USE OF MEL FREQUENCY CEPSTRAL COEFFICIENTS FOR THE IMPLEMENTATION OF A SPEAKER RECOGNITION SYSTEM

Rachel T. Pol, Dhanashri S. Sawant, Priyanka S. Sharma, Bhoomika V. Purohit, Pratibha R. Dumane
B.E Student, B.E Student, B.E Student, B.E Student, Assistant Professor
Electronics and Telecommunication Engineering
Don Bosco Institute of Technology, Mumbai, India

Abstract: The paper proposes a Speaker Recognition system which does the task of validating a user's claimed identity using characteristics extracted from their voices. It is one of the most useful and popular biometric recognition techniques in the world especially related to areas in which security is a major concern. A direct analysis and synthesizing of the complex voice signal is due to too much information contained in the signal. Therefore, the digital signal processes, Feature Extraction and Feature Matching were introduced to represent the voice signal. Mel- Frequency Cepstral Coefficients (MFCC) were extracted from the speech signal which were used to represent each speaker and recognition was carried out using weighted Euclidean distance. MATLABR2017b platform was used to implement feature extraction process.

Index Terms **– Co Feature matching, Feature Extraction, MFCC, Euclidean distance.**

I. INTRODUCTION

Speaker recognition is the process of recognizing the speaker from the database based on characteristics in the speech signal [1], [2]. Generally, speaker recognition can be classified into two processes speaker identification and speaker verification. The main difference between these two categories is that, the speaker verification performs a binary decision to verify the speaker's identity whereas speaker identification performs multiple decisions and it consists of the process of comparing the voice of the person speaking to a database or reference templates in an attempt to identify the speaker. Speaker identification can be further divided into two subcategories; text-dependent and text-independent speaker identification [5]. In text-dependent speaker identification, recognition is performed on a voiced instance of a particular word and in text-independent type, the speaker is free to utter any words for identification. The speaker recognition is being used for various popular applications which include automated dictation and command interfaces. Most of the speaker recognition systems contain two phases.

In the first phase the unique features from the voice signal are extracted which are used later for identifying the speaker. The second phase is feature matching in which the extracted voice features are compared with the database of known speakers [4]. The overall efficiency of the system depends on how efficiently the features of the voice are extracted and the procedures used to compare the real time voice sample features with the database. A general block diagram of speaker recognition system is shown in Fig. 1.

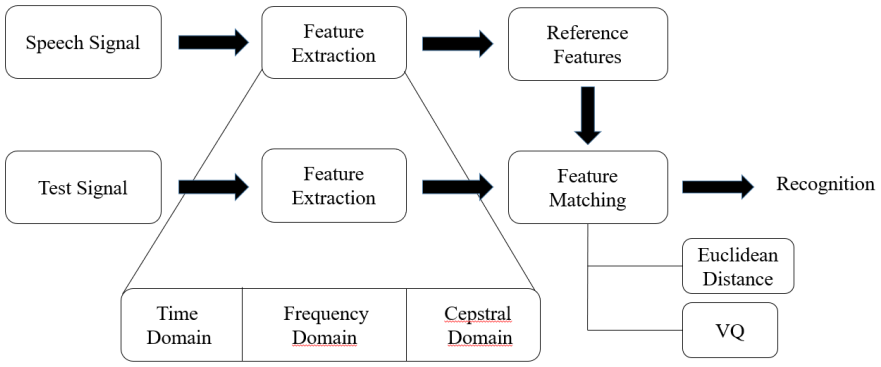

Fig. 1. Block Diagram of speaker recognition system

II. SPEECH RECOGNITION

A voice analysis is done after taking an input through microphone from a user. The design of the system involves manipulation of the input audio signal. At different levels, different operations are performed on the input signal such as Pre-emphasis, Framing, Windowing, Mel Cepstrum analysis and recognition (Matching) of the spoken word.

A. Feature Extraction using Mel Frequency Cepstral Coefficients (MFCC)

Feature Extraction module provides the acoustic feature vectors used to characterize the spectral properties of the time varying speech signal such that its output eases the work of recognition stage. Mel frequency cepstral coefficients (MFCC) is probably the best known and most widely used method for both speech and speaker recognition [9]. MFCC's are based on the known variation of the human ear's critical bandwidths with frequency.

MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. The block diagram showing the computation of MFCC is shown in Fig. 2.

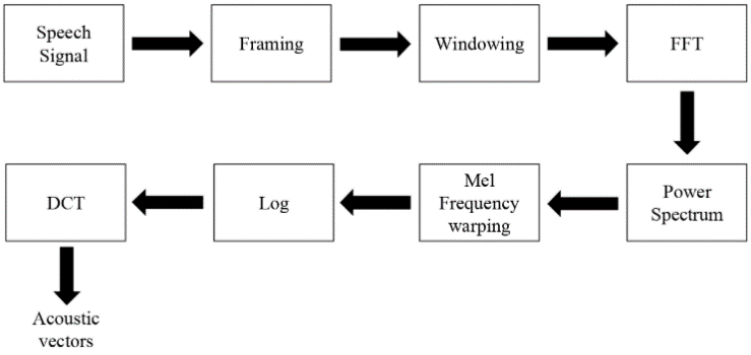

Fig. 2. Block diagram of MFCC

As shown in the block diagram of MFCC Fig. 2. MFCC consists of five main computational steps.

Step 1: Framing

This is 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 40ms and an overlap of 50% to 75% [4]. The voice signal is divided into frames of N samples. Adjacent frames are separated by M (M<N). Typical values used are M = 100 and N= 256

Step 2: Windowing

In this step windowing of each frame with a window function is done to minimize the discontinuities of the signal by tapering the beginning and end of each frame to zero. In this case, hamming window is used to perform windowing function.

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$

Step 3: Fast Fourier Transform

In this step each frame of N samples is converted from time domain into frequency domain. This enables in analyzing the localized frequency domain characteristics that are more useful for speaker recognition and other speech processing tasks. For analyzing the speech signals, normally the modulus square values of FFT are taken which gives the power spectrum [4]. This is motivated by the fact that the human auditory system does not perceive differences in the phase of speech signals. Moreover, the power spectrum is real valued, which facilitates the analysis. FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT) which is defined on the set of N samples {xn}, as follows:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-\frac{i2\pi kn}{n}}; k = 0, \cdots, N-1$$

Step 4: Mel Frequency Warping

Mel frequency warping is done to transfer the real frequency scale to human perceived frequency scale called the Mel-frequency scale [3]. The Mel scale is approximately linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz

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

where f denotes the real frequency and Mel(f) denotes the perceived frequency. The Mel frequency warping is normally realized by triangular filter banks with the center frequency of the filter normally evenly spaced on the frequency axis as shown in Fig.3.

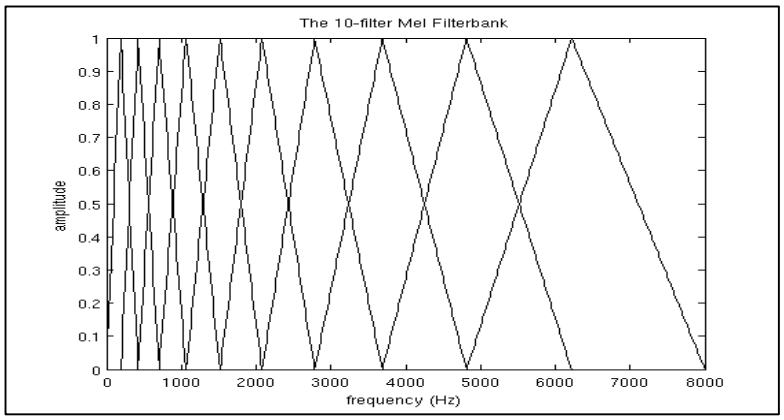

Fig.3. Triangular filter bank

Step 5: Discrete Cosine Transform

After the Mel frequency warping, log of the filter bank output is computed to attenuate the spectrum and finally the log Mel spectrum is converted into time domain using Discrete Cosine Transform (DCT) [3]. The resultant set of coefficients are called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vectors.

B. Feature Matching using Vector Quantization (VQ)

VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by a centroid [7, 8]. The collection of all code words is called a codebook.

One speaker can be discriminated from another based on the location of centroid Fig.4. shows a conceptual diagram to illustrate this recognition process. In the figure, only two speakers and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from the speaker 1 while the triangles are from speaker 2. In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The resultant code words (centroids) are shown by black circles and black triangles for speaker 1 and 2, respectively. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion [8, 9]. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the total VQ distortion is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified.

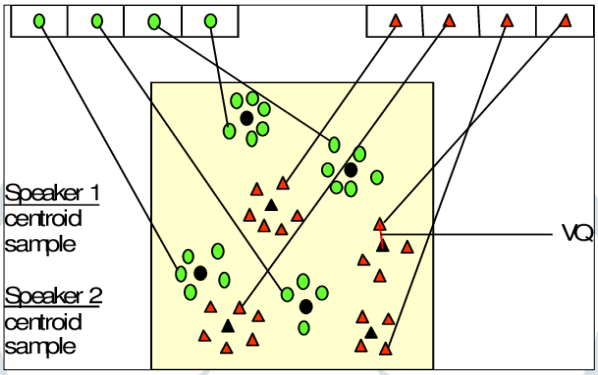

Fig.4 . Vector Quantization

C. Feature Matching using Euclidean Distance

In the speaker recognition phase, an unknown speaker's voice is represented by a sequence of feature vectors {x1, x2…. xi), and then it is compared with the codebooks from the database. In order to identify the unknown speaker, this can be done by measuring the distortion distance of two vector sets based on minimizing the Euclidean distance [6]. The formula used to calculate the Euclidean distance can be defined as following:

The Euclidean distance between two points P = (p1, p2…pn) and Q = (q1, q2...qn),

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.$$

The speaker with the lowest distortion distance is chosen to be identified as the unknown person.

III. Results And Discussion

Some snapshots at different stages of the processing of the speech signal are shown below.

Spoken words: Difficult roads often lead to beautiful destinations

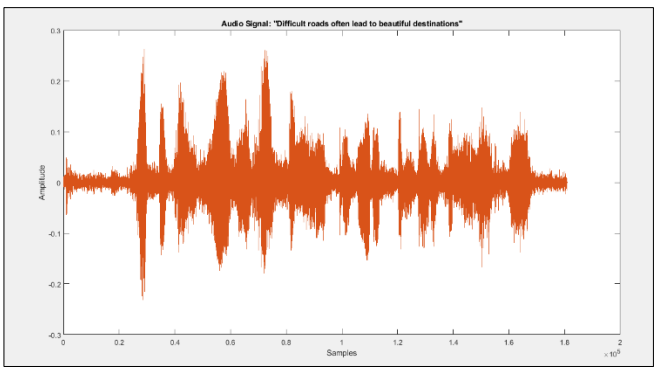

Fig. 5. Original Speech Signal plot

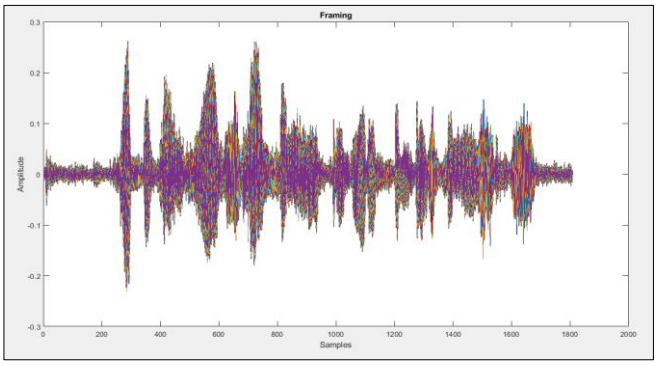

Fig. 6. Framing of the speech signal

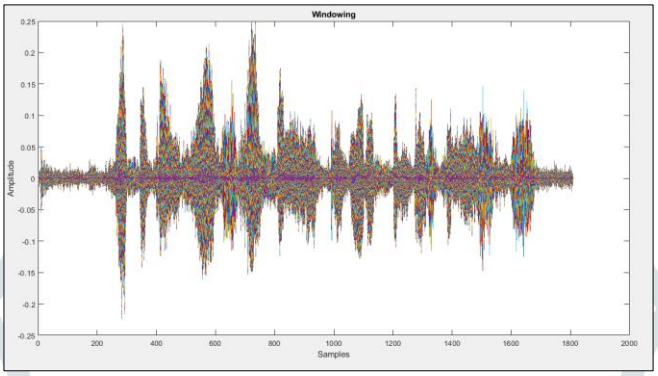

Fig. 7. Windowing

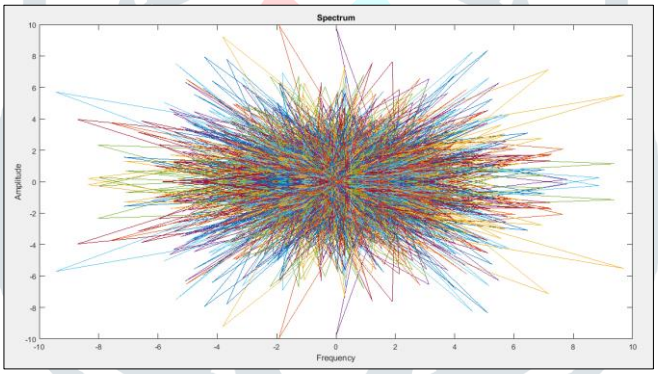

Fig. 8. Fast Fourier transform plot

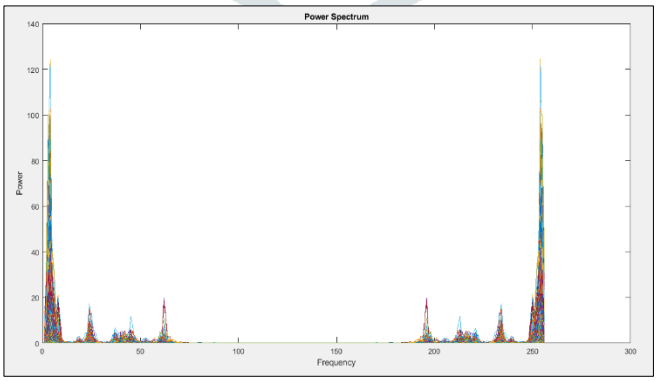

Fig. 9. Power Spectrum plot

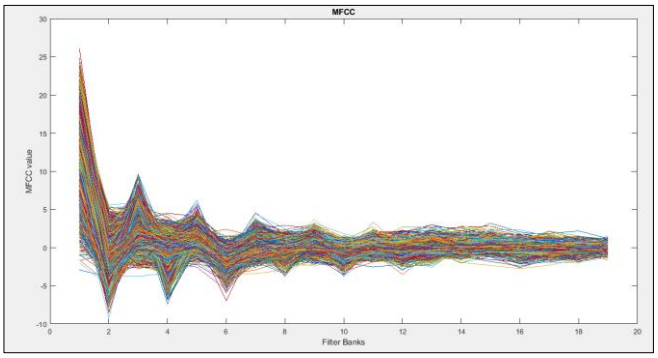

Fig. 10. Mel frequency cepstral coefficient plot

Columns 1801 through 1811

18.2482	16.8168	19.0848	20.5894	15.7622	18.5597	16.6210	15.8591	14.4385	16.5000
-2.7387	-0.8934	-1.0471	-2.7382	-2.8700	-3.8489	-1.6897	-1.7432	-1.2745	1.2984
2.6820	2.3362	0.7586	3.8100	-0.1488	1.5885	2.6354	2.9497	0.8015	-2.3275
0.3899	1.3315	-0.0811	0.8682	0.8198	0.1572	-0.0447	-1.4012	-0.8838	1.3626
1.7456	0.8032	0.1139	1.9835	0.3709	1.4480	0.8564	1.9926	0.8244	-0.7895
-2.0708	-1.5132	-2.2547	-1.6997	-1.3819	-1.7207	-1.0245	-2.2403	-2.1494	-1.1045
-0.2556	-0.3844	-0.7217	0.4924	-0.7086	0.3166	0.6935	-0.0303	-0.6602	-0.9483
-1.5538	-0.2930	0.0488	-0.7410	0.2402	-0.4754	-0.6866	-1.4596	-0.5489	0.2842
1.1936	-0.3193	-0.0952	0.4997	1.0872	-0.3611	0.7769	1.7168	-0.2749	0.0512
-1.3722	-0.8528	-0.1146	-2.0549	0.6513	-0.9166	-0.4069	-0.3468	-0.5215	-1.0963
0.3383	-0.4527	-0.4272	-0.1272	0.6044	-0.0389	0.2647	0.1375	0.6705	-0.2056
-1.5174	-0.4409	-0.1726	-0.6099	0.9569	-0.1873	-0.4360	-0.6339	0.8386	0.2452
-0.1760	-0.3568	0.1722	0.5551	1.5340	0.2929	0.3185	0.4773	0.0654	-0.0008
-0.6790	-0.1849	0.4511	-0.1425	1.4601	-0.0908	-0.4401	-0.6214	0.0842	-0.3253
0.9502	0.0899	0.3891	0.2948	1.0655	-0.0984	0.4793	0.7411	0.1620	0.4073
0.3868	-0.1459	-0.0685	-0.2842	0.9370	0.1129	0.3763	0.0901	0.1951	-0.0808
0.4132	0.0878	0.1706	0.0162	0.8047	-0.1698	-0.2127	-0.1160	0.2635	-0.4480
-0.2468	-0.2423	-0.0786	-0.5176	0.2841	-0.3401	-0.0929	-0.2214	0.5110	-0.3035
-0.1256	-0.3781	-0.0427	0.0302	0.2809	-0.3894	-0.1640	-0.1715	0.4699	0.3292

Fig. 11. MFCC Feature vectors

Feature vectors of a speech signal were extracted using MFCC on linearly spaced filters in Mel scale. Of the total MFCC feature vectors, 12 feature coefficients are used for further processing of the speech signal. The extracted features of the unknown speaker are then compared to the stored extracted features for each different speaker. Weighted Euclidean distance is further used in order to identify the unknown speaker.

REFERENCES

[1] S. A. Fattah, "A Time and Frequency Domain Formant Frequency Estimation Scheme for Noisy Speech Signals," *IEEE*, Dept. of Electrical and Computer Engineering Concordia University, Montreal, Quebec, Canada, 2009.

[2] Zilovic, M.S. Ramachandran, R.P. and Mammone, R.J "Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions". IEEE Transactions on Speech and Audio Processing, Volume 6, May 1998.

[3] Hongyu Xu, Xia Zhang, Liang Jia, "The Extraction and Simulation of Mel Frequency Cepstrum Speech Parameters," International Conference on Systems and Informatics (ICSAI 2012)

[4] Jose Krause Perin, Maria Frank, Neil Gallagher, "Speaker Recognition for Multi-Source Single-Channel Recordings," CS229 Final Paper, 2014.

[5] Fu Zhonghua; Zhao Rongchun; "An overview of modeling technology of speaker recognition", IEEE Proceedings of the International Conference on Neural Networks and Signal Processing Volume 2, Dec. 2003.

[6] Jon Gudnason and Mike Brookes "Voice source cepstrum coefficient for speaker identification" 2008 IEEE.

[7] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman, "Speaker identification using mel frequency cepstral coefficients" ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh.

[8] Sheeraz Memon, Margaret Lech and Ling He "Using Information theoretic vector quantization for inverted MFCC based speaker verification" IEEE CCECE/CCGEI, Saskatoon, May 2005.

[9] "1999 IEEE.Donghoon Hyun and Chulhee Lee "Optimization of mel-cepstrum for speech recognition