SPEECH & SPEAKER RECOGNITION IN REAL TIME FOR SECURITY APPLICATION

[1]MANDAR NITIN KAKADE, [2]DR. D.B.SALUNKE,[3]DR.ANUPAMA DESHPANDE

[1]RESEARCH SCHOLAR, [2]PROFESSOR, [3]PROFESSOR
[1]DEPARTMENT OF ELECTRONICS ENGINEERING, [2]RESEARCH GUIDE, [3]RESEARCH GUIDE
[1]JJT UNIVERSITY, RAJASTHAN, INDIA

Abstract : In this paper, an automatic speech-speaker recognition system is implemented in real time noisy environment. The database creation with personalized voice in noisy environment is done with microphone arrangement. Various techniques in preprocessing step can be used to remove noise from sampled voice signal. Mel Frequency Cepstral Coefficient (MFCC) technique is used to extract Mel Cepstral Coefficients from each speech sample and thus database is created during training phase. For testing purpose, each input sampled speech signal is mapped with stored database using Vector Quantization (VQ) and Dynamic Time Warping (DTW) techniques. Output of mapped VQ is Speaker Recognition and output of mapped DTW is Speech Recognition. Using single sampled voice, real time Speech and Speaker can be recognized. This system is very useful for various applications such as Forensic, Banking where security is at highest priority. Therefore, nowadays for various security applications, automatic real time Speech and Speaker recognition systems in real time are on the verge of commercial success.

***IndexTerms* - MFCC, DTW, VQ, Speech, Speaker, Forensic**

1. INTRODUCTION

In the process of Automatic Speech-Speaker recognition, spoken words or sentences are converted into texts. This is very effective and simple way of communication between computer and human as it does not require devices like Keyboard, Mouse etc. So physically challenged people can also use this way of communication. Main aim of Automatic Speaker Recognition system is to accept or reject the claim of speaker after analyzing their speech sample with stored database. These systems were developed and tested under very low noise conditions so that output efficiency and recognition rate should be very high. Therefore automatic speaker recognition is used in Digital Forensic application.

The process of identification of speech samples can be closed set identification or open set identification. It is also Isolated word recognition or continuous word recognition. Identifying speech sample which is already present in database is called closed set identification. On the other hand, identification of speech sample which does not present in database is called open set identification. For isolated word recognition, presence of silence period in both sides of word is a must, whereas in continuous word recognition, it is difficult to identify silence period after utterance to recognize the spoken word. Automatic Speech-Speaker recognition system has many application areas including – Education sector, Medical sector, Military applications, Banking sector, Forensic applications, Security applications etc. For speech recognition system, various feature extraction techniques such are MFCC, LPC, RASTA and various feature matching techniques are DTW,VQ,HMM,GMM,ANN.

For speaker recognition, techniques are divided into Template modeling & Stochastic modeling which is further classified as Text dependent method and Text independent method. This is summarized as follows –

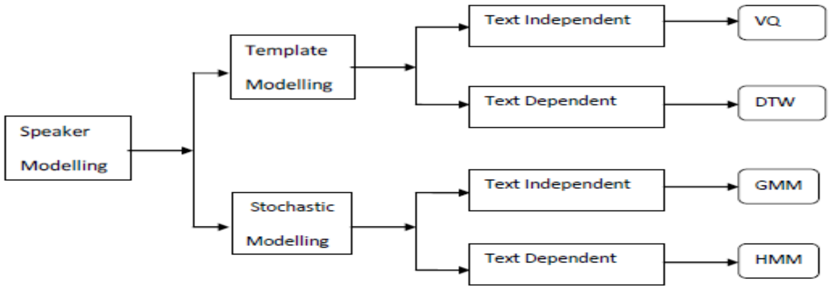


Fig: Speaker Modelling Techniques

Basic block diagram of Automatic Speech-Speaker recognition is as follows –

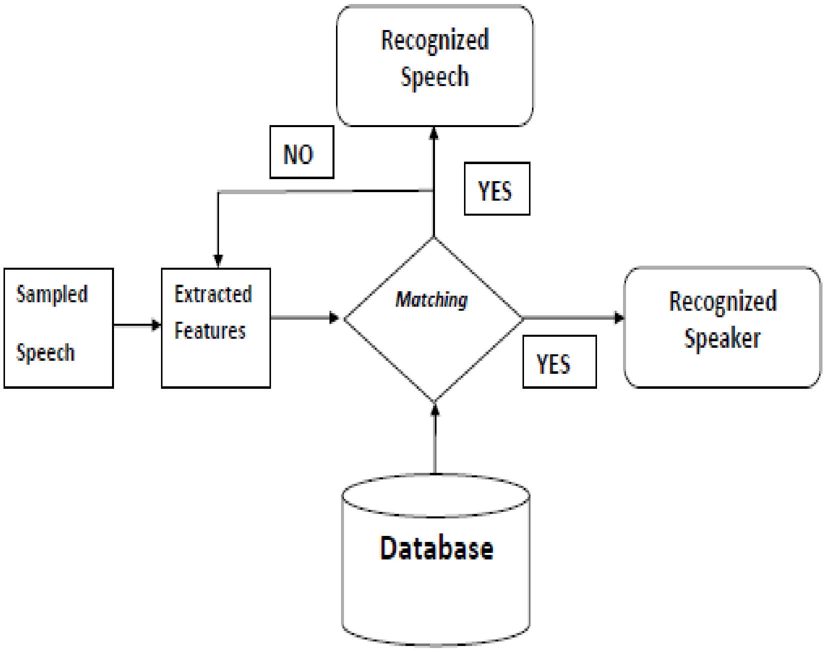
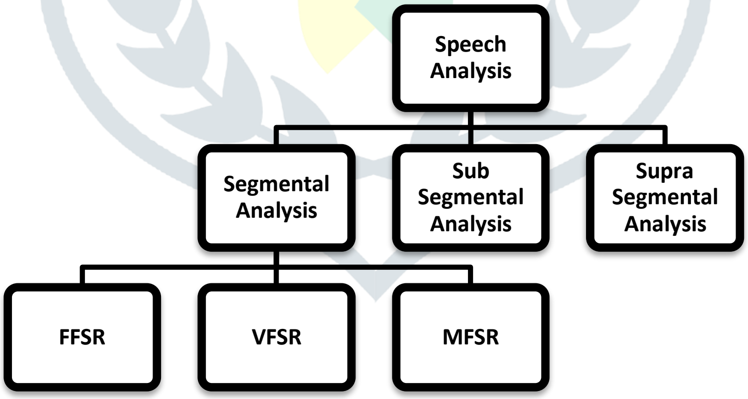


Fig: Block diagram of Automatic Speech-Speaker Recognition System

2. SPEECH FEATURES ANALYSIS TECHNIQUES

From speech, we can extract specific information of speaker. This specific information is due to vocal track & excitation source. Excitation source is Larynx and resonating structure is Vocal track. Speaker specific parameters are due to shape, size & dynamics of vocal track and excitation source. The information present in speech signal contains identity of speaker at various levels. This is summarized as follows –



i) In sub segmental analysis, duration of frame is 3-5msec. It is mainly used to extract information parameters of excitation source.
ii) In supra segmental analysis, duration of frame is 50-200msec. It is mainly used to extract parameters such as Pitch, Intensity, Duration and Speech quality.
iii) In segmental analysis, duration of frame is 20-30msec. It contain techniques such as Fixed Frame Sixe & Rate (FFSR), Variable Frame Sixe & Rate (VFSR), Multiple Frame Sixe & Rate (MFSR).
In FFSR, frame duration is 20-30msec and frame shift is 10msec. Its performance is better in noiseless clean signal and degrades as noise introduces. In VFSR, frame size and rate varies as a function of spectral characteristic of signal hence VFSR is having more advantages than FFSR. It is having better signal acoustic modelling and better performance without degradation in noisy environment.
In MFSR, same speech signal is analysed using different frame size and rate. This is beauty of this method hence it generates more feature vectors and gives better performance than FFSR and VFSR.
To increase the accuracy and efficiency of Automatic Speech-Speaker recognition system, following features are required –

a) Difference between feature vectors should be large between speakers but small within speaker.

b) Difficult to mimic/dummy voice
c) Independent on speaker's health
d) Occurs continuously within speech
e) Robustness against noise and environmental distortions

Various types of features are present for speaker recognition such as Spectral features, Dynamic features, Suprasegmental features, High level features, Dynamic features. This paper focuses on Spectral features which focuses on short term spectrum (MFCC).

3. ALGORITHMS & METHODOLOGY

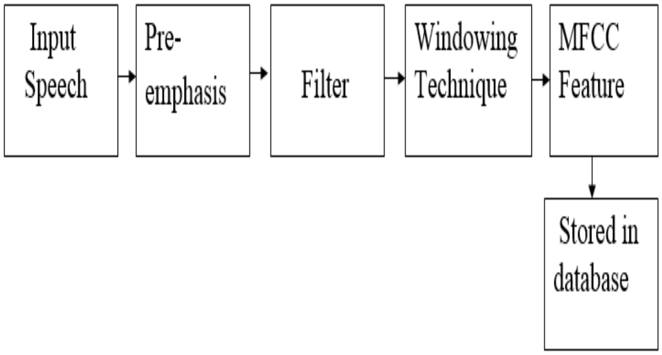


Fig: Training Phase

Figure shows steps in training phase of Automatic Speech and Speaker Recognition system.

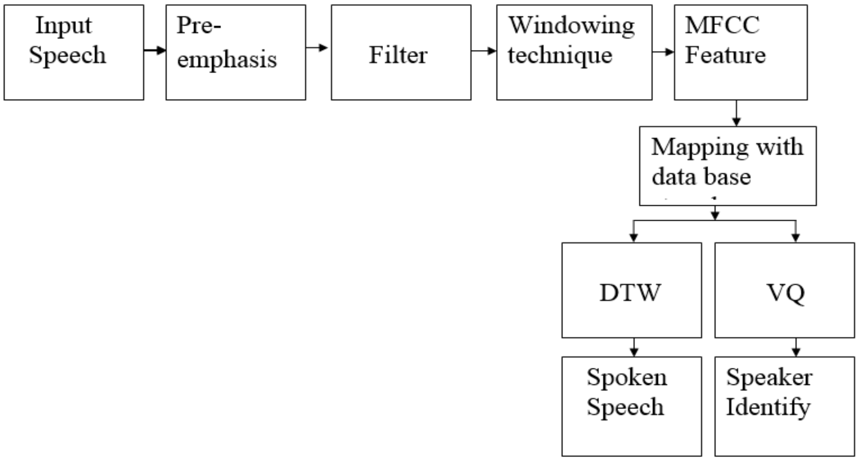


Fig: Testing Phase

Figure shows steps in testing phase of Automatic Speech and Speaker Recognition system.

INPUT SPEECH

The input speech given is the voice sample of a speaker or various speakers to carry out the speech and speaker recognition of the same.

PRE-EMPHASIS

The pre-emphasizing is nothing but boosting of the signal. The desired output signal is boosted in order to get an emphasized output. The low frequency which contains noise is discarded and high frequency signal which contains voice is boosted at a threshold level.

FILTERING

Filtering is done to filter out the noise of the ambient and surrounding. High pass filter is used in speech and speaker recognition, as all the desired data is in the high frequency band. The low frequency band is the noise, which is discarded.

FRAMING & WINDOWING

In framing, speech signal is divided into smaller frames of 20-30msec duration. To avoid aliasing effect and to minimize discontinuities, every frame is multiplied with window function. Various windowing techniques like Hamming, Hanning, Rectangular, Kaiser Window etc. Hamming window is being used because the other windows are less efficient. The rectangular window has got sharp cut-offs and also has got ringing effect.

MEL FREQUENCY WARPING

Mel scale is based on perception of human hearing which is linear below 1KHZ and logarithmic above 1KHZ. Therefore to map these variations, Triangular Mel filter bank which is having overlapping nature is used to produce MFCC coefficients.

MFCC

The MEL Frequency Cepstral Coefficients Features (MFCC) is the feature extraction technique. It is applied to get the coefficients of every speaker's voice samples differently. The graph used to plot such coefficients is called as MEL scale. The Mel scale relates perceived frequency, or pitch, pure tone to its actual measured frequency. Thus, the very first step for speaker recognition is to extract the features.

DTW

The Dynamic time wrapping (DTW) is the technique of comparing the frame wise sequence of vectors from the feature extraction with set of stored templates. It is a technique which "wraps" the time axis to detect the best match between the sequences. In DTW, entire utterance is divided into smaller frames and local distortion is calculated and overall decision is done based on these smaller distortions.

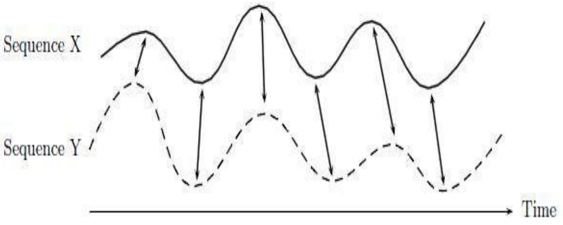


Fig: Dynamic Time Warping

VQ

Vector Quantization converts large set of feature vectors into smaller set of feature vectors which denotes centroid. Vector Quantization is the matching technique. In this, codebooks are generated by using Linde-buzo-gray (LBG) algorithm. In this algorithm, good set of codebooks are selected so that a good match can be done. The codebooks are nothing but the set of vectors which contains different voice samples of each speaker. Speaker specific code is generated containing feature vectors of single speaker. During testing phase, distortion in VQ is calculated over entire speech uttered and based on mapping, recognition decision is done.

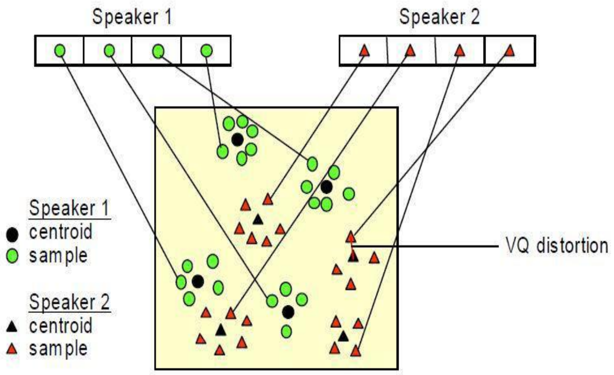


Fig: Codebook of Speaker

SPEAKER

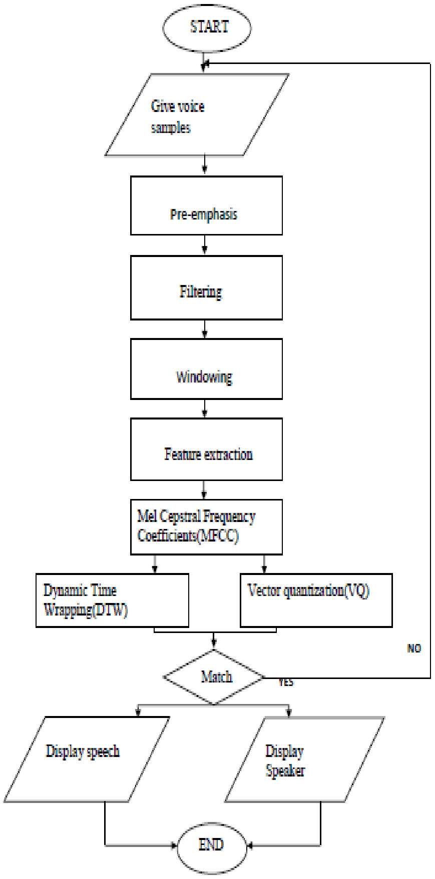
After all the above steps, the accurate speaker is recognized. Speaker recognizing means the identity who is speaking is displayed.

SPEECH

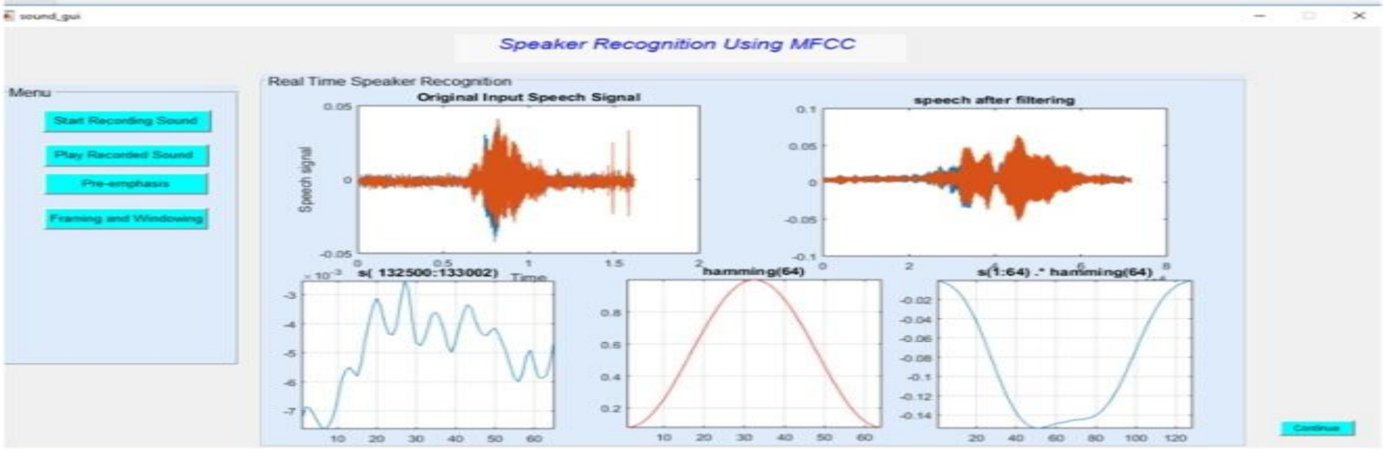
After all the above steps, the accurate speech is recognized. Speech recognizing means the identity what he/she is speaking is been displayed.

4. FLOW DIAGRAM OF AUTOMATIC SPEECH AND SPEAKER RECOGNITION SYSTEM

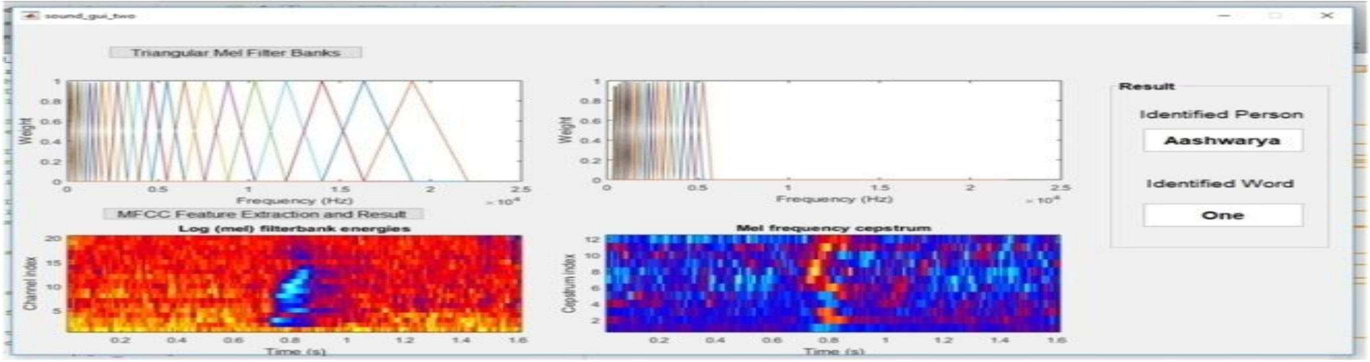
Figure indicates flow of speech sample from input speech up to final speech recognition as well as speaker recognition.



5. RESULTS & DISCUSSIONS



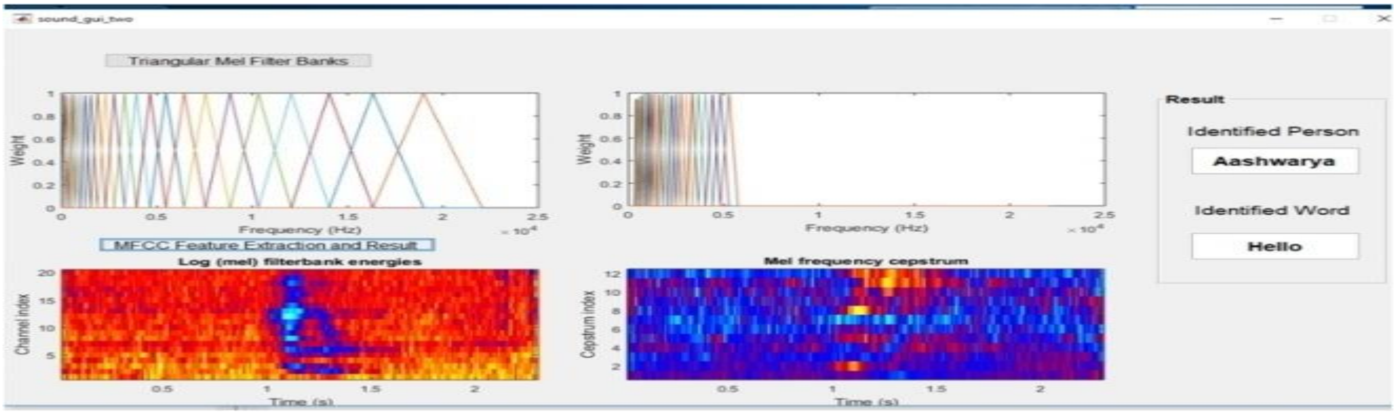
Above figure shows output windows of original speech, speech after filtering, hamming window.



Identified Person
Aashwarya

Identified Word
One

Above figure shows output windows of triangular filter bank, MFCC spectrum.



Identified Person: Aashwarya

Identified Word: Hello

Results of identified speech and identified speaker is displayed.

6. CONCLUSION

After analyzing all the information, we found that the audio signal can be represented as coefficients even if audio signal is a quasi-stationary. The main part of paper is the MFCC extraction technique. Using the MFCC, we try to represent the human voice in a better way. As the MFCC uses the MEL scale, the approximation to the human voice behaviour is good. The results obtained using MFCC and VQ are appreciable.The next method implemented was DTW. It has its own virtues of being very simple and astonishingly computation efficient. Instead of data sample, MFCCs of a test utterance were warped with respect to reference speaker and the least Euclidian distance was taken as basis for speaker identification. DTW doesn't take into account vocal tract information of a particular user. It only tries to align two vectors efficiently in time domain. Still its simplicity and easy hardware implementation has made it a regular tool for mobile applications. We can achieve 82% of accuracy by using MFCC which is more compared to other techniques. Thus, we can recognize the voice with more accuracy and make an useful application out of it.

7. REFERENCES

1) Prajakta P. Dahake, Kailash Shaw, Mrs. P. Malathi**,** "Speaker Dependent Speech Emotion Recognition using MFCC and Support Vector Machine", *(ICACDOT) International Institute of Information Technology (I²IT),* 2016Pune

2) Dany Ishac, Antoine Abche and Elie Karam,Georges Nassar , Dorothée Callens,**"**A Text-dependent Speaker-Recognition System**",** *direction of IEEE Instrumentation and Measurement Society prior to the acceptance and publication.*

3) J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, 1997.

4) Fang-Yie Leu, Guan-Liang Lin, **"**An MFCC-based Speaker Identification System**"** TungHai University*," IEEE 31st International Conference on Advanced Information Networking and Applications*,Taiwan2017.

5) Karthik Selvan, Aju Joseph, Anish Babu K. K,**"**Speaker Recognition System for Security Applications***",*** *IEEE Recent Advances in Intelligent Computational Systems (RAICS)*,2013.

6) Aseem Saxena, Amit Kumar Sinha, Shashank Chakrawarti, Surabhi Charu, **"**Speech Recognition Using Matlab,**"** *Suresh Gyan Vihar University, Jaipur, Rajasthan*, India Nov-2013.