Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI: 10.21474/IJAR01/4779
DOI URL: http://dx.doi.org/10.21474/IJAR01/4779
ISSN NO. 2320-5407

RESEARCH ARTICLE

DEVELOPMENT OF VOICE ACTIVATED SPEAKER RECOGNITION SYSTEM.

Chaitra K N[1], Anjan Kumar B S[2] and Dr .H N Suresh[3].
1. M.tech Student, Department of E&I, Bangalore Institute of Technology, Bengaluru.
2. Assistant professor, Department of E&I, Bangalore Institute of Technology, Bengaluru.
3. PG Coordinator, Department of E&I, Bangalore Institute of Technology, Bengaluru.

Manuscript Info

Manuscript History

Received: 8 May 2017
Final Accepted: 10 June 2017
Published: July 2017

Key words:-
Speaker Recognition, MFCC, speaker verification, feature matching, speaker identification, VQ.

Abstract

In digital signal processing techniques, the first step is the pattern recognition problem, which is essentially solved using recognition system based on speaker method. This method is based on the individual information of the utterer(speaker) stored in the form of speech waves and recognizes the speaker automatically based on the information available. It is important to process the uttered signal for fast and accurate speaker recognition system. It involves authentication of a speaker from a large ensemble of possible speakers. In this paper we implemented feature extraction of speech signal using Mel Frequency Cepstral analysis (MFCC) and the result of MFCC analysis are a series of vector characteristics, used to build Vector Quantization (VQ) codebook.

Introduction:-

Speech signals are represented as a sequence of sounds, which contains several information in it. Speech signal results in multiple variations arising at multiple levels: semantic, linguistic, articulatory and acoustic. Variation occurring due to these levels results in difference in the auditory properties of uttered signal. Utterer related fluctuations are seen because of anatomical variance inherent in the human vocal system and common individuals learned speaking routine. In speaker recognition [1], all the variations can be utilized to eliminate unwanted signals among the utterers. Speaker recognition entitles for a safe and secure way of verification of each utterer. Recognition of speaker is purely based on the individual speech alone. At the initial stage, the training models with respect to each speaker are built in the system. With the help of trained models and the characteristics of a given speech the authentication is done in the identification part, based on the identity unknown speaker.

This paper is designed to ease the communication barrier by helping the human to machine interface through speech. Recognition can be analyzed in two stages: training (verification) stage and testing (identification) stage. It is further branched into speaker authentication and speaker identification. Speaker authentication confirms the speaker's gender. Such a process is applied to the circumstances with the information or to the hampered areas and to the various forms of financial transactions. Speaker identification system decides the respective speaker with in the troupe of speaker's developed in a given speech articulation. Such systems have potential forensic applications. It is acknowledged that uttered signal depends on speaker feature that recognizes us to communicate with other person through telephone.

Corresponding Author:- Chaitra K N.
Address:- M.tech Student, Department of E&I, Bangalore Institute of Technology, Bengaluru.

In upcoming days, it is anticipated that the recognition system will make it possible to identify and authenticate the speakers for accessing the systems. It also make use of the system to eliminate manual control of operations by the use of speech, like financial transactions, exchange of data without knowing to the third party, hacking of confidential and also personal data. Although biometric techniques such as fingerprints and retinal scanning are important means of communication, and identification of data. Uttered signals can also implemented as a non-evasive biometric but it can be misused with or without the speakers intimation. Despite of this, the other types of authentication, like password or credentials, a speaker's speech cannot be misused, lost or misplaced. The purpose of recognition of speaker is to distil, designate and diagnose the knowledge of the individual speaker.

The rest of the concepts about the paper are catalogued as: section-II describes about the proposed method. Feature extraction is explained in Section- III, Section-IV describe about Vector Quantization and Section-V describes Simulation results carried out.

Proposed method for recognition of the System:-

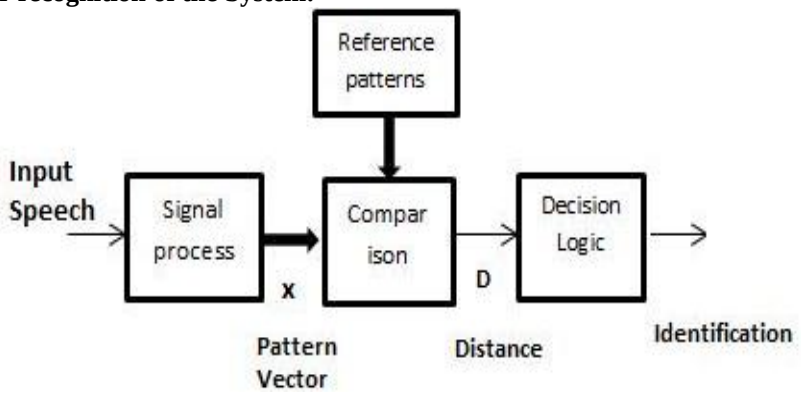

Figure.1:- General Representation of Recognition System.

The above figure shows that the representation of speech signal is obtained using digital speech processing which preserve the information about the uttered signal that are relevant to the identity of speaker. The obtained pattern is compared with the previous reference pattern and decision logic is made among the available alternatives. The two considerable subgroups of speaker recognition: are training (verification) phase and testing (identification) phase [3].

Speaker Verification:-

An identity is claimed by the user for speaker verification, and the decision required for the verification of system is strictly binary i.e to accept or reject the claimed individuality of the speaker.

Speaker Identification:-

The problem of speaker identification differs significantly from the speaker verification problem. In this case the system is required to make an absolute identification among the N speakers in the user population.

Feature Extraction:-

The recognition of speaker can be explained in two modules: one is the feature extraction and the other is the feature matching.

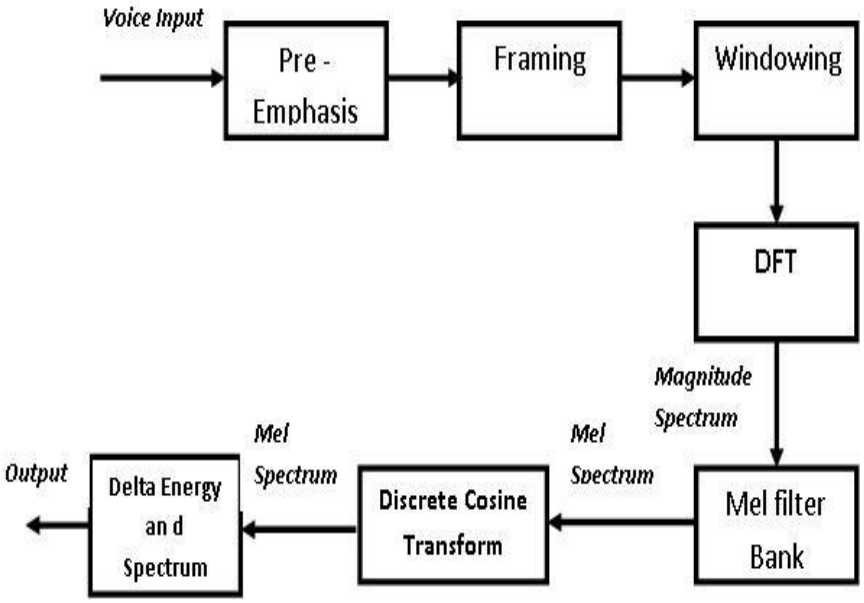

Figure.2:- Illustration of Feature Extraction method.

In this section feature extraction part is explained with the use MFCC coefficients. The other module is used to match the obtained feature vectors with the number of centroids in the Vector Quantization. Cepstral Coefficient (MFCC) are the most popular method for extracting the speech features from the given voice input. The implementing these coefficients depend on the text of the speaker[4] in the identification phase. With the use of K-means clustering[5] algorithm the feature vectors are measured with respect to centroids in VQ.

MFCC feature vectors are calculated in both the phases i.e. in training as well as in testing stage. Eucledian distance between the two Cepstral features are computed with the use of Cepstrum[2] similarity between them. The identification is performed satisfactorily by the code which is developed in the MATLAB.

Extracting speech Features:-

The speech signals are continuous in nature, in order to know the spectral characteristics of the signal it should be converted into some parametric representation. The characteristics and behavior of speech signals varies accordingly with time. For a short period of time (10-30msec) the variations are found to be stationary. In order to determine the certain variation in the signal, it should be determined in the long period of time. Hence, we use short-time spectral analysis for characterizing the speech signal.

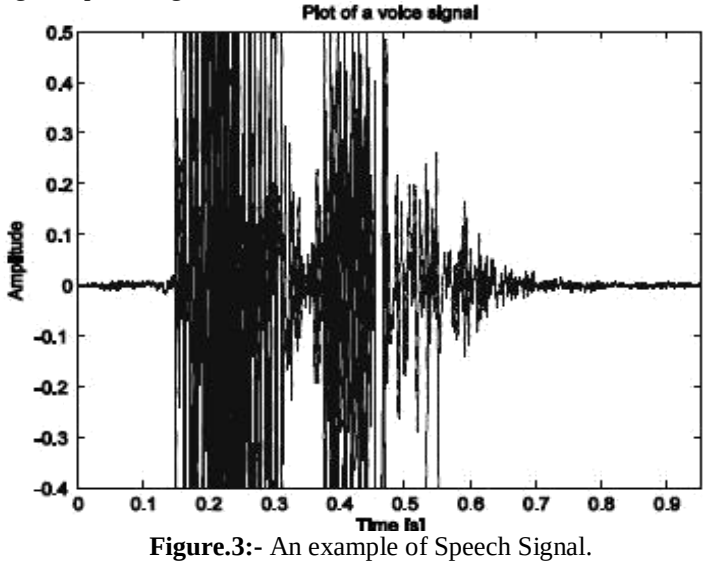

Figure.3:- An example of Speech Signal.

MFCC Extraction Process:-
MFCC process is most popular method used for the extraction of feature vectors from the uttered speech. The scale used to extract the coefficients is the Mel-Scale, which is expressed in the linear frequency and logarithmic spacing of below and above 1 kHz respectively. The plot of Mel-Scale is shown in Figure 4.

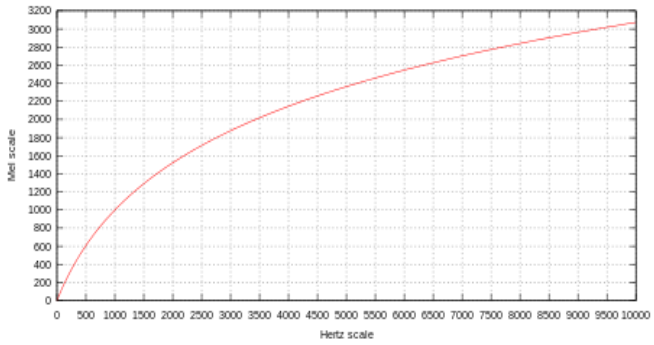

Figure.4:- Mel Scale.

The input signal is recorded at the sampling frequency of 16 kHz. It is chosen in order to reduce the effect of aliasing. This is called as front-end analysis, which involves feature extraction, Voice activity detection and removal of noise. It is formed both in training and testing phase. The steps involved in extraction process are explained as follows:

Pre-emphasis:-
This step involves the pre-processing of the input signal and removal of unwanted signals in the speech. The energy of higher frequency components are increased

Framing:-
The continuous input signal is divided into frames of P samples, and adjacent frames are being separated by Q frames (Q<P). The first frame consists of P frames and the next frame starts with Q frames, overlapped with P-Q samples and so on. The values to be considered for P=256 and Q=100.

Windowing:-
After framing, the next step in the process is windowing. Each individual frame has to be minimized to avoid signal discontinuities at the starting and ending of each frame. Let us define the window as w(n), here we use hamming window. Plot of hamming window is shown in below fig.5.

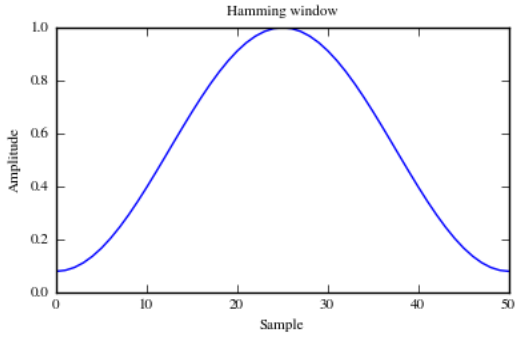

Figure.5:- Hamming Window.

Fast Fourier Transform:-
It uses the P samples for processing the data from time to frequency domain. FFT is used as a fast processing algorithm for computation of DFT. FFT reduces the computation time.

Mel-filter Bank:-
A speech signal does not follow linear scale. Thus it is subjected Mel scale.

Feature Matching:-
There are various techniques used for modeling which includes Vector Quantization (VQ), Hidden Markov Model (HMM), Dynamic Time Wrapping (DTW), Linear Predictive Coding (LPC), and Artificial Neural Network(ANN). In this paper we implemented VQ due to its simplicity and computational cost.

Vector Quantization:-
The technique VQ uses the extracted features from set of measured vectors; these are represented in centroids with equal distance from one vector to other vectors in the centroid.

The VQ consists of training set and clustering vectors. The process of VQ for two speakers is represented in the figure 6.

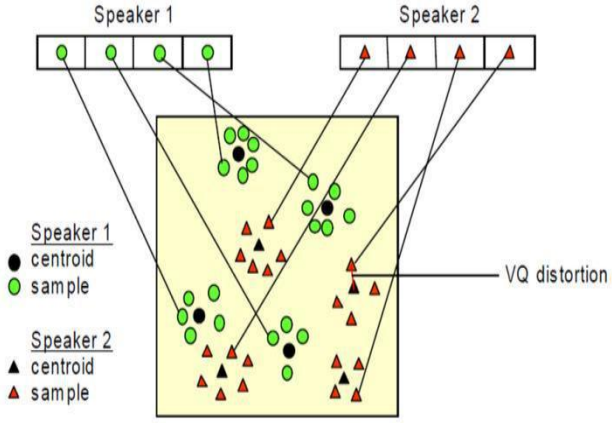

Figure 6:- Vector Quantization of two speakers.

The Eucledian distance of Two points are calculated using.

$$d(p,q) = d(q,p) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Decision Making:-

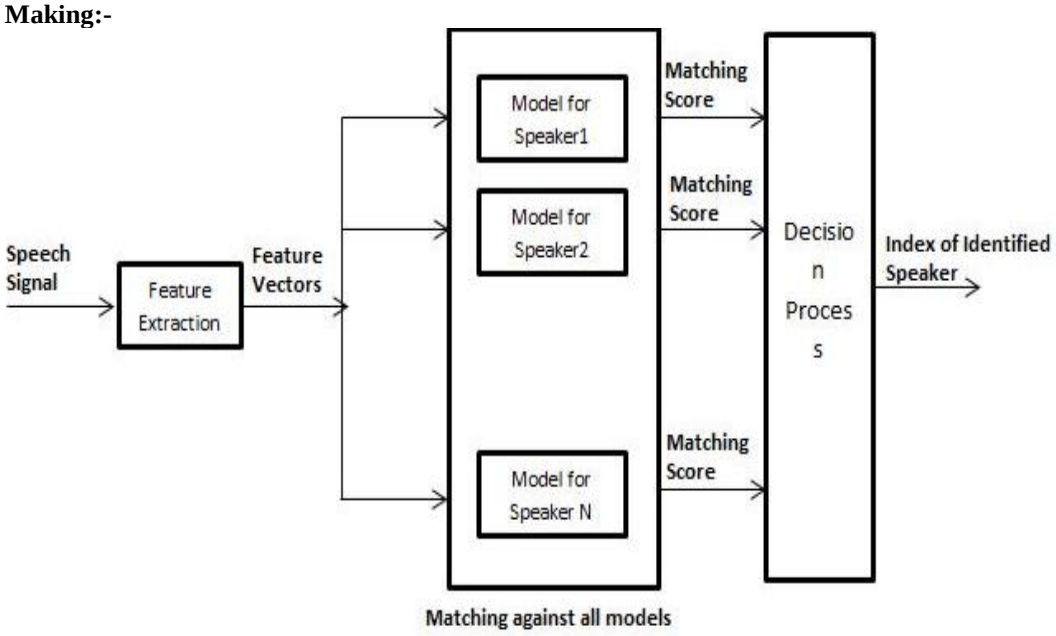

Figure.7:- Decision process

Decision process depends on the selected matching and modeling algorithms. The decision is based on the computed distances.

Results and Analysis:-

The Uttered signals with respect to two utterers are given by U1.wav and U2.wav respectively. These two files in the verification phase i.e. in the training phase are linked with the same utterers in the identification phase (testing). When both the phases are matched, we obtain as below:

Consider correct Utterers:-
Utterer 1 equals Utterer 1
Utterer 2 equals Utterer 2

Eucledian distance among the codebook of different utterers:-
Measurement of utterer 1 with respect to utterer 1 is 2.582456e+000
Measurement of utterer 1 with respect to utterer 2 is 3.423658e+000
Measurement of utterer 2 with respect to utterer 1 is 3.644154e+000
Measurement of utterer 2 with respect to utterer 2 is 2.023527e+000

Figure 8 and 9 shows the plot of Eucledian distance with respect to speaker 1 and speaker 2 respectively.

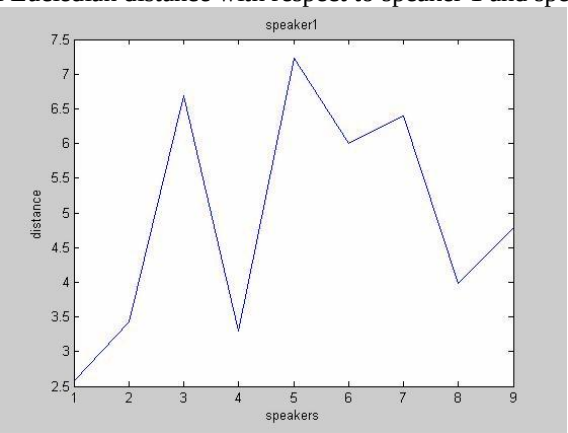

Figure.8:- Plot of Euclidean Measurement of utterer 1 with respect to utterer 2.

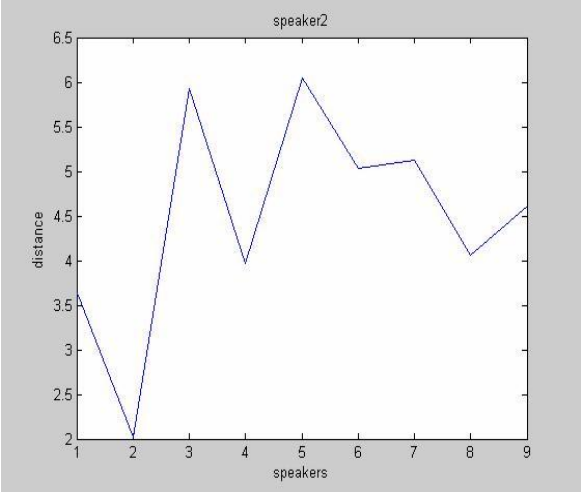

Figure.9:- Plot of Euclidean distance between speaker 2 and speaker 1.

The overall performance is limited by single vector corresponds to VQ codebook. Using high quality audio devices we can optimize the performance factor in the noise free environment.

The testing is conducted in the noise free area for good recognition accuracy of the system. We need to update the recorded database once in 2-3 years for better results of the model.

Conclusion:-

The speaker identification and verification is done with the use of feature extraction method using MFCC technique and Feature Matching method using Vector Quantization. We compared two speakers (utterers) U1 and U2 with the corresponding database stored in the VQ. The original speech signal is compared with the recorded speech signal. The unknown speaker is identified and is used for authentication of speaker identity.

Finally, it concludes that overall performance is better and the recognition rate of the model is also accurate even though it has some limitation.

References:-

1. Campbell, J.P., Jr.; **"Speaker recognition: a tutorial"** Proceedings of the IEEE Volume 85, Issue 9, Sept. 1997 Page(s):1437 – 1462.
2. Childers, D.G.; Skinner, D.P.; Kemerait, R.C.; **"The cepstrum: A guide to processing"** Proceedings of the IEEE Volume 65, Issue 10, Oct. 1977 Page(s):1428 – 1443.
3. Lawrence R. Rabiner and Ronald W. Schafer, "**Digital Processing of Speech Signals**", pearson publication.
4. S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum", IEEE Transactions on Acoustic, Speech, Signal Processing, Vol.34, issue 1, Feb 1986, pp. 52-59.
5. Nakai, M.; Shimodaira, H.; Kimura, M.; "A fast VQ codebook design algorithm for a large number of data", IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 1, Page(s):109 – 112, March 1992.