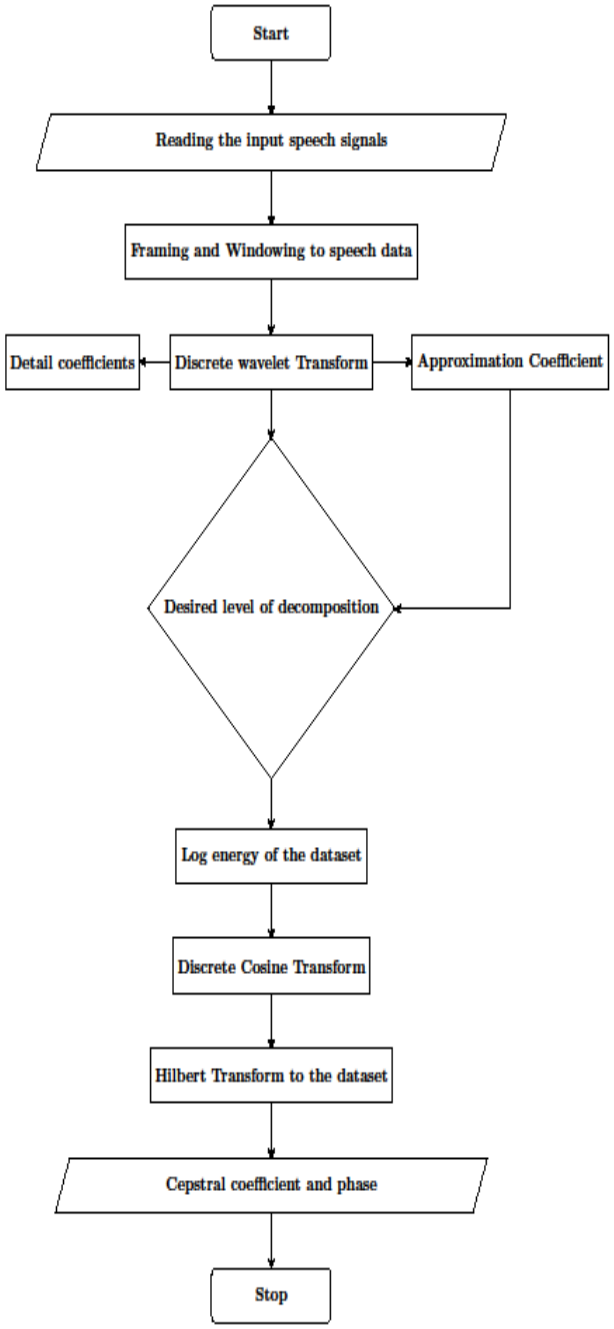

Fig. 4. *Flowchart of the algorithm*

5. Experiment and Results

The experiment is performed using two types of dataset. The first dataset contains English utterances for 25 numerical words (numbers from one to five) uttered by five different speakers. Each individual uttered each number 5 times. The second dataset contains RedDots dataset consists of nine different English sentences uttered by nine different English speakers (Female). Each individual uttered each sentence 9 times.

The testing of the classifier is performed at different stages of the algorithm. A plot of ROC is shown in different stages.When we tested the classifier, an overall recognition accuracy of 96% and 95.1% is achieved by means of FFNN for both the dataset respectively. Table III shows the recognition accuracy for both the self recorded numerical dataset and RedDots dataset with and without Hilbert transformation. The dataset with Hilbert contains both the amplitude and phase value. Recognition accuracy using both amplitude and phase is obtained.

ROC curve without taking DCT

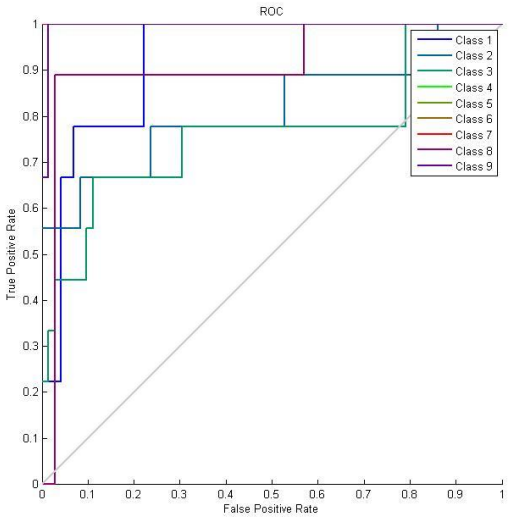

Fig. 5. *Receiver Operating Characteristics curve*

ROC curve after taking DCT

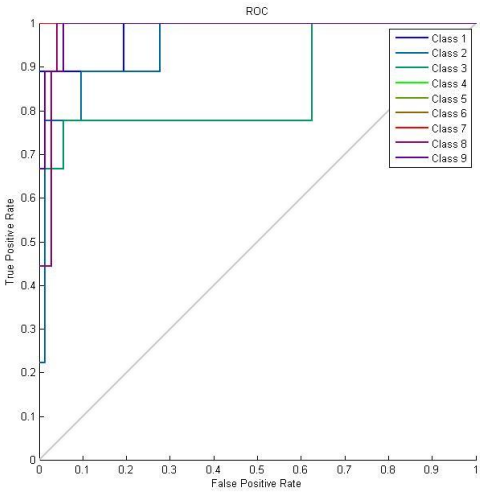

Fig. 6. *Receiver Operating Characteristics curve*

ROC curve after taking hilbert transform

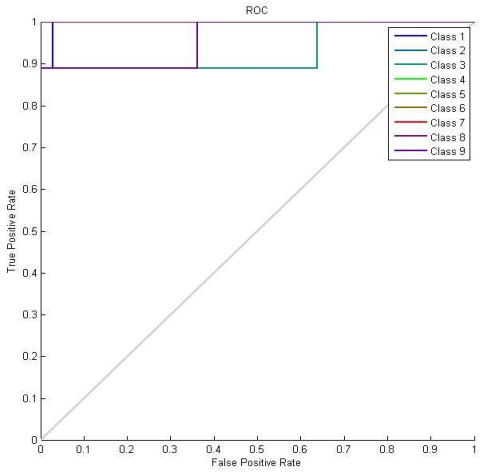

Fig. 7. *Receiver Operating Characteristics curve*

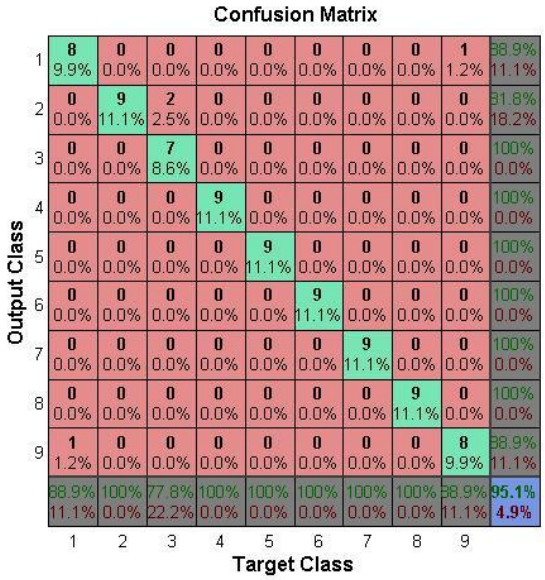

Fig. 8. *A plot of confusion for RedDots data of female speakers*

Table II. Recongition rate using FFNN for neumaric datasets

Data	Correct Classification	Incorrect Classification	Recognition Accuracy	Average Recognition Acccuracy
One	4	1	80%	96%
Two	5	0	100%	
Three	5	0	100%	
Four	5	0	100%	
Five	5	0	100%	

It indicates the reliability and effectiveness of the proposed method for extracting features from speech signals. The performance of phase shows that phase can have a considerable amount of information related to speech and speaker information. To indicate the performance of the system three Receiver Operating Characteristic (ROC) curves are added in Fig. 5, 6 and 7. Recognition rate using FFNN for numeric data is shown in Table II. This table is obtained from confusion matrix. To show the recognition accuracy a plot of confusion is added in Fig. 8. Table III indicates the reliability and effectiveness of the proposed method for extracting features from speech signals.

Table III. Recongition accuracy for the datasets

Datasets	Recognition Accuracy			
	With DCT	Without DCT	With Hillbert	
			Amplitude	phase
Self Recorded	70	90	96	88
RedDots Female	68	89	95.1	84

Conclusion

In this study, a speaker verification system for short utterances based on Discrete Wavelet Transform and Artificial Neural Network techniques is proposed. The experiment is performed using two types of data sets consists of numbers from 1 to5 and nine different English sentences from RedDots datasets. The experimental results show that the introduced method can make an effective analysis with an average recognition rate of about 90% for both amplitude and phase of speech signals. The scope of this paper lies in the further extraction of the information from phase to give better enhancement in the recognition

rate. From the results we can conclude that this method can be used to design an accurate and robust classifier.

References:

[1] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357-366, August 1980.

[2] L. R. Rabiner and R. W. Schafer, "Digital Processing of Speech Signals," Prentice- Hall Englewood Cliffs,1978.

[3] E. R. Rady, A. H. Yahia, El-Dahshan, El-S.A. and H. El-Borey, "Speech Recognition System Based on Wavelet Transform and Artificial Neural Network," Egyptian Computer Science Journal, ECS, Vol. 37 No. 3, ISSN-1110-2586 , 2013.

[4] N. Almaadeed, A. Aggoun and A. Amira, "Speaker Identification Using Multimodal Neuralnetworks and Wavelet Analysis," IET-BMT, 2014.

[5] M. Siafarikas, T. Ganchev and Fakotakis, "Wavelet Packets Based Speaker Verification," Proceedings of the ISCA speaker and language recognition workshop Odyssey, Toledo, Spain, May 31–June 3, pp. 257–264, 2004.

[6] T. B. Adam, M.S. Salam and T.S. Gunawan, "Wavelet Based Cepstral Coefficients for Neural Network Speech Recognition," IEEE International Conference on Signal and Image Processing Applications(ICSIPA), October 2013.

[7] J.B. Buckheit and D.L. Donoho, Wave Lab and Reproducible Research, Dept. of Statistics, Stanford University, Tech. Rep. 474, 1995.

[8] E. Wesfred and V. Wickerhauser, "Adapted local trigonometric transforms and speech processing," IEEE trans. on Signal Proc. 41 N.12, pp-3596-3600, 1993.

[9] E. Visser, M. Otsuka and Lee, "A Spatio-Temporal Speech Enhancement Scheme for Robust Speech Recognition in Noisy Environment," Speech Communication, pp. 393-407, 2003.

[10] Y.A. Alotaibi, Investigation of Spoken Arabic Digits in Speech Recognition Setting. Informatics and Computer Sciences 173 (1–3)105–139, 2005.

[11] J. Lampinen and E. Oja, "Fast Self-organization by the Probing Algorithm," Proceedings of the International Joint Conference on Neural Networks (IJCNN), volume II, pp. 503-507, Piscataway, NJ. IEEE Service Center, 1989.

[12] S. Haykin, Neural Networks: A Comprehensive Foundation, Edition 2, Prentice Hall, 1999.

[13] S. Mallat, A Wavelet Tour of Signal Processing, Elsevier, UK, 1999.

[14] S. Lung and C. Chen, "Further Reduced Form of Karhunen–Loeve transform for Text Independent Speaker Recognition ," Electronics Letters, Volume 34, ISSN 0013-5194, pp. 1380–1382 , July 1998.

[15] M. Vetterli and J. Kovacevic, Wavelets and Subband Coding, Prentice-Hall, NewJersey, 1995.

[16] A. Shukla, R. Tiwari, H.K. Meena and R. Kala, "Speaker Identification using Wavelet Analysis and Modular Neural Networks," J. Acoust. Soc. India (JASI), Volume 36, (1), pp. 14–19, 2009.

[17] M. Sifuzzaman, M.R. Islam and M.Z. Ali, "Application of Wavelet Transform and its Advantages Compared to Fourier Transform," Journal of Physical Sciences, Vol. 13, pp. 121-134, ISSN: 0972-8791, 2009.

[18] V.R. Vimal Krishnan and P. Babu Anto, "Feature Parameter Extraction from Wavelet Sub band Analysis for the Recognition of Isolated Malayalam Spoken Words," International Journal of Computer and Network Security(IJCNS), 1(1), October 2009.

[19] H. Amhia and R. Kumar, "A New Approach of Speech Compression by Using DWT & DCT," IJAREEIE3(7), pp. 10762-10765, 2014.

[20] K.A. Lee, A. Larcher, G. Wang, P. Kenny, N. Li. H. Brummer, T. Stafylakis, J. Alam, A. Swart and J. Perrez, "The RedDots Data Collection for Speaker Recognition," INTERSPEECH, 2015.