**Paper ID: VESCOMM 16**

# IMPLEMENTATION OF SPEECH RECOGNITION SYSTEM

Mr. Maindargi L.C.[1]
1.PG student, Department of E&TC,
VVPIET, Solapur.

Prof. Mantri D.B.[2]
2. Assistant Professor, Department of E&TC,
VVPIET, Solapur

***Abstract*— The aim of this paper is to express the accuracy and time results of speech recognition (SR) system, based on Mel-Frequency Cepstral Coefficients (MFCC). The numbers of speech files were considered for the experimentation, MFCC were extracted and coefficients were statistically analyzed. Audio database is used for training and testing of the algorithm. Also Gaussian filters have been replaced with triangular filters to achieve higher level of accuracy. The speech files of about 2 second duration are to be given as input to training and testing unit. The results will be checked for a number of speech files.**

**The accuracy achieved by the proposed approach is expected higher than previous systems under study and can be implemented using Matlab as the programming tool.**

## I. INTRODUCTION

The speech signal of a person is unique and never changing. The signal taken as an input can be stored in template format and the stored templates can be compared with unknown or the input signal and exact match can be found out. The matched signal can be used for driving different mechanical machines, for machine activation in electronics or any industry and for determination of voice of unknown person for security purpose.

Speech recognition is the process of automatically recognizing the spoken words of person based on information in speech signal. Each spoken word is created using the phonetic combination of a set of vowel semivowel and consonant speech sound units. The most popular spectral based parameter used in recognition approach is the Mel Frequency Cepstral Coefficients called MFCC. MFCCs are coefficients, which represent audio, based on perception of human auditory systems. The basic difference between the operation of FFT/DCT and the MFCC is that in the MFCC, 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 or DCT.

Due to its advantage of less complexity in implementation of feature extraction algorithm, certain coefficients of MFCC corresponding to the Mel scale frequencies of speech Cepstrum are extracted from spoken word samples in database.

## II. LITERATURE REVIEW

Speech Recognition research has been ongoing for more than 80 years. Over that period there have been at least 4 generations of approaches, and a 5th generation is being formulated based on current research themes. To cover the complete history of speech recognition is beyond the scope of this paper.

By 2001, computer speech recognition had reached 80% accuracy and no further progress was reported till 2010. Speech recognition technology development began to edge back into the forrefront with one major event: the arrival of the "Google Voice Search app for the iPhone". In 2010, Google added "personalized recognition" to Voice Search on Android phones, so that the software could record users' voice searches and produce a more accurate speech model. The company also added Voice Search to its Chrome Browser in mid-2011. Like Google's Voice Search, Siri relies on cloud-based processing. It draws on its knowledge about the speaker to generate a contextual reply and responds to voice input. Parallel processing methods using combinations of HMMs and acoustic- phonetic approaches to detect and correct linguistic irregularities are used to increase recognition decision reliability and increase robustness for recognition of speech in noisy environment.

## III. METHODOLOGY

A sensor which makes acquisition of data and its subsequent sampling: in the specific case the sensor is a microphone, possibly with a high Signal to Ratio (SNR) value. Since the input signal is essentially speech, the sampling rate is usually set to 8 kHz. A step of preprocessing that in the voice context is constituted by the signal cleaning: simply de - noising algorithm can be applied to recorded data after a normalization procedure. In order to clean recorded speech signal from environmental additive noise, a spectral subtraction algorithm.

The extraction of the peculiar characteristics (feature extraction): in this stage Mel frequency cepstral coefficients are evaluated using a Mel filter bank after a transformation of the frequency axis in a logarithmic one. The generation of a specific template for each speaker: in this work we have decided to use the Gaussian Mixture Models (GMM) where model parameters are estimated with the maximum similarity making use of the Expectation and Maximization (EM) algorithm. In case of the user is registering (enrollment) for the first time to the system, this template will be added to the database, using some database programming techniques.

Otherwise, in case of test among users already present in the database, a comparison (matcher) determines which profile matches the generated template of the test speech. The matcher utilizes a similarity test, obtaining by a ratio value that can be accepted if it is higher than a decision threshold. The typical ASR system is shown in Figure 1. The technologies used for the development of the biometric system are the MFCC for the extraction of the characteristics and the GMM for the statistical analysis of the data obtained, for the templates generation and for the comparison.

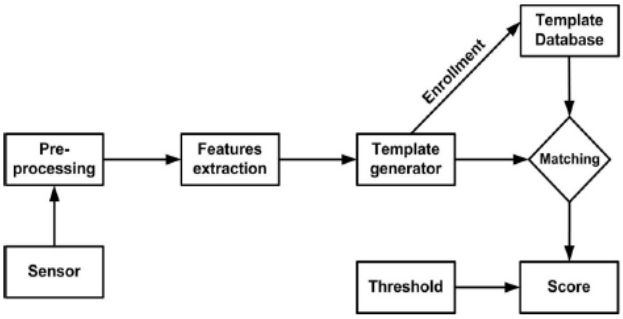


Fig.1 A Typical ASR System

### A. *Recognition Module*

Isolated word detection involves two digital signal processes which are Feature Extraction and Feature Matching. Feature extraction involves calculation of MFCCs for each frame. MFCCs are the coefficients that collectively represent the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear MEL scale of frequency. For feature matching DTW method is used.

### B. *Feature Extraction (MFCC)*

MFCC is chosen for the following reasons

1. MFCC is the most important features, which are required among various kinds of speech applications.
2. It gives high accuracy results for clean speech.
3. MFCC can be regarded as the "standard" features in speaker as well as speech recognition.
4. In the MFCC frequency bands are positioned logarithmically which approximates the human auditory systems response more closely than the linearly spaced frequency bands of FFT or DCT.

MFCC is based on human hearing perceptions which cannot perceive frequencies over 1KHz. Features obtained by MFCC algorithm are similar to known variation of the human cochlea's critical bandwidth with frequency. The process extracting MFCCs for a given voice sample is shown in Fig. 2

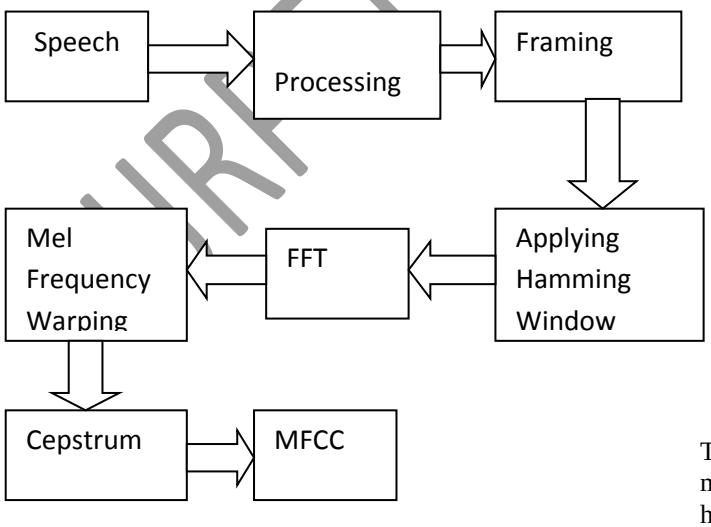


Fig.2 MFCC flow diagram

Pre-emphasis stage increases the magnitude of higher frequency with respect to lower frequencies. FIR filter, used for this purpose and its corresponding discrete output is given in equation 1 and equation 2 respectively.

$$F(z) = 1 - kz^{-1} \qquad 0 < k < 1 \qquad (1)$$

$$y[n] = s[n] - k.s[n-1] \qquad 0 < k < 1 \qquad (2)$$

Where, y[n] is the output and s[n] the signal input of the FIR filter. and frame length is 256. Now each frame is multiplied with hamming window. The Hamming window function is expressed in. Output of each frame after filtering is obtained as in

$$W[n] = 0.54 - 0.46 \cos \frac{2\pi n}{N - 1} \qquad (3)$$

$$Y[n] = X[n] \cdot W[n] \qquad (4)$$

Where, N = number of samples in each frame
Y[n] = Output signal
X (n) = input signal
$W[n]$ = $n^{th}$ coefficient of hamming window

Fast Fourier Transform (FFT) is applied to each frame which transforms signal to frequency domain. We would generally perform a 512 point FFT and keep only the first 257 coefficients. Thus the spectrum for each frame is obtained. But, it still contains lot of information not required for feature matching stage. The feature matching algorithm cannot discern the difference between two closely spaced frequencies. For this reason we take clumps of spectral bins and sum them up to get an idea of how much energy exists in various frequency regions. This can be performed by multiplying each frame with Triangular MEL Filter banks shown in Fig. 3.

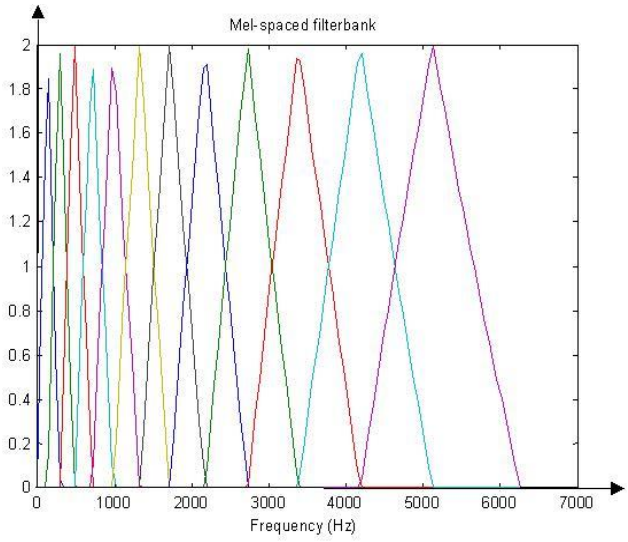


Fig. 3 Mel-spaced filter banks

The first filter is very narrow and gives us indication of how much energy exists near *zero* hertz. As the frequency gets higher our filters get wider as we become less concerned about variations. The equation for calculating MEL for a given frequency is shown in (5).

$$F(MEL) = 2595 \cdot \log_{10}[1 + f/700] \quad (5)$$

We are only interested in roughly how much energy occurs at each spot. Here a set of 26 triangular filters are taken. To calculate filter bank energies we multiply each filter bank with the energy spectrum, and then add up the coefficients. Once this is performed we are left with 26 numbers that give us an indication of how much energy was in each filter bank. Logarithm for these 26 energy values is taken following by Discrete Cosine Transform (DCT). DCT is calculated using equation shown in (6).

$$C_n = \sum_{k=1}^{k} (\log S_k)\, n\, k - \frac{1\Pi}{2\ K} \quad (6)$$

Where, n = 1, 2…K

$S_k$ = FFT coefficients

Value of K is taken to be 26. Thus we are left with 26 coefficients but, for feature matching; only the lower 12-13 of the 26 coefficients are kept. The resulting features (12 numbers for each frame) are called Mel Frequency Cepstral Coefficients. Thus the sample which is in frequency domain after applying FFT is converted back to time domain using MEL filter and DCT as shown in Fig. 4.


Fig.4. Steps in Converting from frequency domain to time domain

Then Delta and Delta-Delta coefficients are calculated for each frame. The first order derivative is called delta coefficient and the second order derivative is called delta-delta coefficient. The $n^{th}$ Delta feature and Delta-Delta feature is then defined by (7) and (8).

$$f_k[n] = f_{k+M}[n] - f_{k-M}[n] \quad (7)$$

$$^2 f_k[n] = f_{k+M}[n] - f_{k-M}[n] \quad (8)$$

Where, M typically is 2 -3 frames. The differentiation is done for each feature vector separately. Thus, for each frame we are left with 36 coefficients (12 MFCCs, 12 Delta, and 12 Delta-Delta).

### C. *Feature Matching (DTW)*

In this stage, the features of word calculated in previous step are compared with reference templates. DTW algorithm is implemented to calculate least distance between features of word uttered and reference templates. Corresponding to least value among calculated scores with each template, the word is detected. DTW finds the optimal alignment between two times series if one time series may be "warped" non-linearly by stretching or shrinking it along its time axis. The extent of matching between two time series is measured in terms of distance factor. Dynamic time wrapping for two voice samples is illustrated in Fig.5.

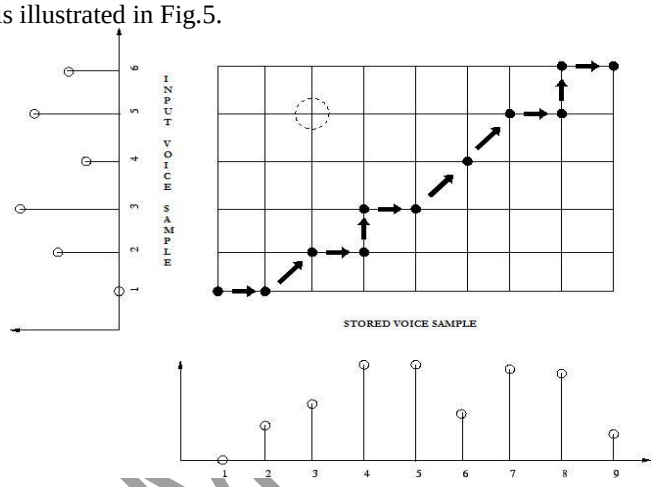


Fig. 5. Dynamic Time Wrapping of two voice samples

A matrix of order *n by m* is created whose *(i, j)* element is distance $d(a_i, b_j)$ between points $a_i$ and $b_j$ of two time sequences. Euclidean computation is used to measure distance between features of input sample and saved template. Then, distance is measured by (9).

$$D(i,j) = \min[D(i-1, j-1), D(i-1, j), D(i, j-1)] + d(i,j) \quad (9)$$

The template corresponding to least distance is the word detected.

## IV. CONCLUSION

This paper introduces a new method for speech recognition system by feature extraction method to improve the recognition accuracy and security. Many parameters can be used for speech recognition but MFCC appears to be the best method for the same. It is found to have a good recognition rate as well as to improve the accuracy of the system to implement it for number of applications. The work is planned to reveal the efficiency of MFCC for speech recognition purpose. Here we can plan for Gaussian filters instead of triangular to get the higher level of accuracy as all the elements of speech signals can be treated effectively without loss of information.

## *REFERENCES*

[1] P.K. Sharma, B.R. Lakshmikantha and K.S. Sundar, "Real Time Control of DC Motor Drive using Speech Recognition", Proceedings of the 2010 India International Conference on Power Electronics (IICPE), Jan. 28-30, 2011, New Delhi, India, pp. 1-5.

[2] MFCC retrieved on Jan 23rd, 2013 from, http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[3] Daryl Ning, "Developing an Isolated Word Recognition System in MATLAB", article retrieved from http://www.mathworks.in/company/newsletters/articles/developing-an-isolated-word-recognition-system-in-matlab.html.

[4] B.P. Das, R. Parekh, "Recognition of Isolated Words using Features based on LPC, MFCC, ZCR and STE, with Neural Network Classifiers", International Journal of Modern Engineering Research, Vol. 2, No. 3, June 2012, pp. 854-858.

[5] L. Rabiner, "A tutorial on Hidden Marcov Model and selected applications in Speech Recognition", Proceedings of the IEEE, Vol.77, No.2, 1989, pp. 257-286.

[6] L. Muda, M.Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, Vol. 2, No. 3, March 2010, pp. 138-143.

[7] A. Bala, Abhijit kumar, Nidhika Birla, "Voice Command Recognition System Based On MFCC And DTW", International Journal of Engineering Science and Technology, Vol. 2, No. 12, 2010, pp.7335-7342
.

[8] M.R. Hasan, M. Jamil, M.G. Rabbani and M.S. Rahman, "Speaker Identification Using MEL Frequency Cepstral Coefficient", Proceedings of 3rd International conference on Electrical and Computer Engineering (ICECE), December,28-30, 2004, Dhaka, Bangladesh, pp. 565-568.

[9] N.N.Lokhande , N.S.Nehe , P.S.Vikhe "MFCC Based Robust Features for English Word Recognition" IEEE International Conference, 978-1-4673-2272-0/12/2012.

[10] Nguyen Viet Cuong, Vu Dinh, Lam Si Tung Ho, "Mel frequency Cepstral Coefficients for Eye Movement Identification", IEEE 2012.

[11] Shivanker Dev Dhingra, Geeta Nijhawan , Poonam Pandit, "ISOLATED SPEECH RECOGNITION USING MFCC AND DTW" , International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Aug 2013.

[12]"Kernel-Based Feature Extraction with Speech Technology Application" Andras Kocsor and Laszlo Toth, *Associate Member, IEEE* Transactions On Signal Processing, Vol. 52, No. 8, August 2004.

[14] S. Chakroborty and G. Saha, "Improved Text-Independent Speech recognition Using Fused MFCC and IMFCC feature Sets Based on Gaussian Filter," *International Journal of Signal Processing*, Vol. 5, No. 1, 2009, pp. 11-19.

[14] Alfredo Maesa, Fabio Garzia, Michele Scarpiniti, Roberto Cusani, "Improved Text-Independent Speech recognition Using MFCC feature Sets & Gaussian Mixer Model," Journal of information security, 2012.