Speech Recognition and Verification using MFCC & VQ

Kashyap Patel, R.K. Prasad

Abstract- ***Speech recognition is very important branch in digital signal processing. Speaker Recognition software using MFCC (Mel Frequency Cepstral Co-efficient) and vector quantization has been designed, developed and tested satisfactorily for male and female voice. In this paper the ability of HPS (Harmonic Product Spectrum) algorithm and MFCC for gender and speaker recognition is explored. HPS algorithm can be used to find the pitch of the speaker which can be used to determine gender of the speaker. In this algorithm the speech signals for male and female ware recorded in .wav(dot wav) file at 8 KHz sampling rate and then modified. This modified wav file for speech signal was processed using MATLAB software for computing and plotting the autocorrelation of speech signal. The software reliably computes the pitch of male and female voice.***

The MFCC algorithm is used to simulate feature extraction module. Using this algorithm the cepstral co-efficient are calculated of Mel frequency scale. VQ (Vector Quantization) method will be used for reduction of amount of data to decrease computation time. In the feature matching stage Euclidean distance is applied as similarity criterion. Because of high accuracy of used algorithm the accuracy of voice command system is high. In this paper the quality and testing of speaker recognition and gender recognition system is completed and analysed.

Keywords: Autocorrelation, Signal, Voice command, Pitch, MFCC, Vector quantization, Euclidean distance.

I. INTRODUCTION

Speech processing is one of most important branches in digital signal processing. Speech signals can be used for speech recognition, speaker recognition or voice command recognition systems. The task of speaker identification is to determine the identity of a speaker by machine. To recognize voice, the voices must be familiar in case of human beings as well as machines. The second component of speaker identification is testing, namely the task of comparing an unidentified utterance to the training data and making the identification.

Depending upon the application the area of speaker recognition is divided into two parts. One is identification and other is verification. In speaker identification there are two types, one is text dependent and another is text independent. Speaker identification is divided into two components: feature extraction and feature classification. In speaker identification the speaker can be identified by his voice, where in case of speaker verification the speaker is verified using database.

Manuscript received on May, 2013.

Mr. Kashyap Patel, M-TECH Student of Electronics Engineering Department of Electronics, Bharati Vidyapeeth College of Engineering, Bharati Vidyapeeth Deemed University, Pune, India.

Dr. R.K. Prasad, Department of Electronics Engineering Department of Electronics, Bharati Vidyapeeth College of Engineering, Bharati Vidyapeeth Deemed University, Pune, India.

The Pitch is used for speaker identification. Pitch is nothing but fundamental frequency of a particular person. This is one of the important characteristic of human being, which differ from each other.

The speech signal is an acoustic sound pressure wave that originates by exiting of air from vocal tract and voluntary movement of anatomical structure. The schematic diagram of human speech production is as shown in figure 1.

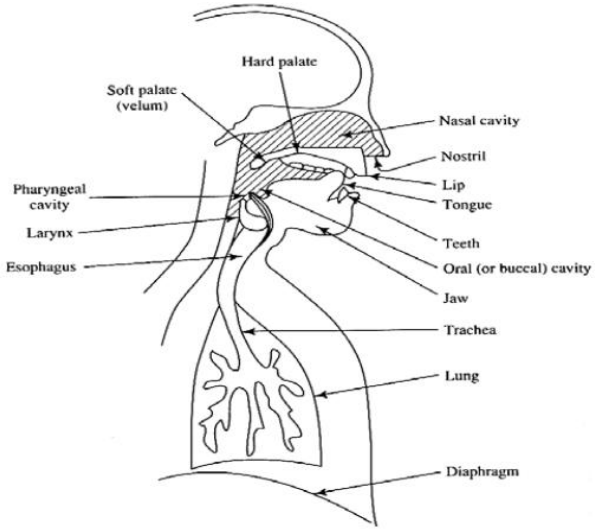

Fig. 1. Schematic diagram of human speech production mechanism

The components of this system are the lungs, trachea larynx (organ of voice production), pharyngeal cavity, oral cavity and nasal cavity. In technical discussion, the pharyngeal and oral cavities are usually called the "vocal tract". Therefore the vocal tract begins at the output of the larynx and terminates at the input of lips. Finer anatomical components critical to speech production.

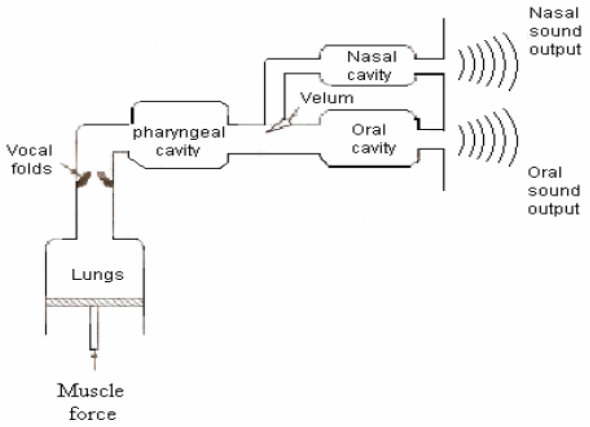

Fig. 2 Technical model for speech production

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

These components can move to different position to change the size and shape of vocal tract and produce various speech sound. The technical model of speech production is as shown in figure 2.

II. GENDER RECOGNITION

The block diagram of gender recognition system is as shown in figure 3.

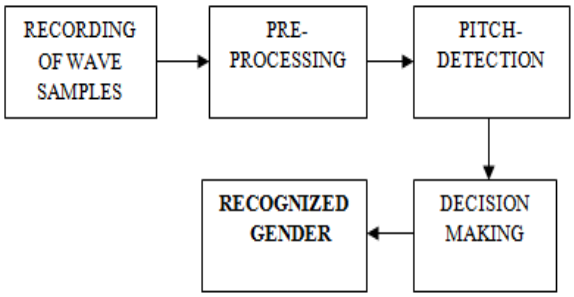

Fig. 3 Block diagram of Gender Recognition

The recording of speaker voice samples is done using Microsoft Sound Recorder (inbuilt software in windows operating system). Standard computer microphone is used for recording. The pre-processing includes noise removal, silence detection and removal and pre-emphasis. Pitch Detection is the main block for gender recognition. Pitch is nothing but the fundamental frequency of sound. The ANSI defines pitch at the attribute of auditory sensation of sounds. For detection of pitch the autocorrelation of speech signals for male and female voices has been computed and plotted using MATLAB software.

The speech signal and its corrologram for male and female voice is shown in figure 4. The female voice pitch computed from auto-corrologram is 431.5 Hz. Similarly the recorded speech signal and its auto-corrologram is shown in figure 5. From this figure the pitch of male voice is 250.3 Hz.

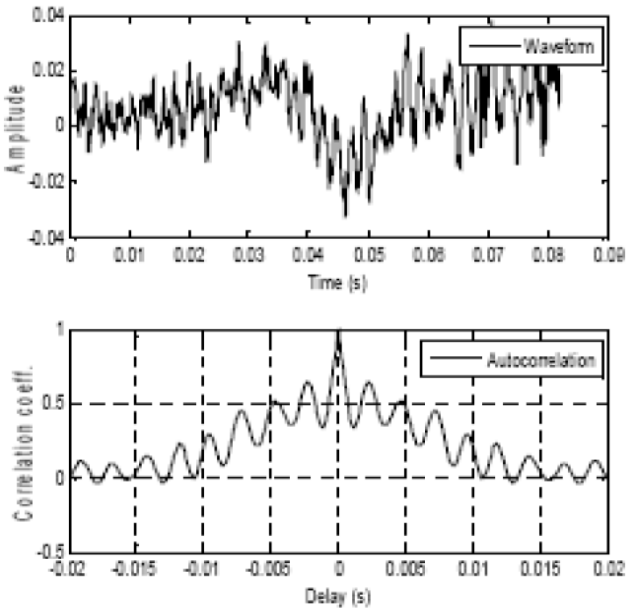

Fig. 4 Correlation coefficients for female voice

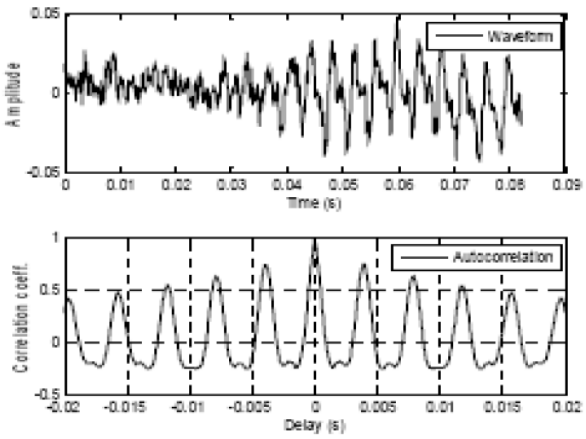

Fig. 5. Correlation coefficients for male voice

As the pitch of female speaker is higher than male speaker some threshold can be selected so as to discriminate male and female speaker. If the value of mean pitch is less than the threshold then the given speaker is male else if mean pitch value is greater than threshold then the given speaker is female.

III. SPEAKER IDENTIFICATION & VERIFICATION

The main aim The main aim of this project is speaker identification, which consists of comparing a speech signal from an unknown speaker to a database of known speaker. The system can recognize the speaker, which has been trained with a number of speakers. Figure 6 shows the fundamental formation of speaker identification and verification systems. Where the speaker identification is the process of determining which registered speaker provides a given speech. On the other hand, speaker verification is the process of rejecting or accepting the identity claim of a speaker. In most of the applications, voice is use as the key to confirm the identities of a speaker are classified as speaker verification.

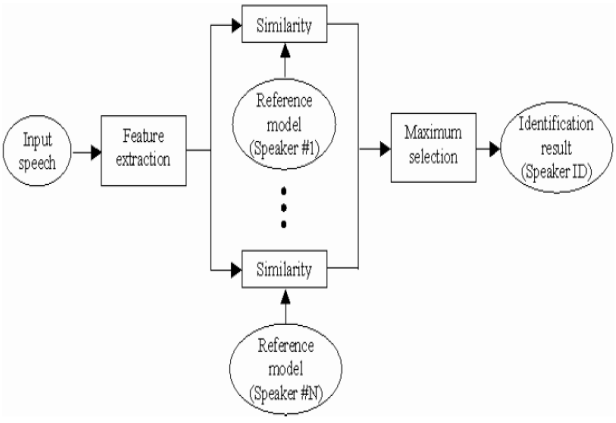

Fig. 6 Conceptual presentation of speaker identification

A. Mel Frequency Cepstrum Coefficient

In this project we are using the Mel Frequency Cepstral Coefficients (MFCC) technique to extract features from the speech signal and compare the unknown speaker with the exits speaker in the database. Figure 7 shows the complete pipeline of Mel Frequency Cepstral Coefficients.

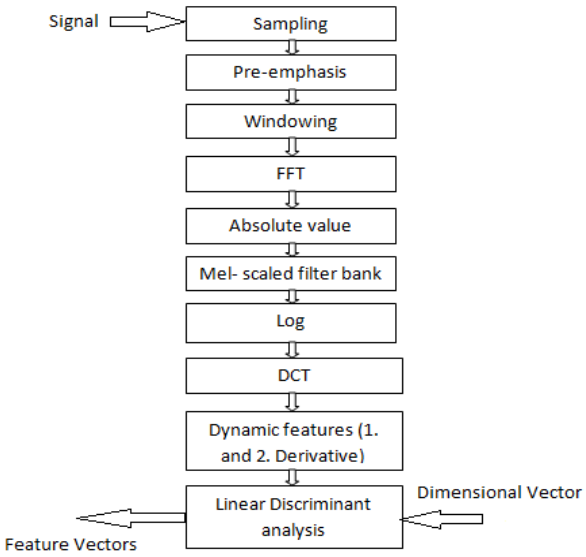

Fig. 7 Pipeline of MFCC

The Mel-frequency Cepstrum Coefficient (MFCC) technique is often used to create the fingerprint of the sound files. The MFCC are based on the known variation of the human ear's critical bandwidth frequencies with filters spaced linearly at low frequencies and logarithmically at high frequencies used to capture the important characteristics of speech. Studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, *f*, measured in Hz, a subjective 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. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels.
The following formula is used to compute the Mels for a particular frequency: $mel(f) = 2595*\log10(1+ f / 700)$. A block diagram of the MFCC processes is shown in Figure 8.

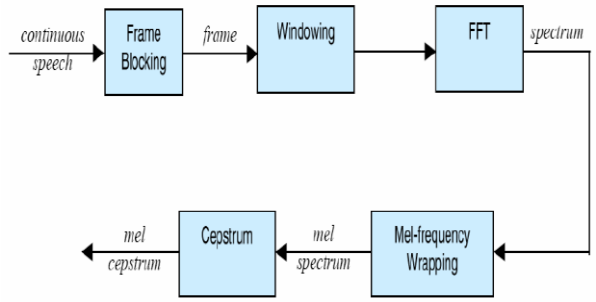

Fig. 8 Block diagram of MFCC

The speech waveform is cropped to remove silence or acoustical interference that may be present in the beginning or end of the sound file. The windowing block minimizes the discontinuities of the signal by tapering the beginning and end of each frame to zero. The FFT block converts 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. In the final step, the Cepstrum, the Mel-spectrum scale is converted back to standard frequency scale. This spectrum provides a good representation of the spectral properties of the signal which is key for representing and recognizing characteristics of the speaker.

After the fingerprint is created, we will also referred to as an acoustic vector. This vector will be stored as a reference in the database. When an unknown sound file is imported into MatLab, a fingerprint will be created of it also and its resultant vector will be compared against those in the database, again using the Euclidian distance technique, and a suitable match will be determined. This process is as referred to as feature matching.

B. Vector Quantization

A speaker recognition system must able to estimate probability distributions of the computed feature vectors. Storing every single vector that generate from the training mode is impossible, since these distributions are defined over a high-dimensional space. It is often easier to start by quantizing each feature vector to one of a relatively small number of template vectors, with a process called vector quantization. VQ is a process of taking a large set of feature vectors and producing a smaller set of measure vectors that represents the centroids of the distribution.
The technique of VQ consists of extracting a small number of representative feature vectors as an efficient means of characterizing the speaker specific features. By means of VQ, storing every single vector that we generate from the training is impossible.
By using these training data features are clustered to form a codebook for each speaker. In the recognition stage, the data from the tested speaker is compared to the codebook of each speaker and measure the difference. These differences are then use to make the recognition decision.

C. K-Means Algorithm

The K-means algorithm is a way to cluster the training vectors to get feature vectors. In this algorithm clustered the vectors based on attributes into k partitions. It use the k means of data generated from Gaussian distributions to cluster the vectors. The objective of the k-means is to minimize total intra-cluster variance, V.

$$V = \sum_{i=1}^{k} \sum_{j \in Si} |x_j - \mu_i|^2$$

where there are k clusters Si, i = 1,2,...,k and μi is the centroid or mean point of all the points, $x_j \in S_i$.
The process of k-means algorithm used least-squares partitioning method to divide the input vectors into k initial sets. It then calculates the mean point, or centroid, of each set. It constructs a new partition by associating each point with the closest centroid. Then the centroids are recalculated for the new clusters, and algorithm repeated until when the vectors no longer switch clusters or alternatively centroids are no longer changed.

D. Euclidean Distance

In the speaker recognition phase, an unknown speaker's voice is represented by a sequence of feature vector {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.

The Euclidean distance is the "ordinary" distance between the two points that one would measure with a ruler, which can be proven by repeated application of the Pythagorean Theorem. 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), is given by

$$\sqrt{(p1-q1)^2+(p2-q2)^2+\cdots+(pn-qn)^2}$$

$$=\sqrt{\sum_{i=1}^{n}(pi-qi)^2}$$

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

IV. EXPERIMENTAL RESULT

To implement proposed speaker recognition system, a system with some voice commands such as 'Hello' is considered.

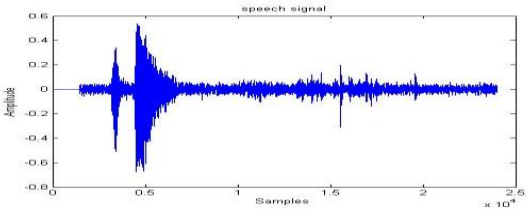

Fig. 9 original speech signal

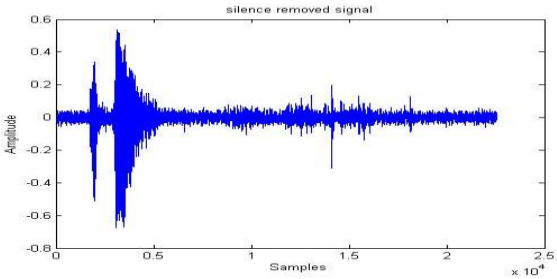

Fig. 10 Silence removal signal

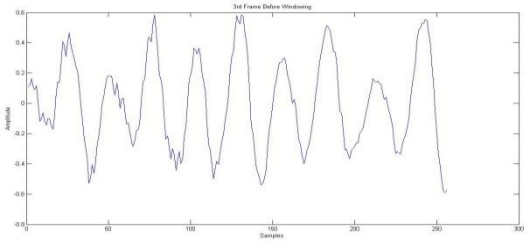

Fig. 11 Framing

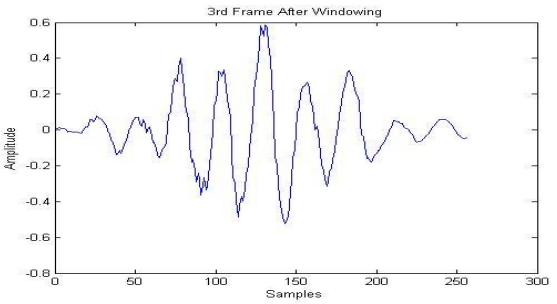

Fig. 12 Windowing

Training phase is done in two forms. First system was trained with one repetition for each command and once in each testing sessions.

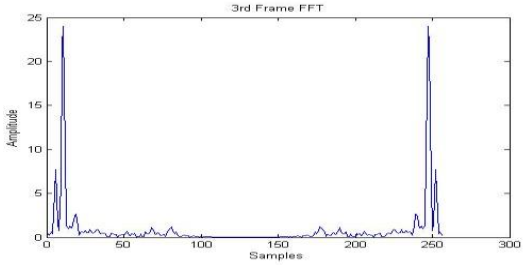

Fig. 13 Fast Fourier Transform

Table I. Pitch and frequency of different speech database

Sr no.	Speech Signal	Peak (P1)	P2	P3	P4	P5	Frequency In Hz (F1)	F2	F3	F4	F5
1	Speaker 1	102.9565	81.1906	71.5883	56.6103	47.3029	522	429	567	415	434
2	Speaker 2	95.4134	77.7759	70.3393	46.3746	44.5413	596	601	605	578	201
3	Speaker 3	73.3376	67.7199	57.7164	39.7373	28.3564	423	389	267	354	397
4	Speaker 4	63.9193	46.2722	33.1897	30.3625	22.5359	392	296	284	188	275
5	Speaker 5	60.6791	49.7088	33.3772	30.5015	30.3435	184	279	188	263	355

With this type of training error rate is about 13%. In second form, speaker repeated the words 4 times in a single training session, and then twice in each testing session. By doing this negligible error rate in recognition of commands is achieved.

V. CONCLUSION

The goal of this project was to create a gender and speaker recognition system, and apply it to a speech of 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). The function 'melcepst' is used to calculate the mel cepstrum of a signal. The speaker was modelled using Vector Quantization (VQ). A VQ codebook is generated by clustering the training feature vectors of each speaker and then stored in the speaker database. In this method, the K means algorithm is used to do the clustering. In the recognition stage, a distortion measure which based on the minimizing the Euclidean distance was used when matching an unknown speaker with the speaker database.

REFERENCES

1. Mahdi Shaneh and Azizollah Taheri, "Voice Command Recognition System Based on MFCC and VQ algorithms" World Academy of Science, Engineering and Technology 33 2009.
2. Ms. Arundhati S. Mehendale and Mrs. M.R. Dixit "Speaker Identification" Signals and Image processing: An International Journal (SIPIJ) Vol. 2, No. 2, June 2011.

3. Jamel Price, Sophomore student, Dr. Ali Eydgahi "Design of an Automatic Speech Recognition System Using MATLAB" Chesapeake Information Based Aeronautics Consortium August 2005.
4. E. Darren. Ellis "Design of a Speaker Recognition Code using MATLAB "Department of Computer and Electrical Engineering-University of Tennessee, Knoxville Tennessee 37996. 9th May 2001.
5. J.S Chitode, Anuradha S. Nigade " Throat Microphone Signals for Isolated Word Recognition Using LPC " International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 8, August 2012. ISSN: 2277 128X.

AUTHORS PROFILE

Mr. Kashyap Patel, M-TECH Student of Electronics Engineering Department of Electronics, Bharati Vidyapeeth College of Engineering, Bharati Vidyapeeth Deemed University, Pune, India.

Dr. R.K. Prasad, Department of Electronics Engineering Department of Electronics, Bharati Vidyapeeth College of Engineering, Bharati Vidyapeeth Deemed University, Pune, India.