Design of Automatic Speech Recognition System using ANN (Artificial Neural Networks) for Ubiquitous Control

Anurag Bajpai[1] Deepam Dubey[2]
[1,2]Department of Electronics Design & Technology
[1,2]NIELIT, Gorakhpur, India

Abstract— This paper is about automatic speech recognition system for controlling specific tasks with the help of a certain number of voice commands. It primarily deals with enhancing the accuracy of the voice based automation system. The feature extraction of speech has been done by the MFCC (Mel Frequency Cepstral Coefficients) and testing or matching has been performed with the help of ANN (Artificial Neural Networks). For enhancing the performance Euclidean distance is also measured to reject the unauthorized voice commands.

Key words: Automatic Speech Recognition, MFCC (Mel Frequency Cepstral Coefficients), ANN (Artificial Neural Networks), Euclidean Distance

I. Introduction

Speech is an easy and efficient mode of communication for people to communicate with each other. Numerous languages are spoken in this world [1]. With the aid of speech a person can interact with the machine, e.g. computer rather than using any interface(s). This can be achieved by an Automatic Speech Recognition (ASR) system which captures the words spoken from a person via a microphone, feeds them to the system model and efficiently identifies the speaker's speech with terrific accuracy.

It has always been a bit of a difficult challenge to implement an efficient ASR system mainly due to the following reasons:

- There is a voiced part and an unvoiced part in the speech.
- Degradation in speech signal due to noise.
- Variation in the frequency, pitch and other spectral characteristics.

II. Automatic Speech Recognition

The automatic speech recognition system is based on the spectral characteristics which can be derived easily from the speech signal. The basic approach to develop this system is listed as follows:

- Recording of the speech
- Feature Extraction
- Feature Matching (Pattern Matching)

A. Recording of the speech

Speech signal consists of frequencies of low level to those around 4 –5kHz. From the Nyquist theorem if fs samples are taken then it can reconstruct the original signal to up to fs / 2 frequency. So speech signal is sampled at the rate of 16000 samples per second for proper frequency coverage.

B. Feature Extraction

Feature extraction is done to capture the spectral characteristics of the speech signal. Those features are extracted which are somewhat invariant to changes in the speaker. Some of the feature extraction techniques are:

- Power Spectral Analysis
- Linear Predictive Analysis (LPC) [4]
- Mel Frequency Cepstral Coefficients (MFCC) [3]
- Mel Scale Cepstral Analysis
- Linear Predictive Cepstral Coefficients (LPCC)
- Relative Spectral Filtering of Log domain coefficients (RASTA) [5]

C. Feature Matching (Pattern Matching)

The second step after feature extraction is to train the model with the help of a training vector. For training and testing purposes we can use a variety of models. Some of them are:

- Hidden Markov Models (HMM) [6]
- Artificial Neural Networks (ANN) [7]
- Correlation Method
- Gaussian Mixture Model (GMM) [9]

III. MFCC

Mel-Frequency Cepstral Coefficients feature extraction technique is being used to mimic the working of human auditory system [2] [8]. MFCC help in changing the spectral information of the speech signal in terms of coefficients that are used as training vectors in the speech recognition. MFCC feature extraction can be done as following steps:

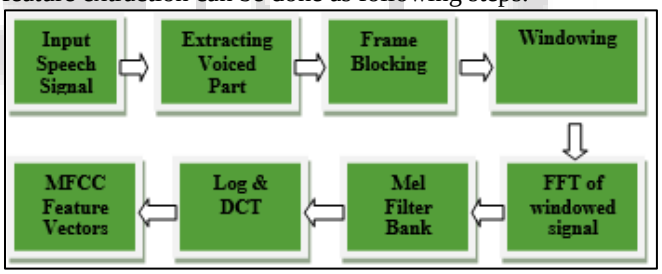

Fig. 1: Block diagram for MFCC computation

A. Step 1

The speech (word ON) signal is recorded with a sampling rate of 16000 samples / sec and stored as on.wav.

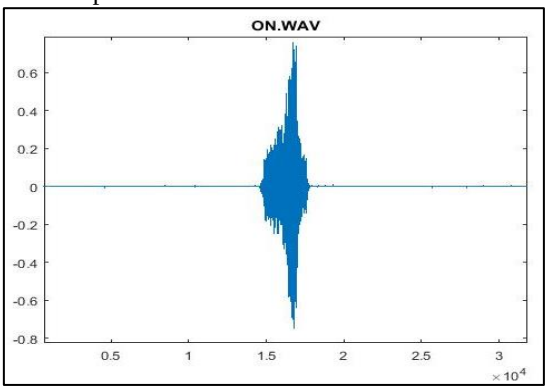

Fig. 2: Recorded ON.wav before processing

B. Step 2

The voiced part is extracted from the recorded signal

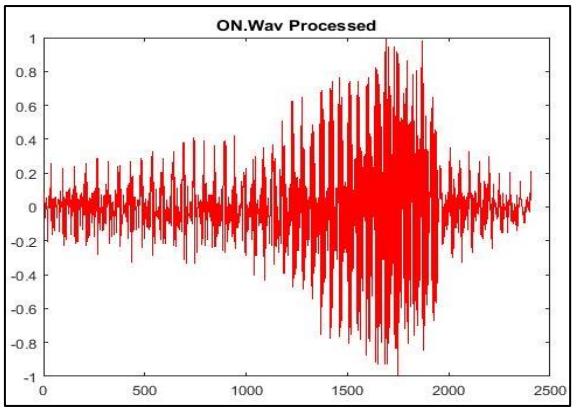

Fig. 3: Voiced part of ON.wav

C. Step 3

Frame blocking and windowing is done to the signal

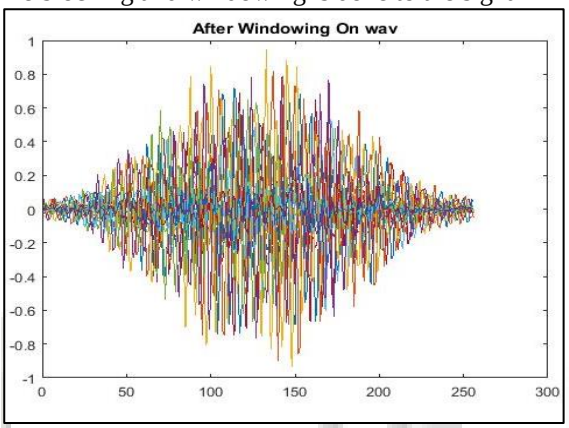

Fig. 4: Windowed On.wav

D. Step 4

FFT of the windowed signal is performed.

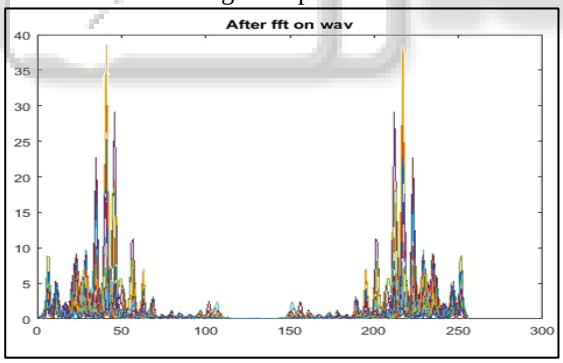

Fig. 5: After FFT on.wav

E. Step 5

The signal frequency is mapped with the Mel frequency with help of Mel filter bank

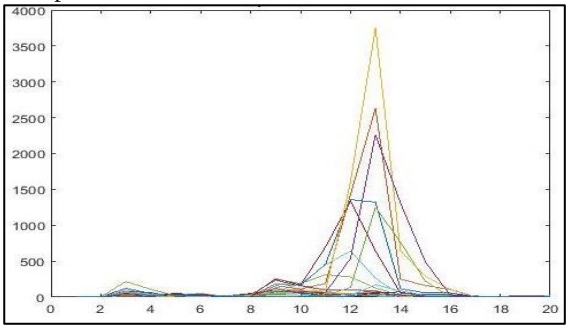
Fig. 6: Mel spectrum of the signal

F. Step 6

Taking log and DCT (Discrete Cosine Transform) of the signal and computing the MFCC of signal

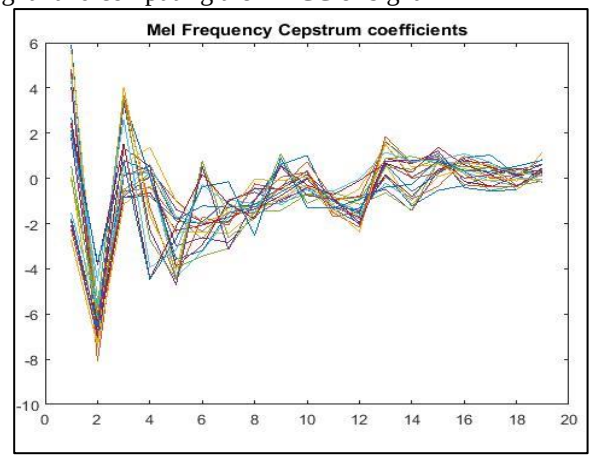

Fig. 7: MFCC of the signal on.wav

After computation the MFCC are normalized and a training feature vector is calculated from normalized MFCC.

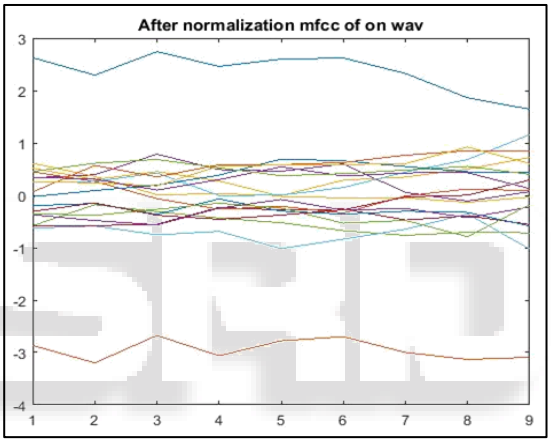

Fig. 8: Normalized MFCC of on.wav

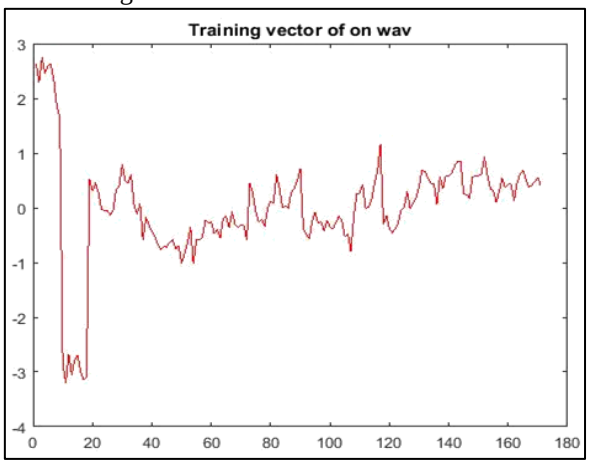

Fig. 9: Training vector of on.wav

IV. ANN

A Neural Network is an entity composed of many simple processing elements operating in parallel that can acquire, save, and utilize experiential knowledge. Artificial neural networks can be classified on account of their interconnections some are as follows:

– Feed forward neural network
– Recurrent neural network

In feed forward neural networks, link goes only in one direction i.e. forward direction while in recurrent neural networks link can go in any direction. Mostly Back Propagation (a type of recurrent neural network) is used for pattern matching in the speech recognition systems. For learning of the neural network some algorithms which are used in simulation are as follows:

- Polak-Ribiere Conjugate Gradient (CGP)
- Resilient Back propagation (RP)
- Conjugate Gradient with Powell \ Beale Restarts
- Scaled Conjugate Gradient Back Propagation (SCG)

Sigmoid and Softmax are used as activation functions.

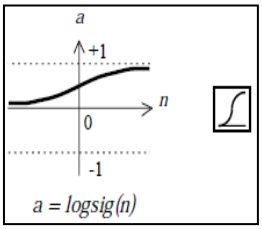

Fig. 10: Log-Sigmoid Transfer Function

V. Experimental Setups

The experiments were performed on three recorded speech signals. 15 samples of each word 'on', 'off' and 'stop' were recorded. Speech samples of three persons were taken.

In the training phase the MFCC was computed and different neural networks such as feed forward and recurrent networks were made with different learning algorithms and were trained with the MFCC feature vectors. The computed MFCC were also stored in a reference matrix.

In testing phase MFCC was computed and its Euclidean distance was computed with each of the feature vectors stored in the reference matrix. If the distance is more than a specified level then mismatch is shown else the tested feature vector class (on, off, stop) is decided by the different neural networks.

A hardware setup was implemented to demonstrate the voice based automation with the help of MATLAB and Arduino interface. The components are listed as follows:

- Arduino Mega 2560
- Bluetooth Module
- LCD
- LED

A. *Arduino Mega2560*

The Arduino Mega 2560 is a microcontroller board based on the ATmega2560. It contains 54 digital input/output pins (of which 15 can be used as PWM outputs), 16 analog inputs, 4 UARTs (hardware serial ports), a 16 MHz crystal oscillator, a USB connection, a power jack, an ICSP header, and a reset button.

B. *Bluetooth Module*

Bluetooth module HC-05 is used for serial communication with Matlab.

C. *LCD*

A 16*2 LCD (1602) was used to display the spoken words or mismatched output.

D. *LED*

The state of LED was controlled with the help of voice commands 'on' and 'off'.

The network gives a good performance when it is trained with limited data set just as it was trained for three inputs for this work.

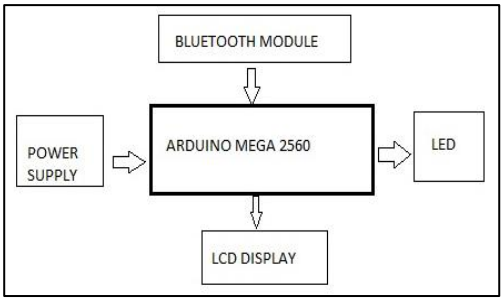

Fig. 11: Hardware assembly

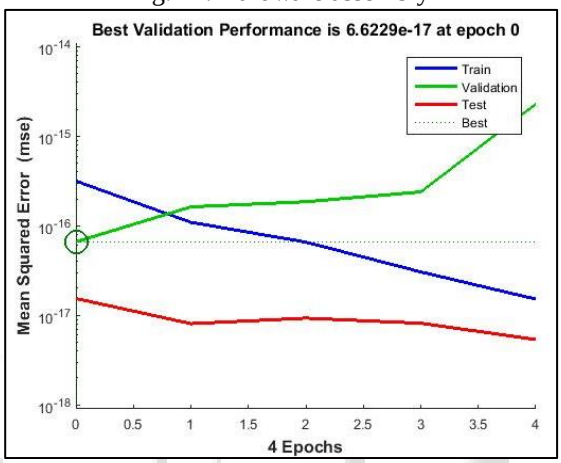

Fig. 12: Performance of network

VI. Conclusion

Due to the use of Euclidean distance concept the accuracy of the system is increased and it is able to reject the unauthorized spoken words. As the voiced part of the speech is being used so the Signal to Noise ratio (SNR) of the system is also improved and it can detect signals even in the noisy atmosphere. The system can be used for real time control with a specific number of commands. The performance of the system can be further improved by increasing the number of training data set.

References

[1] Hajer Rahali, Zied Hajaiej, and Noureddine Ellouze, "Robust Features for Impulsive Noisy Speech Recognition Using Relative Spectral Analysis", International Journal of Computer, Information, Systems and Control Engineering Vol.8 No.9, pp. 1421-1426, 2014.

[2] Santosh Gaikwad, Bharti Gawali, and Pravin Yannawar, "Performance Analysis of MFCC & DTW for Isolated Arabic Digit", International Journal of Advanced Research in Computer Science, Vol. 2 (1), pp. 513-518 Jan. –Feb, 2011.

[3] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, August 1980.

[4] Young-Giu Jung, Mun-Sung Han, and Sang Jo Lee, "Development of an Optimized Feature Extraction Algorithm for Throat Signal Analysis", Electonics and Telecommunication research Institute Journal, Volume 29, Number 3, pp.292-299, June 2007
[5] Ram Singh, and Preeti Rao, "Spectral Subtraction Speech Enhancement with RASTA Filtering", Proceeding of National Conference on Communications (NCC), Kanpur, India, 2007.
[6] Lawrence R. Rabiner,"A tutorial on Hidden markov models and selected applications in speech recognition"
[7] 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.
[8] Z. András, K. Daniil, S. Ralf and N. Hermann, "Using Multiple Acoustic Feature Sets for Speech Recognition", Speech Communication, Vol. 49, No. 6, 2007, pp.514-525.
[9] 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)

