International Journal of Advanced Science and Technology
Vol. 29, No. 7s, (2020), pp. 5871-5877

Speaker Independent Robust Speech Recognition Using Wavelet Transform and Neural Network

Irshad[1*],ZainulAbdin Jaffery[2],Khwaja M Rafi[3], Shaheen Khan[4], Nadeem Ahmad[5]

[1,2,5]*Department of Electrical Engineering, JamiaMilliaIslamia, New Delhi, India.*
[3]*Department of EEE, Mewat Engineering College (WAQF), Nuh, India.*
[4]*Department of ECE, Mewat Engineering College (WAQF), Nuh, India.*

Abstract:

For smooth interaction between humans and machines, speech recognition can play a big role in future.In this paper, a Speaker Independent Speech Recognition system is designed with the help of Artificial Neural Network.Wavelet transform is used to extract the features from the speech samples. These features were used to train a 10 layer, 7 input feed-forward neural network to classify the speech samples. Samples were recorded from 10 different persons using a mobile phone for different speech words. Advantages of Wavelet transform and Neural Network over conventional HMMs are combined to design a Speaker Independent Speech Recognition system. An accuracy of more than 99% was achieved in word recognition using multilayer feedforward Neural Network.

Keywords: *Speech Recognition, Wavelet Transform, Feature Extraction, Neural Network.*

1. Introduction:

In modern age automation systems, speech recognition plays an important role. It is one of the fastest developing fields in the framework of speech science and engineering such as text-to-speech (TTS) and supporting Interactive Voice Response (IVR) systems [1-3].The Speech Recognition systems are mainly classified into four major categories named as Speaker Dependent, Speaker Independent, Isolated word recognition system and phrase recognition system. In speaker dependent systems, a word recognition rate of 90% has been achieved by using speech features. However, Speaker Independent Speech Recognition (SISR) Systems are much more complex due to difficulty in training the system to suit all the speakers.

With the remarkable growth in the use of computers to process speech data, researchers in the area of speech analysis have long sought to extract features of speech waveform for speech recognition. Speech Recognition System is one which essentially converts the speech signal into words in the form of text. The recognized words may be the final output, or the input to natural language processing. An utterance that is given as audio waveform to the proposed Speaker Independent Speech Recognition (SISR) System is digitized and processed to extract representational vectors $X = x_1, x_2...x_t$ (speech features) [4-9]. Hence the problem of speech recognition can be formally stated as to recognize the utterance recorded by different speakers. Recognition will be done on the basis of speech signal features which are calculated using wavelet transform.

In this paper, a speaker independent speech recognition system has been proposed to combine the advantages of Artificial Neural Networks and Wavelet Transform. To design speaker independent speech recognition system, initially speech utterances were recorded from various speakers with the help of a microphone. The microphone was kept at a distance of 5-7 cm from the mouth of speaker. A database of 300 samples was prepared. Speech recognition system was implemented in three stages. In first stage,

ISSN: 2005-4238 IJAST

Pre-emphasis and Silence removal operation was successfully preformed. In second stage, speech features were extracted with the help of Wavelet Transform. A vector of 7 elements was prepared for every input sample. In the third stage, A Neural network was designed for the purpose of pattern recognition. Speech features extracted from speech samples were given as the input to train the neural network. Scaled conjugate gradient algorithm was used to train the neural network. Feedforward neural network was configured with one hidden layer. The adaption of wavelet transform, filter banks and neural networks were studied in the further sections of the paper.

2. Proposed Algorithm

Implementation of proposed algorithm is done into three steps named asPre-processing, Feature Extraction, and Classification as shown in fig. 1. All the blocks have been simulated with the help of MATLAB software. R2017a version of MATLAB has been used to execute all the programs. Minimum requirements for R2017a version is 64-bit Windows based operating system with 1GB internal memory. In practical, speech signal vary in different factors which create challenge to accurate speech signal recognition. Variation may be in the form of age groups, environmental conditions, media of communication, gender etc. A robust speech recognition system should accurately identify the words and the speaker because SRs have been installed for security purposes at various facilities in daily life.

Training and testing a speech recognition system needs a collection of utterances appropriate for the task on hand.Audio samples of English numerical from 'Zero' to 'Nine' were recorded from 3 speakers. 10 utterances of each numerical was recorded by each speaker. Hence 300 total utterances were recorded. Live speech was recorded using high quality microphone at a distance of 5-7 cm from the mouth while recording the utterance. The data recorded from the microphone may contain background noise. This data can be used for studying and comparing various speech enhancement techniques for speech recognition. The speech data was sampled at 16 kHz.

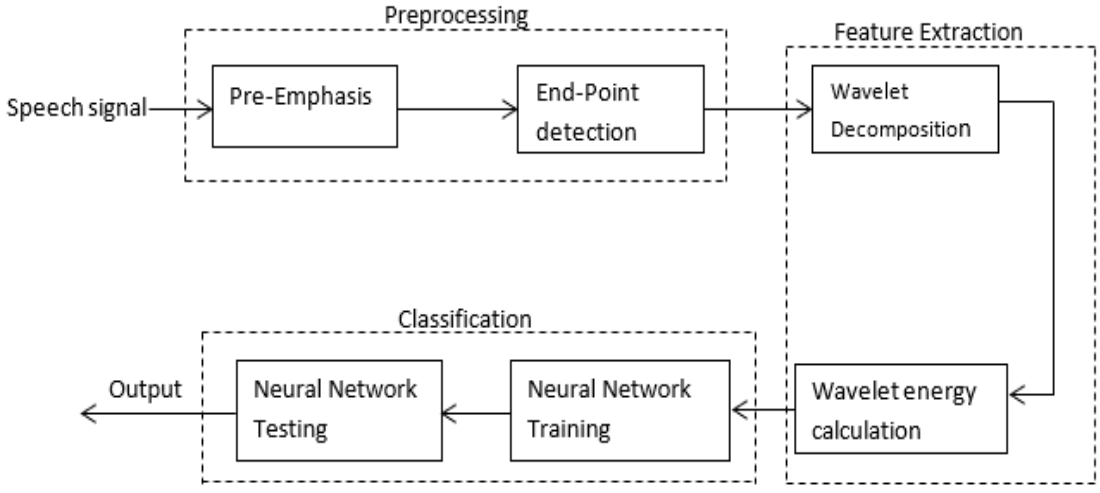


Fig. 1:Block diagram of proposed scheme

To remove the unwanted signal from the speech signal, pre-processing operations are performed on the voice signal. It is important to remove noise signals and silence as it may result in false detection of speech signal.In the present work, two tasks were performed to in pre-processing part. First, to remove the

ISSN: 2005-4238 IJAST

noise signal and second, end point detection to remove the silence part of the signal. End point detection is done by detecting the energy level of the signal as shown in fig. 2.

Efficient feature extraction techniques are required as different speaker may have some common attribute in their voice samples. A good feature extraction tool can separate these common features. In the presented work, two operations are used for pre-processing, Pre-emphasis and Endpoint detection. Pre-emphasis is used to boost the energy level at higher frequencies [1] while Endpoint detection is used to remove the unvoiced part from the signal [6].In speech signal, original signal has very less energy at higher frequencies. So, to boost the energy at higher frequencies, we do pre-emphasis operation so that higher frequency components can be detected well. Pre-emphasis is done as per the following formula [7]:

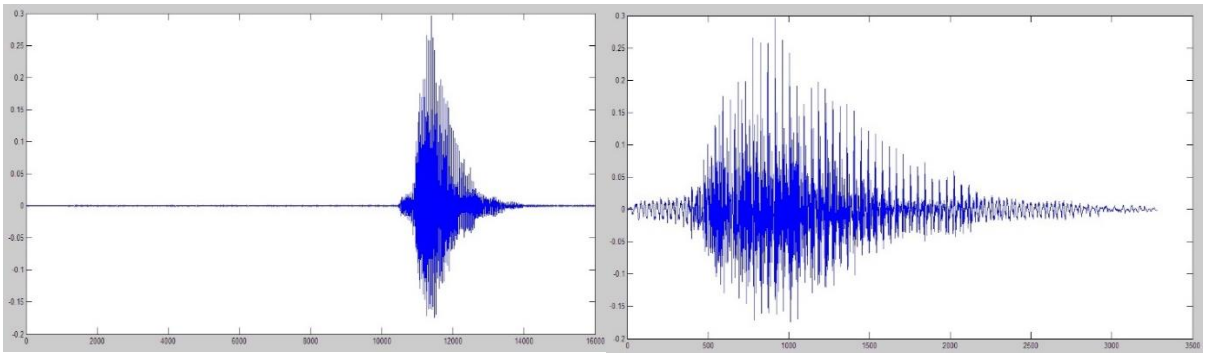
$$y[n] = x[n] - \alpha * x[n-1]; \quad \text{Value of } \alpha \text{ is chosen between 0.1 to 0.9}$$

The speech fames may overlap with the adjacent frames to produce a smooth energy line which needs to be separated efficiently. Fig. 2 shows the energy plot of signal *"Nine"* before emphasis. In the presented method of end-point detection, Standard deviation and mean value of the samples is calculated and then based upon the formula given below, Voiced and unvoiced samples are separated. Suppose the magnitude of i^{th}sample is $x(i)$, then following method is used to separate unvoiced and voiced parts.

$$if \qquad \frac{x(i)-meanVal}{SDev} > Threshold \qquad (1)$$

$$then, \qquad Voiced(i) = 1$$
$$else, \qquad Voiced(i) = 0$$

values which are equal to Zero, are discarded and the remaining part represent the voiced part of speech as shown in the fig.2 (a) and (b).



(a) (b)

Fig. 2: (a) Original energy plot of voice signal 'Nine' (b) Energy plot of signal 'Nine' after end point detection

ISSN: 2005-4238 IJAST

2.1 Feature Extraction

Speech signal differs in various attributes like pitch, short time energy, zero crossing rate and power spectral density. These attributes are extracted in the form of features and are used for recognition purpose [2].

A robust word recognition system is adapted based on multi-resolution analysis. The Wavelet Transform is used to decompose the input speech signal into approximation and detailed coefficients as shown in figure 3, [8-12], [15]. Each Wavelet is a shifted scaled version of mother wavelet. The property of Wavelet transforms which make it suitable for use over conventional methods is the ability of Wavelets to capture the localized features. The feature of speech signal is obtained from approximation and detailed coefficients by decomposing the voice signal up to 6 levels [5].

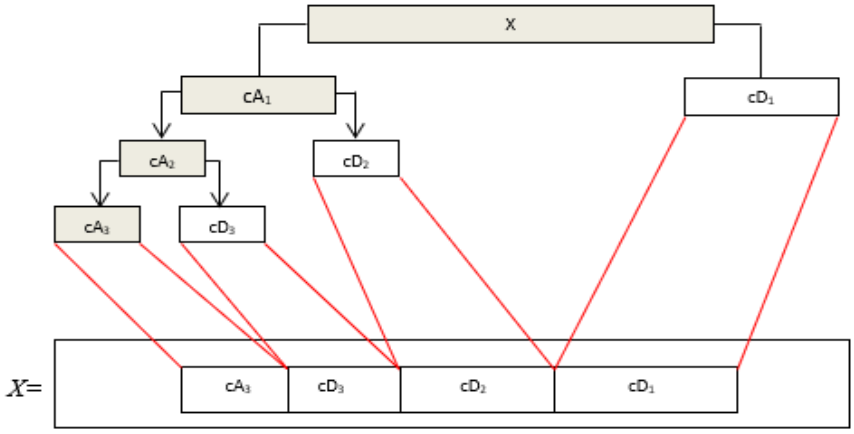


Fig. 3: Wavelet decomposition of a signal up to four levels

To calculate the feature of speech signal, percentage energy corresponding to each Wavelet coefficient is calculated. These energy values are further used to train the neural network which is used for the classification purpose.Decomposing a signal up to level six gives one approximate coefficient and six detailed coefficients. The wavelet decomposition of signal nine before and after decomposition is shown in figure 4 (a) and (b).

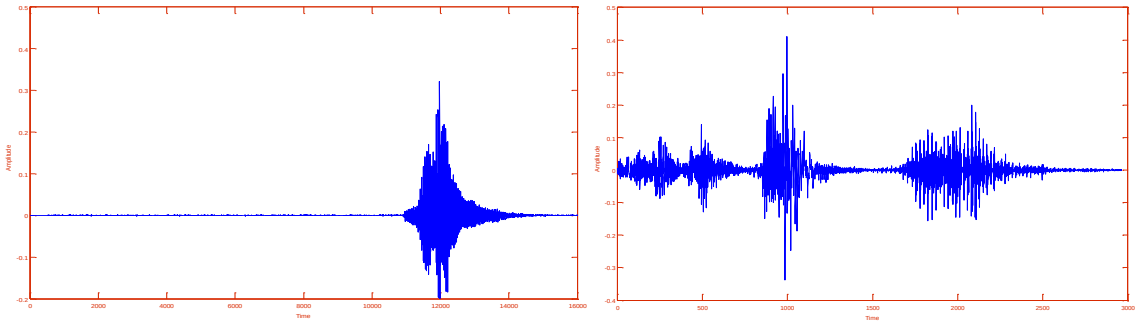


Fig. 4: (a) Signal 'Nine' before wavelet decomposition(b) Signal 'Nine' after wavelet decomposition

ISSN: 2005-4238 IJAST

2.2 Classification

After feature extraction, samples are to be classified into various categories. Classification is done with the help of Neural Network [11], [13, 14]. The multilayer feedforward neural network is used for pattern recognition problem. The work flow for the general neural network design process has five primary steps[1] Collection of data, Create and Configure the network, Training of the network, Validation of the network (post-training analysis), and Use the network.Before beginning the network design process, sample data collection was done.

Once the data has been collected, the next step in training a network is to create the network object. The function *feedforwardnet* creates a multilayer feedforward network. We can configure the network by providing the arguments to the network [4].Two arguments can be provided to the function feedforward to configure the network. The first is about number of neurons in hidden layer. More the number of neurons more is the computation power required. The second argument is the training function. The default transfer function for hidden layers is *tansig* and the default for the output layer is *purelin*[10]. The Configuration of proposed Neural Network used for speech recognition has 7 Inputs,10 outputs, 1 Hidden layer, 9 Hidden Neurons andScaled Conjugate GradientTraining Algorithm. The development of neural network configuration using nntool in MatLab is depicted in figure: 5.

Table 1: Configuration of proposed Neural network

	Input Layer	**Hidden Layer**	**Output Layer**
No of Neurons	09	09	10
Activation function	Linear	Tan-Sigmoid	Linear

In this paper, SCG (Scaled Conjugate Gradient) was used to train the neural network. For optimized learning, a global error function is minimized in NN which depends on the weights in the network and a multivariate function.

$$F = mse = \frac{1}{N}\sum_{i=1}^{N}(e_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(t_i - a_i)^2 \qquad (2)$$

Network gives the best performance when 9 Neurons taken in Hidden layer and transfer function is chosen as Tan-Sigmoid (fig. 6).

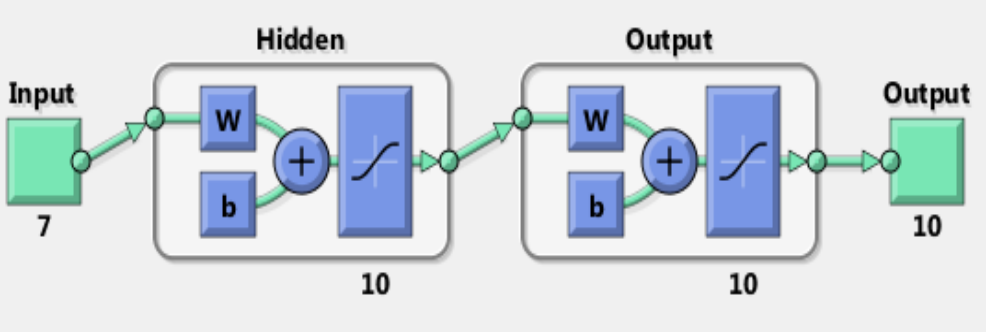


Fig. 5: Diagram of a feed-forward network used for training in project

ISSN: 2005-4238 IJAST

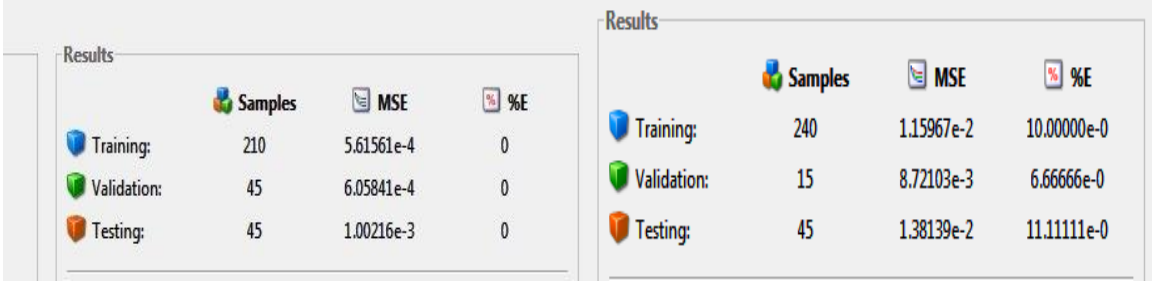


Fig. 6.Training results with 9 Hidden Neurons and 10 hidden neurons in a layer

After the network is trained and, the network object can be used to calculate the network response to any input.

Table. 2: Output values for some input vectors

Output (y)	5th Input (‘*Zero*’)	34th input (‘*One*’)	65th input (‘*two*’)	96th input (‘*Three’*)	127th input (‘*Four*’)	168th input (‘*Five*’)	201th input (‘*Six*’)
y_0	**0.9968**	0.0001	0.0004	0.0000	0.0081	0.0000	0.0000
y_1	0.0489	**0.9293**	0.0092	0.0000	0.0350	0.0029	0.0158
y_2	0.0000	0.0000	**1.0000**	0.0170	0.0003	0.0000	0.0000
y_3	0.0000	0.0000	0.0011	**0.9967**	0.0001	0.0013	0.0022
y_4	0.0019	0.0000	0.0169	0.0000	**0.9893**	0.0015	0.0000
y_5	0.0000	0.0000	0.0000	0.0150	0.0000	**0.9996**	0.0050
y_6	0.0003	0.0053	0.0000	0.0000	0.0000	0.0004	**0.9988**
y_7	0.0000	0.0168	0.0000	0.0000	0.0000	0.0001	0.0000
y_8	0.0039	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
y_9	0.0000	0.0000	0.0000	0.0018	0.0000	0.0233	0.0009

From the table 3, we can verify that for each given input vector, a trained network classifies with an accuracy of more than 99%. In table 3, NN performance parameters have been given.

Table. 3: Performance table

No. of Hidden Neurons	No of Epochs	MSE	Percentage error		
			Training	Validation	Testing
6	91	0.05405	50.00	53.00	46.00
7	118	0.11118	9.04	11.11	13.33
8	90	0.02354	20.00	22.22	17.17
9	74	0.00872	0.00	0.00	0.00
10	57	0.01314	10.00	6.66	11.11
11	101	0.05355	48.09	53.33	55.55

ISSN: 2005-4238 IJAST

3. Conclusion

A speaker independent speech recognition systemis designed and tested successfully to take the advantages of ANN and Wavelets. The results show reasonably good success in recognizing words from various speakers. An accuracy of more than 99% was achieved in recognizing words. Key research challenges for the future are acoustic robustness, efficient constraints for the access of a very large lexicon and well-organized methods for extracting conceptual representations from word hypotheses.

References:

[1] J. Kacur, G. Rozinaj and S. Herrera, "Speech Signal Detection in a Noisy Environment Using Neural Networks and Cepstral Matrices", Journal of Electrical Engineering.55, pp.5-6, 2004.
[2] Xiaolan Zhao, Zuguo Wu and JihaiNiu, "Speech Signal Feature Extraction Based on Wavelet Transform", International Conference on Intelligent Computation and Bio Medical Instrumentation, Wuhan Hubei, , pp. 179-182, 2011.
[3] G. Gosztolya, G. Kovacs, "Spoken Term Detection from Noisy Input", IEEE International Symposium on Applied Computational Intelligence and Informatics, Timisoara, pp. 91-96, 2011.
[4] N.S. Dey, R. Mohanty and K.L. Chugh, "Speech and Speaker Recognition System using Artificial Neural Networks and Hidden Markov Model", IEEE International conference On Communication Systems and Neural Network Technologies, Rajkot, pp. 311-315, 2012.
[5] C.J. Long, S. Dutta, "Wavelet based feature extraction for phoneme recognition", Conference of spoken language processing, Vol. 1, pp. 264-267, 2011.
[6] G. Saha, S. Chakraborty and S. Senapati, "A new Silence Remove and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications", Proc.Of the NCC, 2005.
[7] H. Zhang, H. Hu, "An Endpoint Detection Algorithm based on MFCC and Spectral Entropy using BP NN", International Conference on Signal Processing System, Dalian, pp. 509-513,2010.
[8] L.R. Rabiner and R.W. Schafer, Digital Processing of Speech Signals, 1st Edition, Pearson Education, Prentice Hall of India, 2001.
[9] K.R. Borisagar, D.G. Kamdar, B.S. Sedani, "Speech Enhancement in Noisy Environment Using Voice Activity Detection and Wavelet Thresholding", IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, pp. 1-5, 2010.
[10] W.J. Palm, "MATLAB for engineering application", New York: Mc-Grow Hill, 1999.
[11] T. Hughes, K. Mierle, "Recurrent Neural Networks for Voice Activity Detection", IEEEInternational conference on Acoustics, Speech and Signal processing, Vancouver, pp.7378 7382, 2013.
[12] CHENG Hao, LüMing and TANG Bin etc, "DSSS Signal Detection Based on Wavelet Decomposition," Telecommunication Engineering, pp.67-70, 2006.
[13] R. Vergin, D. O'Shaughnessy, and A. Farhat, "Generalized melfrequencycepstral coefficients for large-vocabulary speaker independent continuous-speech recognition," IEEE Transactions on Speech and Audio Processing, pp.525-532, 1999.
[14] MuzhirShaban Al-Ani, Thabit Sultan Mohammed and Karim M. Aljebory, "Speaker Identification: A Hybrid Approach Using Neural Networks and Wavelet Transform" Journal of Computer Science Publications, pp.304-309, 2007.
[15] Aik Ming Toh, Roberto Togneri, Sven Northolt, "Spectral Entropy as Speech Features for Speech Recognition", Proceedings of PEECS, pp. 22-25, September 2005.

ISSN: 2005-4238 IJAST