Performance Enhancement of Automatic Speech Recognition System using Euclidean Distance Comparison and Artificial Neural Network

Anurag Bajpai[1]
Department of Electronics Design & Technology
NIELIT Gorakhpur,
Gorakhpur, India
anuragsietm@gmail.com

Umang Varshney[2]
Department of Electronics Design & Technology
NIELIT Gorakhpur,
Gorakhpur, India
uumangvarshney@gmail.com

Deepam Dubey[3]
Department of Electronics Design & Technology
NIELIT Gorakhpur,
Gorakhpur, India
deepamdubey@nielit.gov.in

***Abstract*–The paper shows how an Automatic Speech Recognition (ASR) system be efficiently designed for ubiquitous control. The design is based on an algorithm to extract isolated words from a continuous speech signal. The feature extraction of the voiced part of speech signal is done byMel Frequency CepstralCoefficients (MFCC) whereas Artificial Neural Network (ANN) is used for training and pattern. The increased rejection of unauthorized speech commands is obtained by the decisions based on Euclidean distance, measured between trained and tested voiced commands. SNR of the signal is also improved,at the pre-processing stage.**

Index terms:Automatic Speech Recognition (ASR), Mel Frequency Cepstral Coefficients (MFCC), Artificial Neural Networks (ANN), Euclidean distance, SNR.

I. INTRODUCTION

For humans, speech plays an important role to communicate efficiently. Several languages are used for communication in different parts of the world. In recent years, with advancement in machine learning technologies, thedemand of Automatic Speech Recognition Systems has risen. ASR systems have basically three steps: Acquisition of speech signal, Feature extraction and Pattern matching.

Feature extraction is done to extract spectral characteristicsof speech signal and there are numerous ways to do it, such as Linear Predictive Analysis (LPC)[1], Power Spectral analysis, Mel Frequency Cepstral Coefficients (MFCC) [2], Relative Spectral Filtering of log domain Coefficients (RASTA) [3], Linear Predictive Cepstral Coefficients (LPCC) and Mel Scale Cepstral. Pattern matching is done to match the extracted feature vector to the previously trained vectors in the model.There are various techniques such as Gaussian Mixture Model (GMM) [4], Correlation Method, Hidden Markov Models (HMM) [5] and Artificial Neural Networks (ANN) [6] used for pattern matching.

One of the main challenges in the ASR systems has been to recognize a continuous speech signal. To overcome this problem, an algorithm based model, is developed which can extract isolated words from any continuous speech signal. MFCC as feature extraction technique and ANN as pattern matching technique are used in it. Rejection of unauthorized commands is a strenuous taskfor building such systems, which are voice controlledby very few authorized commands. Euclidean distance measurement technique is a solution to overcome such problems. The use of Euclidean distance concept makes it feasible for the system to reject the unauthorized commands during the first stage of processing and system can skip the neural network based second stage comparisons to enhance the speed.

In this paper, a brief overview of speech recognition techniques is provided. Section II describes about the methodology for automatic speech recognition and about the feature extraction and pattern matching techniques, which are used. It provides information regarding the algorithm to extract isolated words from continuous speech signal. Section III analyses the training and testing results of ASR system and the SNR is improved. Section IV is about the conclusion and need for further research.

II. METHODOLOGY

The automatic speech recognition system is based on the spectral characteristics of speech signal. The basic

978-1-5090-6785-5/18/$31.00

approachto develop the system is shown in the Fig.1.

A. Acquisition of Speech Signal

The maximum energy content of the speech signal is in the range of 300Hz to 5 KHz. According to the Nyquist

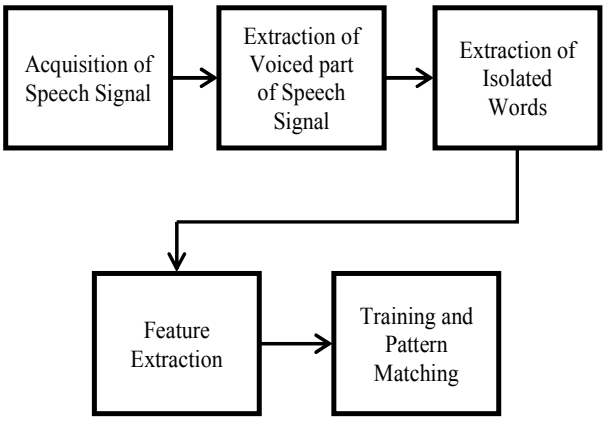

Fig.1. Block Diagram of ASR System

theorem, to reconstruct a signal of frequency f_s, it should be sampled at least at the rate of $2*f_s$ samples/sec. So, for this paper work the acquired speech signal is sampled with a sampling rate 16000 samples/sec as shown in Fig.2.

B. Extraction of Voiced part of Speech Signal

In the speech signal, there is a voiced and an unvoiced part. To improve the SNR, the spectral characteristics are extracted from the voiced part of the speech signal. The complete speech signal is divided into frames of a fix length. The maximum amplitude of samples in the frame is used as the decision logic for selecting or dropping the frame. Finally, all the selected frames are serially stored as voiced part of the speech.

C. Extraction of Isolated Words

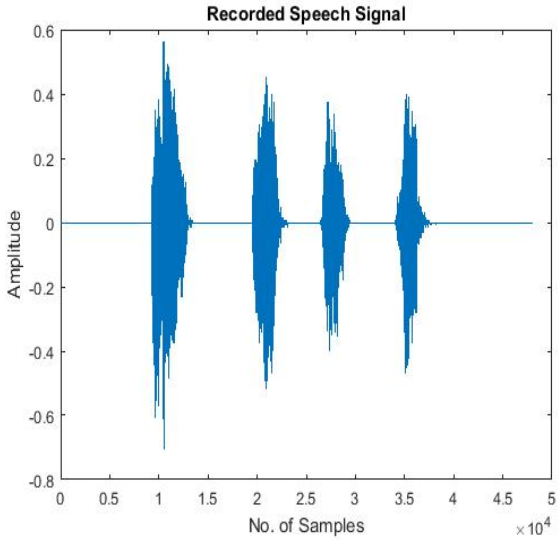

Fig.2. Recorded speech signal

After acquiring the voiced part of speech signal, isolated words are extracted as shown in Fig.3. . For this, the serial number of frames, selected in the voiced part of signal, is used. The difference between the two consecutive serial numbers is used as the decision logic. If the difference is equal or less than two, theframe is in the same word with which the previous frame is associated else a new word starts and the frame is stored in the new word. The process is repeated till the last frame to extract all isolated words. All

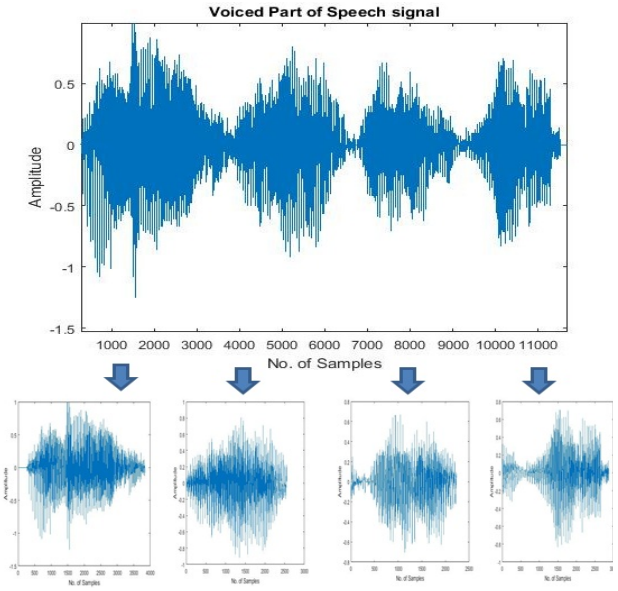

Fig.3. Voiced part of speech signal and extracted isolated words

the isolated words are stored in different variables.

D. Feature Extraction

MFCC is used for feature extraction in which the frequency is converted from Hz scale to Mel scale. This frequency wrapping is a better representation of a speech signal as in Mel scale, calculations are done in decibels and the hearing capacity of human ear is also measured in decibels. The cepstrum is denoted mathematically as following equation:

$$c(n) = \text{ifft}(\log|\text{fft}(s(n))|) \quad \text{.....(1)}$$

Where s(n) is the sampled speech signal, and c(n) is the signal in the Cepstral domain. Cepstral transform of speech signal makes the analysisincredibly simple.The extracted spectral characteristics are used in the training and testing of the ASR system.The block diagram for MFCC extraction is shown in the Fig.4. The extracted feature vector of word FOUR is shown in Fig.5.

E. Training and Pattern Matching

The extracted MFCC feature vectors are normalized and a fixed size feature vector is calculated to train the system. Neural networks are used for training and pattern matching. A neural network has simple processing elements operating in parallel which can acquire, save, and utilize experiential

978-1-5090-6785-5/18/$31.00 © 2018 by IEEE

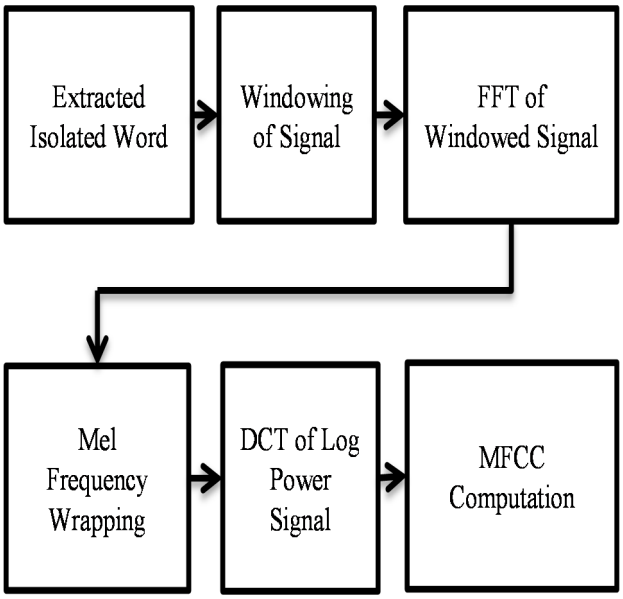

Fig.4. Block diagram for MFCC Extraction

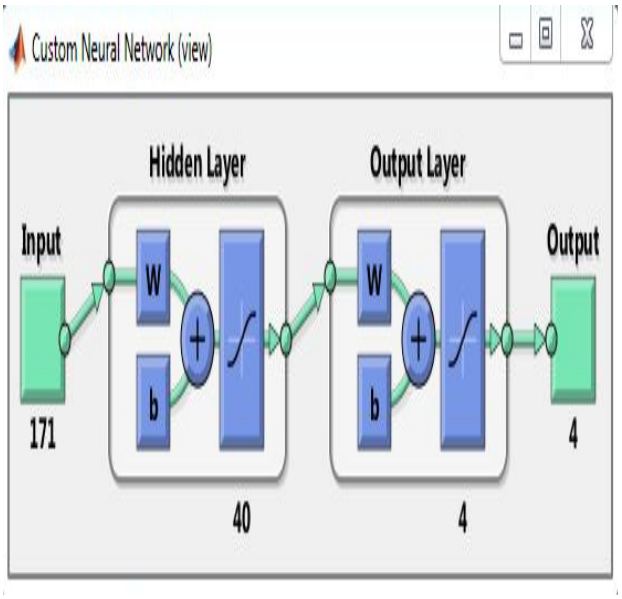

Fig.6. Neural Network Model

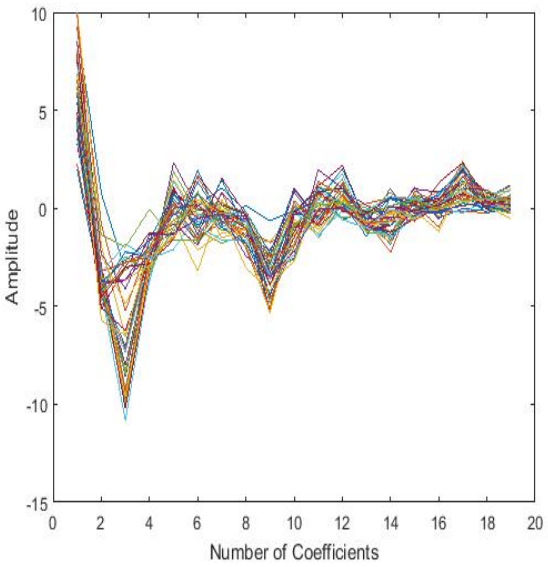

Fig.5. MFCC for word "FOUR"

knowledge. There are basically two types of neural networks based on their interconnections: Feed forward neural network and Recurrent Neural Network. The algorithms used in the learning of neural networks are as follows:

1. Scaled Conjugate Gradient Back Propagation (SCG)
2. Resilient Back propagation (RP)
3. Polak-Ribiere Conjugate Gradient (CGP)
4. Conjugate Gradient with Powell \ Beale Restarts (CGB)

With the use of above algorithms four different neural networks are made and trained. Sigmoid is used as the activation function of the neural network. One of the neural network model used in the system is shown in Fig.6. The network is used to train words "ONE", "TWO", "THREE" and "FOUR". The size of feature vector of each word is fixed to 171. There are 40 hidden layers, used in the neural network model. 40 hidden layers are the minimum number of hidden layers at which the system gets maximum accuracy in recognition of the words. This conclusion was made after different number of trials of the neural network with different hidden layers. The network adjusts its weights, which enhances the recognition accuracy through training. The switching of four electric bulbs is controlled with the help of the spoken command, and to perform this, two networks are trained. The first network is trained with the MFCC feature vectors of words "ON" and "OFF". The second network is trained with the MFCC feature vectors of words "ONE", "TWO", "THREE" and "FOUR" with a training data set consisting 20 samples of each word. The extracted feature vectors are also stored for use in Euclidean distance measurement to reject unauthorized voice commands.In the testing phase, the user speaks a continuous command such as "ON ONE", with the help of the proposed algorithm the isolated words "ON" and "ONE" are extracted. Now the Euclidean distance is measured between the extracted words and the previously stored words. If the distance is more than a specified value, the command is rejected as an unauthorized command; otherwise the "ON" word is given to the first network and "ONE" is given to the second network. Now based on the output given by both the networks, the system understands that the spoken command is "ON ONE". Now the system sends a command via Bluetooth (HC-05) to Arduino2560. The microcontroller understands the command and generates an output signal to control the switching state of the bulbs via relay control.

III. RESULTS AND DISCUSSION

The performance of the designed voice controlled system is tested for four different training algorithms. 276 samples for four different words are taken. It is observed that training vectors utilizing the SCG algorithm tend to produce more accurate voice recognition than other methods, as clearly evident from Table I.

Table I. Recognition Accuracy of 276 samples of 4 different words

S. No.	Accuracy Calculation	
	Training Algorithm	*Accuracy*
1.	SCG	**96.74%**
2.	RP	95.65%
3.	CGP	95.29%
4.	CGB	95.65%

Euclidean distance of speech recognition is estimated after matching the extracted feature vectors with previously stored coefficients of different words. The recognition accuracy is indicated in Table II.

Table II. Euclidean Distance Measurement

S. No.	Euclidean DistanceCalculation	
	Recognition	*Accuracy*
1.	Authorized word	95.5%
2.	Unauthorized word	96%

After SCG training, a performance plot of the neural network is obtained for one of the words is shown in Fig.7.

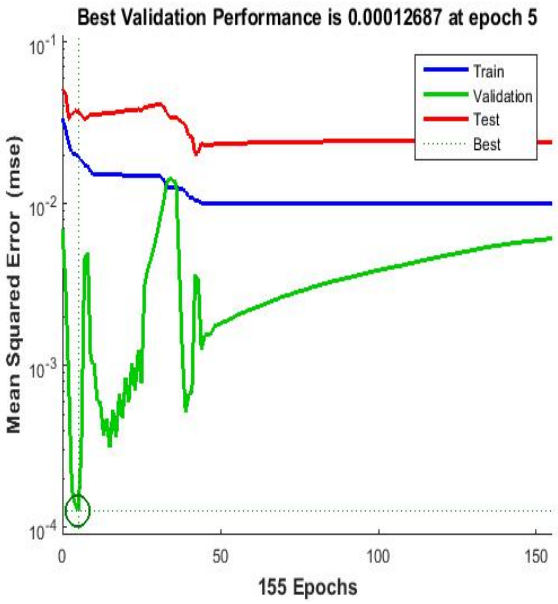

Fig.7. Performance Plot of Neural Network

The minimum mean square error occurs at a certain epoch.The SNR of the speech signal, before processing and after processing, is computed. Since the preprocessing stage consists of extraction of only the voiced part of the signal, a significant level of SNR improvement is achieved. The SNR computation plot is depicted in Fig.8.

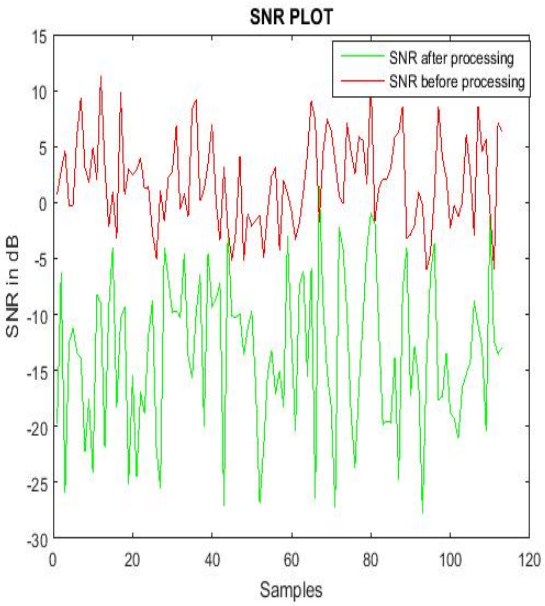

Fig.8. SNR Plot

The input SNR and the output SNR after MFCC extraction is indicated in Table III.

Table III. SNR calculation of Speech signal (Unit: dB)

S. No.	SNR Calculation	
	Input SNR(dB)	*Output SNR After MFCC(dB)*
1.	-5	1.1517
2.	-10	-.0971
3.	-15	-.109

IV. CONCLUSION

The system is speaker independent and is suitable for real time processing.Each time the testing command is matched with the trained commands, the system includes the input speech data into its training data-set, thereby enriching the training database, for next comparisons. The use of Euclidean distance measurement and after that neural networks make the system two level secure. Due to this, the accuracy of recognition and rejection of unauthorized voice gets improved,with repeated use of the system.

978-1-5090-6785-5/18/$31.00 © 2018 by IEEE

V. REFERENCES

[1] M. R. Sambur, N. S. Jayant, "LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise", *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-24, pp. 488-494, Dec. 1976.

[2] Jihyuck Jo , Hoyoung Yoo and In-Cheol Park "Energy-Efficient Floating-Point MFCC Extraction Architecture for Speech Recognition Systems", IEEE Transactions on Very Large Scale Integration (VLSI) Systems ,Volume: 24, Issue: 2,pp. 754 – 758, Feb. 2016 .

[3] Ram Singh, Preeti Rao, "Spectral Subtraction SpeechEnhancement with RASTA Filtering", Proceeding of NationalConference on Communications (NCC), Kanpur, India, 2007.

[4] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Statistical approach to enhancing esophageal speech based on Gaussian mixture models," Proc. ICASSP2010, pp. 4250–4253 (2010).

[5] Myungjong Kim, Younggwan Kim,Joohong Yoo "Regularized Speaker Adaptation of KL-HMM for Dysarthric Speech Recognition" , IEEE Transactions on Neural Systems and Rehabilitation Engineering Volume: 25, Issue: 9, pp: 1581-1591, Sept. 2017.

[6] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," In Proc. ICASSP'13, pp. 7893–7897, 2013.

978-1-5090-6785-5/18/$31.00 © 2018 by IEEE