A MATLAB based improved Speech Enhancement and Recognition using Spectral Subraction Method,MFCC,GMM

[1]Edla Manichandana, [2]Ch Manasvini Abhigna, [3]Bojja Soumya, [4]Ch Anusha

[1]Student, [2]Student, [3]Student, [4]Assistant Professor
[1234]Department of Electronics and Communication Engineering,
[1234]G. Narayanamma Institute of Technology & Science, Hyderabad, Telangana, India

Abstract—Speech recognition is the process of automatically recognizing a particular word from a particular speaker based on the information contained in each voice wave. This document reviews inventions and technological advances in the field of speech recognition and also focuses on the various steps required to identify speaker using MATLAB programming. In this document, we first apply Spectral Subtraction Method to perform speech enhancement and noise removal. This background noise is successfully removed by the Wiener Filter application. In addition, the technique adopted here develops code using MATLAB programming. In this article, nine audio samples were recorded through a microphone and the system was trained according to the recoded audio samples. The MFCC function of the voice sample was calculated and the words were distinguished according to according to the energy associated with each sampled word and recognized using GMM.

***IndexTerms*—Wiener filter, Speech Recognition, Speech Enhancement, MFCC, GMM**

I. INTRODUCTION

Speech is one of the most important medium by which a communication can take place. With the invention and widespread use of mobiles, telephones, data storage devices etc. has provided a major help in setting up of speech communication and its analyzing. The proposed model of speech recognition technology contains signal pre-processing , which describes a procedure to process signal with endpoint detection, pre-emphasis, framing and windowing; And then it's characteristic parameter extraction technology, author mainly used Mel Frequency Cepstral coefficient extraction and related speech recognition algorithm in the experiment. For analyzing the extracted parameter, cross-correlation was utilized.

II. SPEECH ENHANCEMENT

Noise plays a vital role in speech enhancement. Thus noise estimation is one of the major part while performing the speech recognition task. Therefore, it is understood if the estimated noise is low it will not affect the speech signal but if the noise is high then speech will get distorted and loss intelligibility. So to remove the noise we have two techniques i.e. speech degradation and speech enhancement. In this paper we use speech enhancement technique that enlightens upon the major use Speech Degradation technique i.e. removal of Gaussian noise from the original speech wave.

In this technique firstly the degraded signal i.e. original signal mixed with Gaussian noise is first converted to the frequency domain with the help of FFT tool in MATLAB Programming. Then higher frequency noise components are then removed with the help of 3rd order Butterworth low pass filter. The reason to choose Butterworth filter here because it has the capability to filter the Gaussian noise more closely & approximates an ideal low pass filter as the order, n, is increased. This methodology is referred as Spectral Subtraction Method. In this method, the noise spectrum is estimated during speech pauses, and is subtracted from the noisy speech spectrum to estimate the clean speech.

Consider a noisy signal which consists of the clean speech degraded by statistically independent additive noise as,

$$y[n] = s[n] + d[n] \quad (1)$$

Where y[n],s[n] and d[n] are the sampled noisy speech, clean speech, and additive noise, respectively. It is assumed that additive noise is zero mean and uncorrelated with the clean speech. Because the speech signal is non-stationary and timevariant, the noisy speech signal is often processed frame-by-frame. Their representation in the short-time Fourier transform (STFT) domain is given by

$$Y(\omega, k) = S(\omega, k) + D(\omega, k) \quad (2)$$

Where k is a frame number. Throughout this paper, it is assumed that the speech signal is segmented into frames, hence for simplicity, we drop k. Since the speech is assumed to be uncorrelated with the background noise, theShort-term power spectrum of y[n] has no cross-terms.

Hence,

$$|Y(\omega)|^2 = |S(\omega)|^2 + |D(\omega)|^2 \quad (3)$$

The speech can be estimated by subtracting a noise estimate from the received signal.

$$|S(\omega)|^2 = |Y(\omega)|^2 - |D(\omega)|^2 \quad (4)$$

The estimation of the noise spectrum $|D(\omega)|^2$ is obtained by averaging recent speech pauses frames, where M is the number of consecutive frames of speech pauses (SP). If the background noise is stationary, it converges to the optimal noise power spectrum estimate as a longer average is taken. The spectral subtraction can also be looked at as a filter, by manipulatingsuch that it can be expressed as the product of the noisy speech spectrum and the spectral subtraction filter (SSF) as:

$$|S(\omega)|^2 = |H(\omega)|^2 Y(\omega)|^2 \quad (5)$$

Where H (ω) is the gain function and known spectral subtraction filter (SSF). The H (ω) is a zero phase filter, with its magnitude response in the range of $0 \leq H(\omega) \leq 1$. To reconstruct the resulting signal, the phase estimate of the speech is also needed. A common phase estimation method is to adopt the phase of the noisy signal as the phase of the estimated clean speech signal, based on the notion that short-term phase is relatively unimportant to human ears5. Then, the speech signal in a frame is estimated.

$$S(\omega) = |S(\omega)| \wedge (|e\ j<Y(\omega)) = H(\omega)\ Y(\omega) \quad (6)$$

The estimated speech waveform is recovered in the time domain by inverse Fourier transforming S (ω) using an overlap and add approach. The spectral subtraction method, although reducing the noise significantly, has some severe drawbacks. The effectiveness of spectral subtraction is heavily dependent on accurate noise estimation, which is a difficult task to achieve in most conditions. When the noise estimate is less than perfect, two major problems occur, remnant noise with musical structure and speech distortion.

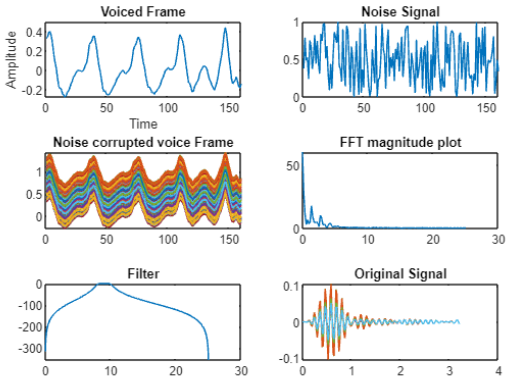

Fig.1 Result of Spectral Subtraction Method

Both Pitch analysis and Formant analysis is achieved using cepstral analysis. Pitch in terms of speech analysis can be defined as a technique which allows the ordering of sounds on a frequency-related scale. Pitch analysis helps us in identifying the state of speech of a person. The considered states are neutral, happy, and sad. Therefore it is very important to understand the concept of pitch analysis. The paper describes a technique that involves the extraction of basic parameters of Pitch Analysis. The general representation of pitch analysis with respect to time is shown in Fig.2.

Fig.2 Pitch analysis

Further with the help of MATLAB Programming we had prepared a code for Formant Analysis. With the help of this code we can easily calculate the first five formants that are present in .wav speech file, calculation of difference between the vector peak positions of these five formants, vector position of the peaks in the power spectral density were easy calculated and can be used to determine the speech file. The general waveform of formant analysis can be shown in Fig.3.

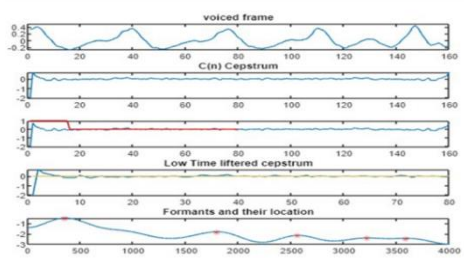

Fig.3 Formant Analysis

Formants in normal language can be defined as the spectral peaks of the sound spectrum. With the help of above discussed Pitch and Formant Analysis, a waveform comparison code was written with the help of MATLAB Programming. Thus, based on this code we can easily characterized Speech waveform files. In this process a reference .wav file was used which is then compared with the remaining .wav files. Moreover, a sorting routine is performed in which sorting and comparison of the average pitch of the reference file with all the other 5 .wav files. The technique further includes the comparison of formant vector of the reference file to all .wav files, and thus sorting for the top 3 average pitch correlations and then again sort these files by formant vectors correlations and then sort these by average pitch. In this way, we can easily recognize the speaker.

III. Speech Preprocessing

Framing is the process of breaking the continuous stream of speech samples into components of constant length to facilitate block-wise processing of the signal. In the same vein, speech can be thought of been a quasi-stationary signal and is stationary only for a short period of time. As a result, the speech signal is slowly varying over time (quasi-stationary) that is when the signal is examined over a short period of time (5-100msec), the signal is fairly stationary. Therefore, speech signals are often analyzed in short time components, which are sometimes referred to as short-time spectral analysis in speech processing.

Windowing is the process of multiplying a waveform of speech signal segment by a time window of given shape, to stress pre-defined characteristics of the signal. At this stage the signal has been framed into segments, each frame is multiplied with a window function w (n) with length N, where N is the length of the frame. To reduce the discontinuity of speech signal at the beginning and end of each frame, the signal should be tapered to zero or close to zero, and hence minimize the mismatch. Moreover, this can be arrived at by windowing each frame of the signal to increase the correlation of the Mel Frequency Cepstrum Coefficients (MFCC). To obtain good frequency resolution, a long window is desirable but the linguistic importance of some short transients makes a short window desirable and effective. The normal Compromise that is always available to settle for is the frame lengths of about 20 or 30 ms, with a frame spacing of 5 to 10 ms. The proper selection in the choice of window w(n) is a grade-off between different factors:

(i) The shape of the window may reduce differences, but it may increase signal shape alteration. The length is proportional to the frequency resolution and inversely proportional to the time resolution.

(ii) The signal overlap is proportional to the frame rate, but it is also proportional to the correlation of subsequent frames.

Where w(n) designates the window function. Types of common window functions used in FIR filter design for speech are

- Hamming Window
- Rectangular Window
- Hanning Window

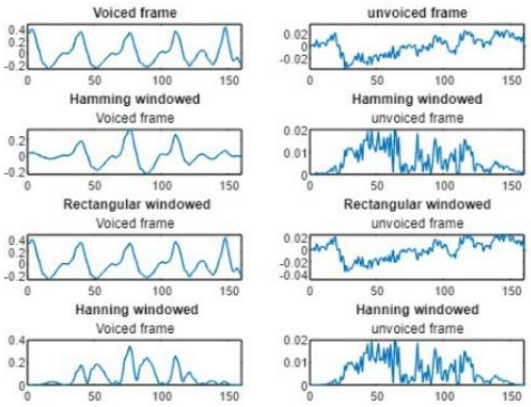

Fig.4Result of Framing and Windowing

Silence Removal is a pre-processing Technique used to remove silence(background noise) and unvoiced segments from the speech signals. Silence Removaland End point detection are the main part of many applications such as speaker andspeech Recognition.

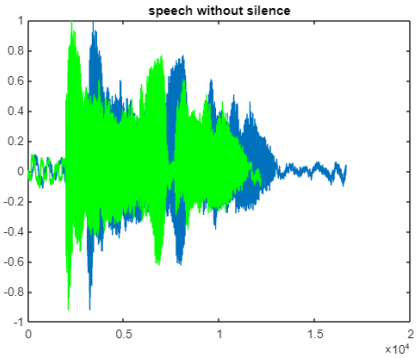

Fig.5 Result of Silence Removal

IV. SPEECH RECOGNITION

Speech features are extracted from recorded speech of a male or female speaker and compared with templates available in the database. Speech can be parameterized by Linear Predictive Codes (LPC), Perceptual Linear Prediction (PLP), Mel Frequency Cepstral Coefficients (MFCC). But in our project, Feature Extraction is done by using MFCC. Mel Frequency Cepstral Coefficients (MFCC) is one amongst the most normally used feature extraction methodology in speech recognition. The use of Mel Frequency Cepstral Coefficients can be considered as one of the standard methods for feature extraction.The MFCC feature extraction technique basically includes windowing the signal, applying the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the inverse DCT.

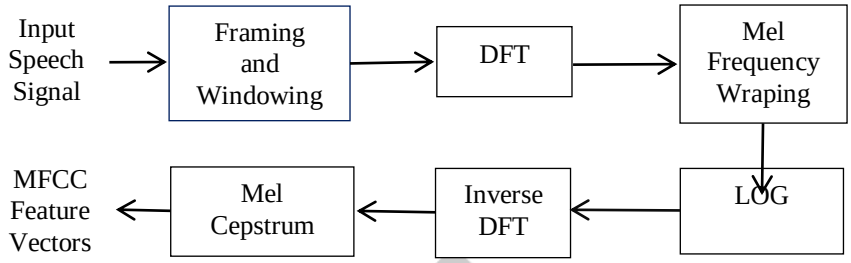

Fig 6: Block Diagram of MFCC

GMM, the abbreviation of Gaussian Mixture Model, which can be seen as a probability density function. The method of GMM is widely used in many fields, such as recognition, prediction, clustering analysis. The parameters of GMM are estimated from the training data by Expectation-Maximization (EM) algorithm or Maximum a Posteriori (MAP) estimation. Compared with many other model and method, the Gaussian Mixture Model has many advantages, hence it is widely used in the speech recognition field. And the Gaussian mixture model which is calculating the characteristic parameters space distribution by the weight sum of multi Gaussian density function when compared with the VQ method, from this point of view, the GMM has more accuracy and superiority. It is very hard to match the process of human's pronunciation organs, but we can simulate a model which express the process of sound, and it can be implemented by building a probability model for speech processing, while Gaussian mixture model is the very probability model which can qualified the condition.

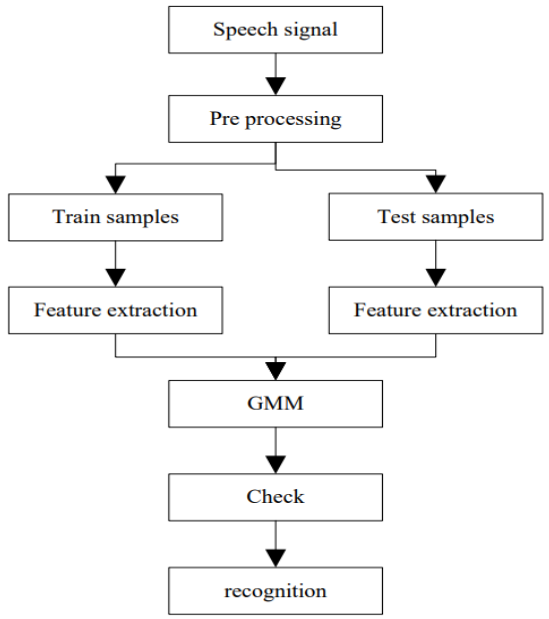

Fig 7: Flow Chart of Speech Recognition.

As far as the results you can see that the autocorrelation of the input signal matches the cross-correlation between the audio sample 3 and the input signal (sample 6) are the same, meaning that the input audio (sample 6) and the sample 3 are equivalents.

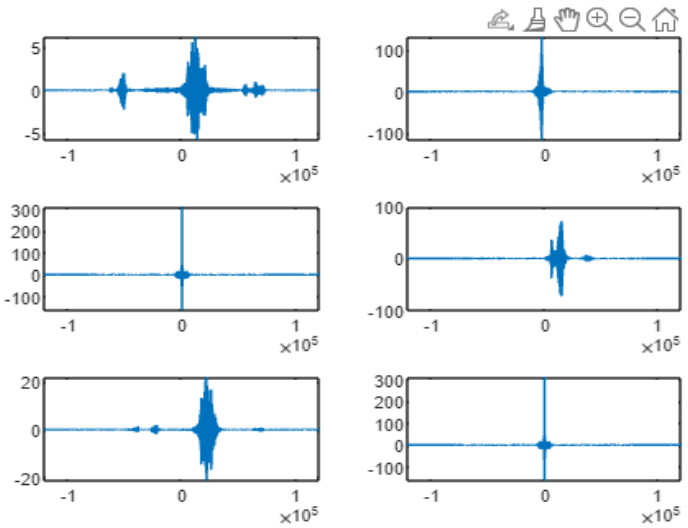

Fig 8: Result of Speech Recognition

V. Conclusion

The speech improving approach primarily based totally at the spectral subtraction set of rules is introduced. It may be visible from the experimental outcomes that the proposed approach correctly reduces heritage noise in assessment with the generally used spectral subtraction kind set of rules.The spectral subtraction technique seems to allow us to remove a fair amount of noise, including cancelling almost completely musical noise.Along with the spectral subtraction method, cepstral analysis has also been demonstrated over the same speech signal to determine important parameters of the speech which are pitch, pitch frequency, formant locations. These help to further processing of speech. This approach may be carried out in embedded structures associated with speech processing or communication- primarily based applications like speech to text conversion. The feature MFCC is extracted for recognition, then GMM training is applied for extracting μ, Σ, C is stored. In testing sessions voice print of unknown speaker is taken then silence removed and likelihood is found, corresponding to the maximum like hood person in database is identified.Experiments have been conducted on the database stored in the lab. And it has been observed that the system is accurate to a value of 90%.

VI. Future Scope

The Automatic speaker recognition system using MATLAB is an efficient program giving almost 90% of accuracy still there are chances to improve it. The main problem with the system is from the external noise. By using some other noise elimination methods, the performance of the system can be improved.

References

[1] Barry Commins “Signal Subspace Speech Enhancement with Adaptive Noise Estimation” National University of Ireland, Galway, September 2005.

[2] Furui, S., Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1981. 29(2): p. 254-272.

[3] Gulzar, T., et al., A systematic analysis of automatic speech recognition: an overview. Int. J. Curr. Eng. Technol, 2014. 4(3): p. 1664-1675.

[4]. J.W. Picone, Signal modeling techniques in speech recognition. Proc. IEEE 81, 1215–1247 (1993).

[5] Dr. Shaila D. Apte, “Speech Processing Applications”, in Speech and Audio Processing, Wiley India Edition.