Feature extraction method - MFCC and GFCC used for Speaker Identification

Miss. Sarika S. Admuthe[1] Dr. Prakash H. Patil[2]
[1]G. S. M. COE, Balewadi, Savitribai Phule Pune University, India [2]Indira College of Engineering and management, Pune, Savitribai Phule Pune University, India

Abstract— To recognition the person by using human speech is main task of Speaker identification. In many organizations like bank, industry, government office, education field speaker identification is most useful biometric techniques to identify speaker. This paper is discuses about the extraction of feature method -MFCC and GFCC for speaker identification.

Key words: Feature Extraction, MFCC, GFCC

I. INTRODUCTION

Speaker identification identifies the speaker as one out of the group known to the system. The speaker identification is used for identifying a person who is suspected to be criminal, to grant the access to a person in restricted areas. Many organizations like banks, institutions, industries etc are currently using this technology for providing greater security to their vast databases. Speaker Identification systems can be either text dependent and text independent. Text dependent system will use the same test pharse whereas text independent systems have no restriction on the test pharse. The Speaker idenification mainly involves two modules namely feature extraction and feature matching. Feature extraction is the process that extracts a small amount of data from the speaker's voice signal that can later be used to represent that speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing the extracted features from his/her voice input with the ones that are already stored in speech database.The Speaker identification has two phase: Train Phase or Enrollenment Phase and Test Phase or Identification Phase as shown in figure 1 and figure2.

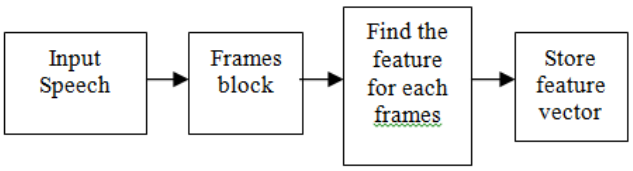


Fig. 1: Train phase or Enrollenment phase

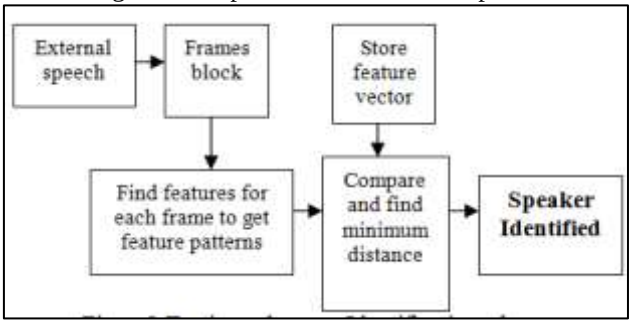


Fig. 2: Testing phase or Identification phase

II. FLOW CHART OF SPEAKER IDENTIFICATION

Speaker Recognition is a system that automatically recognizes speakers on the basis of individual spectral information presented in speech waves. This technique used the speaker's speech to verify their identity.Figure 3 shows the flow chart of speaker identification system[4].

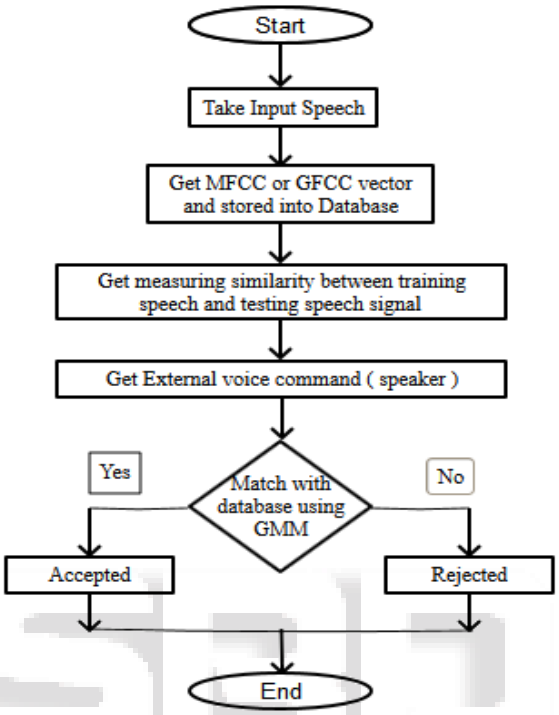


Fig. 3: Flow chart

III. PROPOSED SYSTEM

The goal is to perform Speaker identification in noisy environments. Auditory features- Gammatone Features, GFCC (Gammatone Frequency Cepstral coefficient) and MFCC (Mel frequency Cepstral coefficient) are used as feature extracting method for SID. For speaker modeling we will use here GMM (Gaussian Mixture Model). A Gaussian mixture density is a sum of weighted component densities.

IV. FEATURE EXTRACTION METHODOLOGY

Speaker Recognition mainly involves two modules namely feature extraction and feature matching. Feature extraction is the process that extracts a small amount of data from the speaker's voice signal that can later be used to represent that speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing the extracted features from his/her voice input with the ones that are already stored in our speech database.Speech feature extraction is the signal processing frontend which has purpose to converts the speech waveform into some useful parametric representation. These parameters are then used for further analysis in speaker identification system. Following feature extraction method are used for speaker identification-

1) Mel-Frequency Cepstral Coefficients (MFCC)
2) Gammatone Frequency Cepstral Coefficients (GFCC)

A. (MFCC) MEL Frequency Cepstral Coefficients:

The MFCC is the most evident Cepstral analysis based feature extraction technique for speech and speaker recognition tasks. It is popularly used because it approximates the human system response more closely than any other system as the frequency bands are positioned logarithmically.

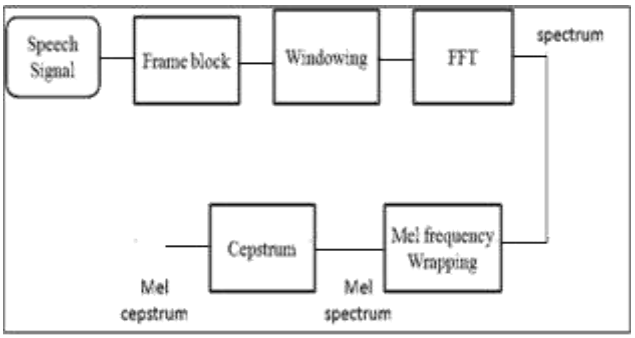


Fig. 4: Block diagram of MFCC

The calculation of the MFCC includes the following steps [3]:

1) Pre-emphasize input signal
2) Perform short-time Fourier analysis to get magnitude spectrum
3) Wrap the magnitude spectrum into Mel-spectrum
4) Take the log operation on the power spectrum (i.e. square of Mel-spectrum)
5) Apply the discrete cosine transform (DCT) on the log-Mel power spectrum to derive Cepstral features and perform Cepstral.

1) Frame Block:

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec. In this step, the continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M (M < N). The first frame consists of the first N samples. The second frame begins M samples after the first frame, and overlaps it by N – M samples. Similarly, the third frame begins 2M samples after the first frame (or M samples after the second frame) and overlaps it by N - 2M samples.

2) Windowing:

The next step in the processing is to window each individual frame so as to minimize 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. Typically the Hamming window is used which is given in equation 1,

$$w(n)=0.54 - 0.46 \cos(2\pi n/N-1) \quad 0< n < N-1 \quad \ldots\ldots 1$$

3) 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. The result after this step is often referred to as spectrum.

4) Mel-frequency Wrapping:

The frequencies range in FFT spectrum is very wide and human perception of the frequency contents of sounds for speech signals does not follow a linear scale. The each tone with an actual frequency is measured in Hz, and a pitch is measured on a scale called the 'Mel' scale. The Mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. After that equation 2 is used to compute the Mel for given frequency f in HZ.

$$\text{mel}(f) = 2595 * \log10(1+ f/700) \quad \ldots\ldots\ldots 2$$

5) Cepstrum:

Cepstrum is defined as the Fourier transform of the logarithm of the spectrum. It is the inverse Fourier transform of the logarithm of the power spectrum of a signal. the Cepstrum domain correspond to multiplication in the frequency domain and convolution in the time domain. This step can convert the log Mel spectrum (as shown in figure 5b) into time domain using DCT (Discrete cosine transform)[4]. The result after this step is referred as Mel Frequency Cepstrum Coefficient as shown in figure 5c.

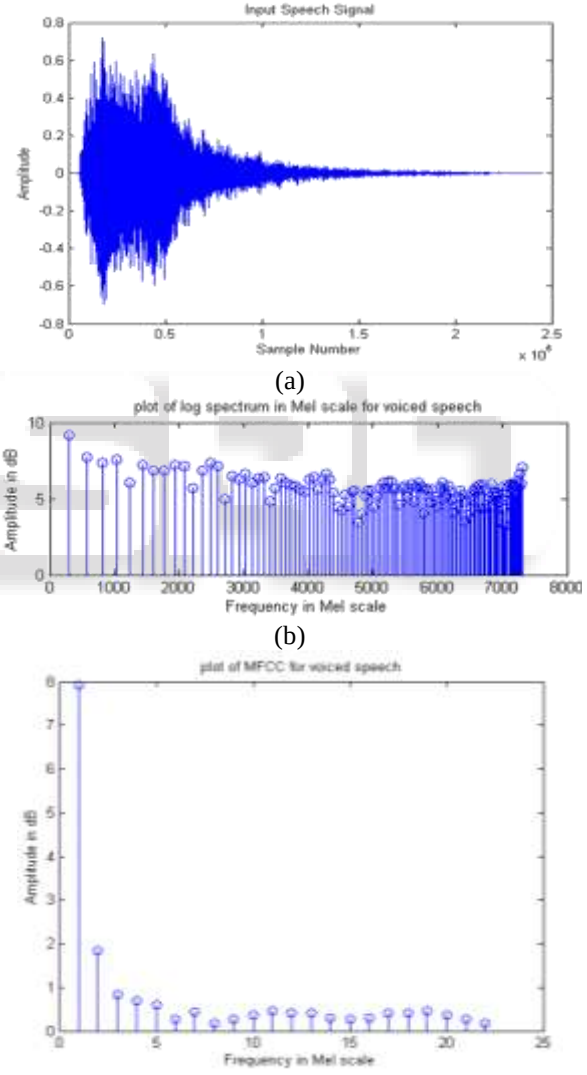


Fig. 5: Speaker1 top to bottom (a) Input Speech Signal (b) Log Mel Spectrum (c) Plot of 22 MFCC.

B. GFCC: Gammatone Frequency Cepstral Coefficients:

Gammatone filters are realized purely in the time domain. Specifically, the filters are applied directly on time series of speech signals by simple operations such as delay, summation and multiplication. This is quite different from the widely adopted frequency-domain design, where signals are transformed to frequency spectra first and the

gammatone filers then applied upon them. The time domain implementation avoids unnecessary approximation introduced by short-time spectral analysis, and saves a considerable proportion of computation involved in FFT.

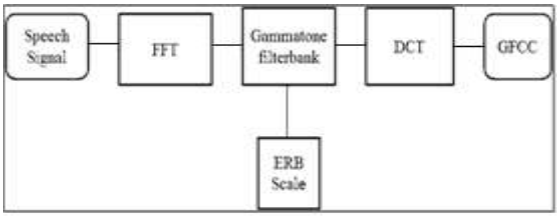


Fig. 6: Block diagram of MFCC

The calculation of the GFCC includes the following steps [3]:

1) Pass input signal through a 64-channel gammatone filter bank.
2) At each channel, fully rectify the filter response (i.e. take absolute value) and decimate it to 100 Hz as a way of time windowing.
3) Then take absolute value afterwards. This creates a time frequency (T-F) representation that is a variant of cochleagram.
4) Take cubic root on the T-F representation.
5) Apply DCT to derive Cepstral features.
6) Performance of System

1) Fast Fourier Transform (FFT):

After windowing the signal then FFT (fast Fourier transformation) is applied to each frame to analyze the spectrum.

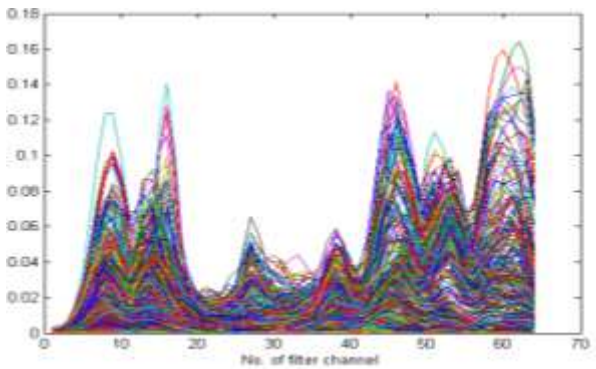
2) Gammatone Filter Bank:

Then Gammatone filter bank is applied to the each Fast Fourier transformed Signal. Gammatone filter-bank is a group of filters. The impulse response of a Gammatone filter is similar to the magnitude characteristics of a human auditory filter. The auditory filter's bandwidth is the value of ERB (Equivalent Rectangular Bandwidth) centered at frequency. The filter center frequencies (fc) and bandwidths are derived from the filter's equivalent rectangular bandwidth (ERB).The ERB of an auditory filter as shown in equation 3:

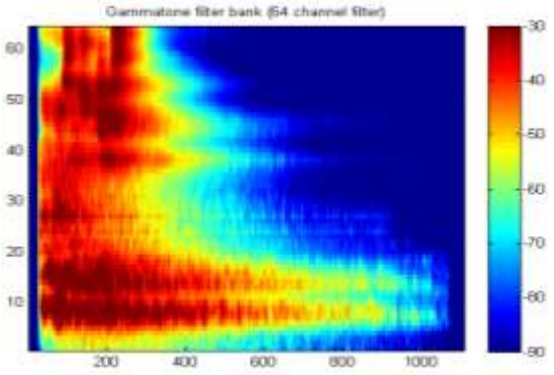
$$ERB(fc) = 24.7\,[(4.37\,fc / 1000) + 1] \quad \text{..... 3}$$

3) Discrete Cosine Transformation:

Discrete cosine transform are applied to model the human loudness perception and decorrelate the logarithmic compressed filter outputs. After Applying DCT we get the GFCC (Gammatone Frequency Cepstral Coefficient) features. The features are stored in the vectors. Theses vectors are stored in the Matlab Database.



(a)



(b)

Fig. 7: Speaker 1 top to bottom (a) 64 channel filter bank (b) spectrogram

Broadly speaking, there are two major differences between MFCC and GFCC. The obvious one is the frequency scale. GFCC, based on ERB scale, has finer resolution at low frequencies than MFCC (Mel scale). The other one is the nonlinear rectification step prior to the DCT. MFCC uses a log while GFCC uses a cubic root. In addition, the log operation transforms convolution between excitation source and vocal tract (filter) into addition in the spectral domain

V. Conclusion

MFCC is the commonly used algorithm for feature extraction of speech because MFCC has better success rate. Also GFCC is superior noise-robustness compared to other speaker features. By investigating these feature vectors along with the speaker identification techniques it was found that the GFCC features gave better results and accuracy for training and testing for all noise speech data.

Acknowledgement

We would like to thank the publishers, researchers and teachers for their guidance. We would also thank the college authority for providing the required infrastructure and support. Last but not the least we would like to extend a heartfelt gratitude to my family members for their support.

References

[1] Md. Moinuddin Arunkumar N. Kanthi "Speaker Identification based on GFCC using GMM", IJIRAE 2014.
[2] Dr. Shaila D. Apte, "Speech and audio processing", Wiley India, 2012.
[3] Xiaojia Zhao and Deliang Wang,"Analyzing Noise Robustness Of Mfcc And Gfcc Features In Speaker Identification", IEEE 2013, Page No - 7204-7208.
[4] Kshamamayee Dash , Debananda Padhi , Bhoomika Panda, Prof. Sanghamitra Mohanty"Speaker Identification using Mel Frequency Cepstral Coefficient and BPNN", Volume 2, Issue 4, April IJARCESE 2012,Page No- 326-332.
[5] Miss. Sarika S. Admuthe," Survey Paper on Automatic Speaker Recognition Systems", IJECS Volume 4, Issue3 March, 2015.

[6] R. W.Schafer and L. R. Rabiner, "Digital representations of speech signals", IEEE, vol. 63, pp. 662-677, Apr. 1975.
[7] Kishore Prahallad, Sudhakar Varanasi, Ranganatham Veluru, Bharat Krishna M, Debashish S Roy "Significance of Formants from Difference Spectrum for Speaker Identification", INTERSPEECH-2006, paper 1583-Tue1CaP.1.
[8] L. R. Rabiner and R.W. Schafer, "Digital Processing of speech signals" (sixth impression), Pearson Education and Dorling Kindersley Pvt. Ltd, 2011.
[9] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in Proc. ICASSP, 2012, pp.4257–4260
[10] A.Revathi1, R. Ganapathy and Y. Venkataramani , "Text Independent Speaker Recognition and Speaker Independent Speech Recognition Using Iterative Clustering Approach", International Journal of Computer science & Information Technology (IJCSIT), Vol 1, No 2, November 2009
[11] Tudor Barbu "A Supervised Text-Independent Speaker Recognition Approach", World Academy of Science, Engineering and Technology 33, 2007.

