Speech Analysis and Feature Extraction using SCILAB

Sunil Sharma, Naveen Jain, Isha Suwalka

Abstract— **With the advent of latest technologies, speech analysis has changed the potential in many security and confidential systems. Such systems are employed in automatic recognisation system who is speaking on the basis of individual information included in speech waves. This technique enhances the chances of identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers. Many systems have been made which costs a lot due to expensive software which checks the voice characteristics of the input utterance, using an automatic speaker recognition system. This paper depicts analysis side of the system implemented using freeware language SCILAB. The synthesis side includes speech production with the extraction of MFCC parameters employing articulation algorithm. The digital filters have been implemented for extraction feature .The signal processing application of SCILAB has become boon in comparison to MATLAB for all programmers.**

***Index Terms*— Speech analysis, Extraction feature, MFCC parameters, SCILAB, comparison.**

I. Introduction

With the increase in large number of industrial equipment used such as engines, a blower, fans, motors etc. The auditory perception system can be assumed in two major components: the peripheral auditory system (ears), and the auditory nervous system (brain). The basic structure of articulation algorithm describes the systems as text-independent speaker identification system since its task is to identify the person who speaks. At the highest level, all speaker recognition systems contain two main modules: feature extraction and feature matching. Feature extraction reveals extraction process involving small amount of data from the voice signal that can later be used to represent each speaker [2]. Feature matching involves the actual procedure to recognize the unknown speaker by comparing extracted features from his/h voice input with the ones from a set of known speakers. In this paper Mel scale parameters have been extracted using articulation algorithm taking a recorded wav format file. The SCILAB programming is used to implement this system for speech extraction feature. The following sub sections of the paper describe all the keywords and details of the methodology. The received acoustic pressure signal is processed by peripheral auditory system into two steps: initially transformed into a mechanical vibration pattern on the basilar membrane; finally, the auditory nervous system is responsible for extracting the perceptual information.

The auditory canal performs as acoustic resonator whose principal effect is to increase the ear’s sensitivity to sounds in the 3-4 KHz range. Speech analysis can be classified into identification and verification. The algorithm employed involves the process of determining registered speaker which provides a given utterance [1]. Another process involves accepting or rejecting the identity claim of a speaker.

II. Speech Analysis

The traditional solution to the speech system was only comparing recorded voice using perception of auditory system. Now through different user friendly softwares speech systems have been designed for different applications among which security system is the main [1]. The main issue to model for one speech generative model is the nonlinear character of the human hearing system. That is why psychoacoustic experimental works have been undertaken to find frequency scales that can model the natural response of the human perceptual system.

Since that moment, several different experiments have been carried out to investigate critical band phenomena and to estimate critical bandwidth. There are two outstanding classes of critical band scales: Bark frequency scale and Mel frequency scale. Mel-frequency scale has been widely used in modern speech recognition system. So we have used a technique to analyse the speech signal. In this paper we have high lightened the use SCILAB in articulation algorithm employing SCILAB script to serve our purpose. The algorithm we proposed is using different functions which use recorded voice as one input signal.

III. MFCC Extraction

The complete process is to extract the MFFC vectors from the speech signal. Here it is to be emphasized that the process of MFCC extraction is applied over each frame of speech signal independently.

Firstly with the pre-stage and the frame conversion ,the algorithm the followed with windowing stage, the MFCC vectors will be obtained from each speech frame. The process of coefficient extraction is explained below considering in

Manuscript received March 19, 2013.

Sunil Sharma, M.Tech (Pursuing), Digital Communication from Arya college of Engg. & Technology, Jaipur

Isha Suwalka , Asst. Prof. in Geetanjali Institute of technical Studies, Udaipur.

Naveen Jain is M.Tech. Scholar and pursuing his M.Tech. in Digital Communication from GITS, Udaipur.

any instant that all the stages are being applied over speech frames. The process is as follows:

A. MFCC extraction process involves computation of the Fast Fourier Transform (FFT) of each frame and obtain its magnitude[2]. This transform is a computationally efficient algorithm of the Discrete Fourier Transform (DFT) whose poweris of of two (K=2n), a faster algorithm is used, so a zero-padding to the nearest power of two within speech frame length is performed.

B. The next step is to adapt the frequency resolution to a perceptual frequency scale which satisfies the properties of the human ears , such as a perceptually mel-frequency scale. This issue corresponds to Mel filterbank stage. The filter-bank analysis consists of a set of bandpass filter whose bandwidths and spacings are roughly equal to those of critical bands and whose range of the centre frequencies covers the most important frequencies for speech perception [3].

The filterbank is basically a set of overlapped triangular bandpass filter, that according to mel-frequency scale, the centre frequencies of these filters are linear equally-spaced below 1 kHz and logarithmic equally-spaced above. The mel filterbank emphasizes these centre frequencies which correspond to mel centre frequencies uniformly spaced on mel-frequency domain. Thus, the input to the mel filterbank is the power spectrum of each frame, Xframe[k], such that for each frame a log-spectral-energy vector, Eframe[m], is obtained as output of the filterbank analysis. Such log-spectral-energy vector contains the energies at centre frequency of each filter. So, the filterbank samples the spectrum of the speech frame at its centre frequencies that conform the mel-frequency scale.

Let's define Hm[k] to be the transfer function of the filter m, the log-spectral energy at the output of each filter can be computed according to the equation [4]; where M is the number of mel filter bank channels. M can vary for different implementations from 24 to 40. The choice of the filterbank energies as input of filterbank analysis has been widely used in early recognition system. However, another approaches based on further transformations have been nowadays proposed to gain substantial advantages respect the filterbank energies input.

IV. SIGNAL PROCESSING SCILAB

Scilab is stated as scientific laboratory which basically software package for numerical computation for engineers and scientists. Scilab is endowed with powerful tools and easy syntax. Matrix being the basic fundamental object for calculation matrix manipulation can be easily handled. [8]
It is basically an interpreted language and is multiplatform so available on different OS such as Linux, Windows & MacOSX. The associated facilities in conjunction with open environment are helpful for signal processing, optimization & control It is easily compatible with Matlab so m-files can be easily used as it has Matlab to Scilab translator tool. So, also known as MATLAB CLONE. [8]. It provides visualization functionalities including 2D&3D graphics, contour and parametric plots and animations. Graphics can be easily exported in various formats like GIF, Xfig, LaTeX etc.[8] Scilab signal processing involves toolbox for handling signal processing function such as wavread etc .This is used for fourier transforms which has inbuilt function such as fft, dft etc. The transformation process is same as of MATLAB. The other toolboxes involved are in SCICOS which can also implement wavelet transform[9]. In this work different functions involving different stages of the algorithm have been implemented using script file. The recorded voice is read using wavread function and other functions have been made for the purpose.

V. ARTICULATION ALGORITHM

The algorithm involves speech feature extraction on mel scal which involves following stages :

A. Frame conversion:

In this step the continuous speech signal is blocked into frames with *N* samples, with adjacent frames being separated by *M* (*M* < *N*). The first frame consists of the first *N* samples and second frame begins with *M* samples after the first frame, which overlaps by *N* - *M* samples . The process continues till all the speech is accounted for within one or more frames. Typical values for *N* and *M* are *N* = 256 (which is equivalent to ~ 30 msec windowing and facilitate the fast radix-2 FFT) and *M* = 100.

B. Windowing

This involves the processing of window in each individual frame which can be hanning , hamming , rectangular etc. which minimizes the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame[4]. If we define the window as w(n),0<=n<=N-1, where *N* is the number of samples in each frame, then the result of windowing is the signal

y(n) = x(n) * w(n),0<=n<=N-1

In this work ,the *Hamming* window is used, which has the definition :

W(n) = 0.54 – 0.46cos((2*pi n)/N-1),0<=n<N-1

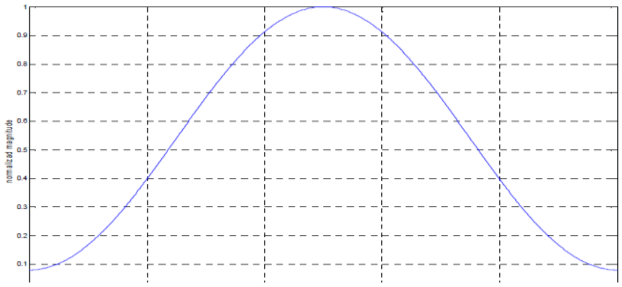

Fig.1 window function.

C. *Fast Fourier Transform (FFT)*

The next processing step is the Fast Fourier Transform, which converts each frame of N samples from the time domain into the frequency domain[4]. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT), which is defined on the set of N samples $\{x_n\}$, as follow:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \qquad k = 0,1,2,\ldots,N-1$$

[5]

In general X_k's are complex numbers and we only consider their absolute values (frequency magnitudes). The resulting sequence $\{X_k\}$ is interpreted as follow: positive frequencies 0<=f<=F_s/2 correspond to values 0<=n<=N/2-1, while negative frequencies -F_s/2<f<0 correspond to N/2+1<=n<N+1. Here, F_s denotes the sampling frequency.

D. *Mel-frequency covering:*

Since human perception of the frequency contents of sounds for speech signals does not follow a linear scale so for each tone with an actual frequency, *f*, measured in Hz, a subjective pitch is measured on a mel scale. The *mel-frequency* scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

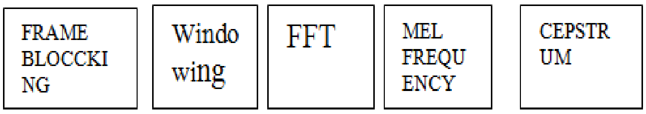

Fig-2 Flow chart for MFCC extraction

VI. Results

The speech utterances feature extraction is done using Mel Frequency Cepstral Coefficients (MFCCs) whose coefficients represents sound based on perception. It is consequent of the Fourier Transform or the Discrete Cosine Transform of the audio clip. The basic distinction between the FFT and the MFCC is the mel scale, the frequency bands are positioned logarithmically (on the mel scale) which approximates the human auditory system's response more closely than the linearly spaced frequency bands of FFT . The process is divided into diifferent blocks. In the frame blocking section, the speech waveform is more or less divided into frames of approximately 30 milliseconds[5]. The windowing block minimizes the discontinuities of the signal by tapering the beginning and end of each frame to zero.

The Fast Fourier block transforms each frame from the time domain to the frequency domain. In the Mel frequency wrapping block, the signal is plotted against the Mel-spectrum to mimic human hearing. As human hearing does not follow the linear scale so Mel-spectrum scale is involved which is a linear spacing below 1000 Hz and logarithmic scaling above 1000 Hz. In the final step, the Mel-spectrum plot is converted back to the time domain by using the following equation:

Mel (f) = 2595*log10 (1 + f /700)

The consequential matrices are called as the Mel-Frequency Cepstrum Coefficients [6].This spectrum provides a fairly simple but unique representation of the spectral properties of the voice signal which is the key for representing and recognizing the voice characteristics of the speaker.

A sample voice patterns may exhibit a certain degree of variance: same words, uttered by the identical speakers but at different laps of time, which may in yet different sequence of MFCC matrices. The purpose of speaker modelling is to build a model that can cope with speaker variation in feature space and to create a fairly unique representation of the speaker's characteristics.

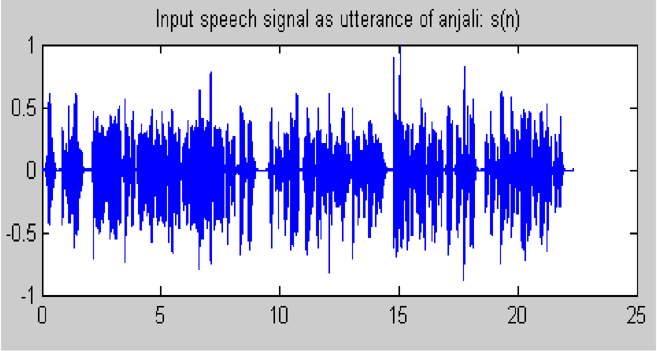

Fig-3 Input speech Signal

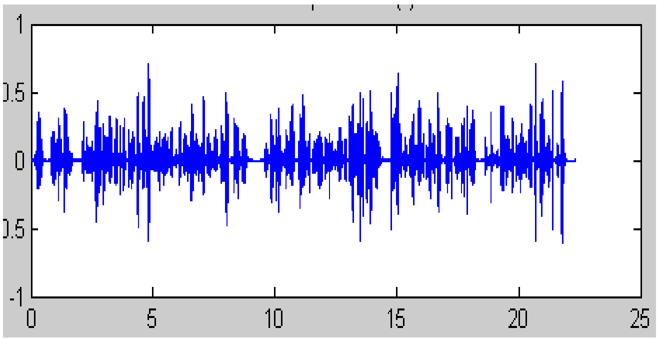

Fig-4 Pre stage analysis which involves filtering process

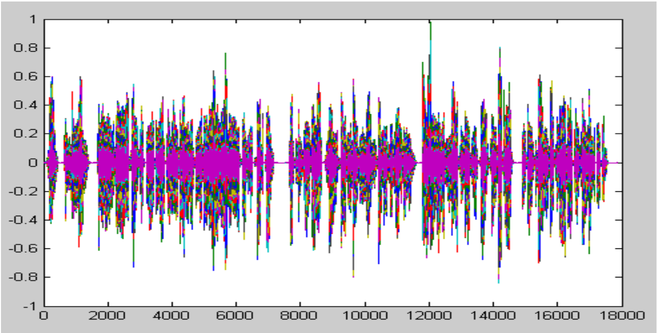

Fig-5 frame blocking output

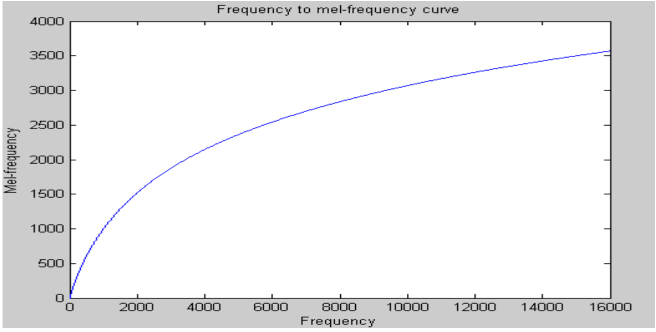

Fig-6 Frame blocking output

VII. CONCLUSION

This paper is compared with the speech system which has been constructed using the speech data and simple utterances recorded locally using SCILAB software. This paper involves feature extraction process using freeware for students so that they can achieve same efficient as that of MATLAB. The objective of this paper was to create a system which is efficient enough in comparision to different expensive softwares, and making speech analysis easier for an unknown speaker. By investigating the extracted features of the unknown speech and then compare them to the stored extracted features for each different speaker in order to identify the unknown speaker. The feature extraction is done by using MFCC (Mel Frequency Cepstral Coefficients).

VIII. FURTHER WORKS

The algorithm proposed in this paper presents a solution for implementation of a speech generative model; whereby the speech is synthesized and recovered from its MFCC representation. Synthesizing speech from parametric representations allows performing an investigation on the intelligibility of the synthesized speech as compared to natural speech. During the MFCC extraction process, much relevant information was lost due to reduction of the spectral resolution in the filterbank analysis and the next truncation into the MFCC components. However, that allowed recovering a smoothed spectral representation in which phonetically irrelevant detail had been removed. For that, the log mel power spectrum could be computed from its MFCCs by an inverse DCT. This mel power spectrum actually represented the envelope of the magnitude spectrum, where the harmonics appeared flattened [9].

In the generative model implementation labview can be used interfacing SCILAB for real time processing of speech. This can be used in many applications such as security systems, identification models etc.to implement the source-filter model for speech production; and in the other hand, to compute a spectral model that could be compared with the one derived directly from the original speech waveform [10]. Previously to the subjective evaluation of the generative model, the goodness of the synthesized speech was measured by computing the spectral distance between the original signal and the one produced from the MFCC coefficients. The two spectral models can be used which is obtained from the LPC coefficients computed from the original signal, and the one obtained from the LPC coefficients computed from the MFCC coefficients[9]. In this evaluation extraction can also be done using SCICOS for the minimum spectral distortion. A spectral distortion mean of the generative model was calculated using equations.

REFERENCES

[1] .L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, N.J., 1993.

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

[3] S.B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, *IEEE Transactions on Acoustics, Speech, Signal Processing*, Vol. ASSP-28, No. 4, August 1980.

[4] Y. Linde, A. Buzo & R. Gray, “An algorithm for vector quantizer design”, *IEEE Transactions on Communications*, Vol. 28, pp.84-95, 1980.

[5] F.K. Song, A.E. Rosenberg and B.H. Juang, “A vector quantization approach to speaker recognition”, *AT&T Technical Journal*, Vol. 66-2, pp. 14-26, March 1987.

[6] Emmanuel C. Ifeachor and Barrie W. Jervis, *Digital Signal Processing, A Practical Approach*, ISBN 0201-59619-9.

[7] Rulph Chassaing, *Digital Signal Processing with C and the TMS320C30*, ISBN 0-471-55780-3.

[8] Scilab manual.

[9] Scilab speech processing manual.

[10] Spectrum Rabiner L, Juang B. H, Fundamentals of speech recognition [4]Chap. 2,pp. 11-65, Pearson Education, First Indian Reprint, 2003.

Sunil Sharma is M.Tech. Scholar and pursuing his M.Tech. in Digital Communication from Arya college of Engg. & Technology, Jaipur affiliated to RTU Kota. His research area is Image processing and Image segmentation.

Isha Suwalka is Asst. Prof. in Geetanjali Institute of technical Studies, Udaipur and received her M.Tech in Digital Wireless Communication System from Suresh Gyan Vihar Universe, Jaipur. She has participated in various workshops on Tivoli, MEMS, MIC & DSP. She has huge interest in the field of Digital Signal Processing , Image processing ,Radar processing and its related softwares like Matlab , Scilab with its interfacing processors. She has already presented several papers in National and International Conferences of IEEE and also got the opportunity to act as Co-chairman in the sessions of National Conference of IEEE and IETE. Her 4 projects haven sponsored by government of Rajasthan through DST.She has recently published one paper in IEEE explore.

Naveen Jain is M.Tech. Scholar and pursuing his M.Tech. in Digital Communication from GITS, Udaipur affiliated to RTU Kota. His research area is Image processing.