Industrial Science
Vol.1,Issue.5/June. 2014 **ISSN** : 2347-5420

SPEAKER RECOGNITION USING MEL FREQUENCY CEPSTRAL COEFFICIENT AND VECTOR QUANTIZATION

Ravi. G and Prasad Kumar B. M.

4thsem M.Te ch (L.S.P) , SJCIT, Chickballpur, India. .
Asst. Prof,. ECE Dept. SJCIT, Chickballpur, India.

Abstract

In this paper we are going to use Mel Frequency Cepstral Co-efficient (MFCC) and vector Quantization (VQ) methods for feature extraction and feature matching process. Further vector quantization is used to reduce the amount of data to be stored in the database for verification. Here we are going to recognize for the English word "zero" is obtained. The experiment results are analyzed using MATLAB. By using this approach recognition rate is reached up to 89% and distortion reduced to 69%.

KEYWORDS:

Speaker Recognition, feature extraction, feature matching, MFCC, VQ.

INTRODUCTION

Speaker recognition is very famous and active area of research. The process of speaker recognition involves the translation of the words spoken by humans in a way that must be understandable by computer. But, the speech is a very random and also same word can be spoken in many ways because it depends on speaking styles, accents, regional, gender etc. Therefore, speaker recognition with good precision is very difficult to approach and perform. So it is better to use letters, alphabets and/or words for recognition, than using sentences. In this paper, an English word "Zero" is used for speaker recognition purpose. The objective of Speaker Recognition is to extract and recognize the information about the speaker identity. Speaker Recognition is divided into two tasks: Speaker identification and Speaker Verification. Speaker Identification determines which registered speaker provides given utterance from among the set of known speakers. Speaker verification accepts or rejects the identity of speaker. This technique makes it possible to use speakers' voice to verify their identity and get control access to services like voice dialing, telephone banking, remote access to computer and security control confidential areas. Feature extraction involves pattern representation of speech signal, which involves mathematical operations to determine their contents. The key mathematical operation of almostall speaker recognition systems is the detection of beginning and end of a word to be recognized. This problem of detecting the endpoints can be easily done by human ear; but not by computers. Therefore from last few years, a number of endpoint detection methods have been developed, aiming to improve the accuracy of speaker recognition system. In addition, the background noises and random properties of speech pose major problem in speaker recognition. This paper, therefore, uses two algorithms' Mel Frequency Cepstral Coefficient (MFCC) to provide estimate vocal tract filter; and Vector Quantization (VQ) to detect recorded voice.

II METHODOLGY

The methods involved in speaker recognition are as follows:

1.Feature extraction (MFCC)

The extraction of the patterns that represent signal is an important task to produce a better recognition performance. This efficiency of this part is important for next part since it affects its behavior. MFCC is based on human hearing perceptions which cannot hear over 1KHz. In other words, MFCC is based on known variation of human ears' with frequency. MFCC has two filters which spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000 Hz. A subjective pitch is present on Mel frequency Scale to capture important characteristic of phonetic in speech. The overall process of the MFCC is shown in Fig 1.

Pre-emphasis

Pre-emphasis is done to emphasize the higher frequencies in order to increase the energy of signals at higher frequencies. For this purpose, a first order FIR filter is used with the characteristics given by: x(n) = x(n) – a*x(n-1) for 0 ≤ a ≤ 1In this paper, the value of a is set to 0.95 which means >20dB gain for higher frequency. A threshold of 0.25 is also set input signal so that a good higher level of signal can be obtained.

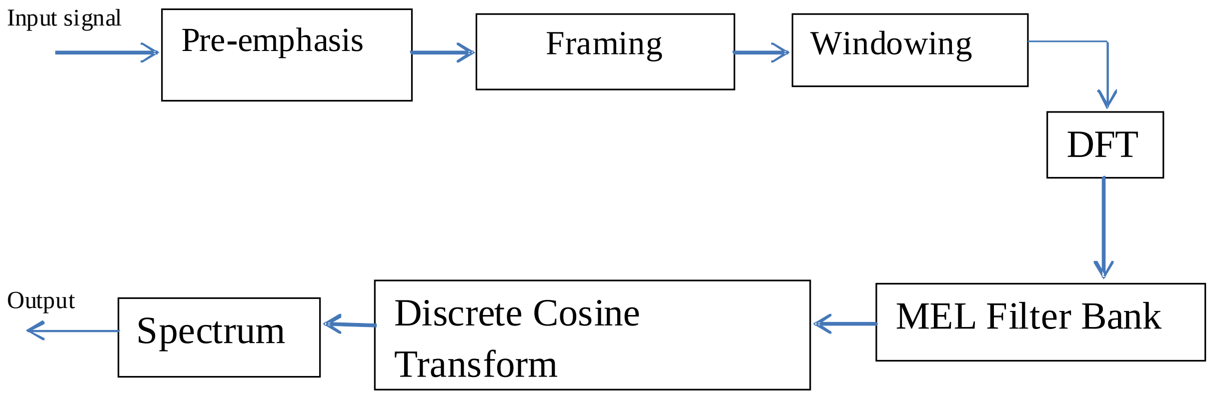


Figure 1: Block Diagram of MFCC process.

Framing: Framing refers to 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.

Windowing: The next step is to do windowing, in this paper, hamming window is used; whereby, its analytical representation is given by:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{L-1}\right)$$

for 0 ≤ n ≤ L-1

Where, w(n) is window operation, n is the sample and L is the total length of the signal.

Discrete Fourier Transform: It is used to convert samples from time domain into frequency domain, using:

$$X[k] = \sum_{n=0}^{N-1} \mathrm{x}(n) e^{-j2kn(\frac{\pi}{N})}$$

Where, x(n) is signal in time domain and the value of N is taken as 256 and K=0,1,……N-1.

MEL Filter Bank: Because of the peculiar shape of the human ear, it is not equally sensitive to all frequency bands. It is less sensitive at higher frequencies, roughly around greater than 1000 Hz. This means human perception of frequency is non-linear. Therefore for better perception, a set of filters is used, which are called MEL Filter banks, it is represented by:

F (MEL) = 2595*log [1+f/700]

Where f is frequency in Hz. These are usually triangle in shape and are used to compute a weighted sum offilter spectral components so that the output of process approximates to a MEL scale. Each filters magnitude frequency response is triangle in shape and equal to unity at the center frequency and decrease linearly to zero at the center frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components.

Discrete Cosine Transform: This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform. The result of the conversion is called Mel Frequency Cepstral Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

Spectrum: As speech signals are random, so there is a need to add features related to the change in Cepstral features over time. For this purpose, in this paper, energy and spectrum features are computed over small interval of frame of speech signals. Mathematically, the energy in a frame for a signal x in a window from time sample t1 to t2, represented as:

Energy = $\Sigma X^2[t]$

Feature Matching: Vector quantization (VQ) is a lossy data compression method based on principle of block coding. Its application in speaker recognition technology assists by creating a classification system for each speaker. It is the process of taking a large set of feature vectors and producing a smaller set of the feature vectors that represents the centroids of distribution. VQ is used here because it would be impractical to store every single feature vectorthat is generated through MFCC algorithm. The obtained feature vectors are clustered into a set of code words. The set of code words is codebook. The clustering is done using the K-Means Algorithm.

The K-means algorithm partitions the T feature vectors into M centroids. The algorithm first randomly chooses M cluster-centroids among the T feature vectors. Then each feature vector is assigned to the nearest centroid, and the new centroids are calculated for the new clusters. This procedure is continued until a stopping criterion is met, that is the mean square error between the feature vectors and the cluster centroids is below a certain threshold or there is no more change in the cluster-center assignment. In other words, the objective of the K-means is to minimize total intra-cluster variance V. In the recognition phase unknown speaker, represented by a sequence of feature vectors $\{X_1, X_2….,X_t\}$, is compared with the codebooks in the database. For each codebook a distortion measure is computed, and the speaker with the lowest distortion is chosen, one way to define the distortion measure, which is the sum of squared distances between vector and its representative (centroid), is to use the average of the Euclidean distances. Each feature vector in the sequence X is compared with all the codebooks, and the codebook with the minimized average distance is chosen to be the best. The speaker with the lowest distortion distance is chosen to be identified as the known person.

III RESULT

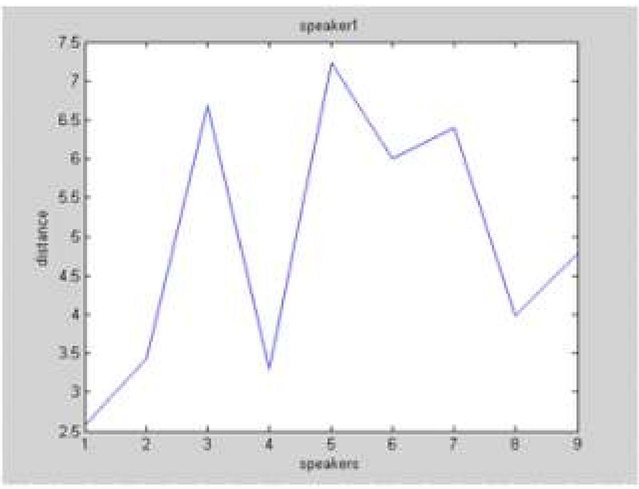
Speech signals corresponding to eight speakers i.e. S1.wav, S2.wav, S3.wav,in the training folder are compared with the speech files of the same speakers in the testing folder. The matching results of the speakers are obtained as follows.

WHEN ALL VALID SPEAKERS ARE CONSIDERED

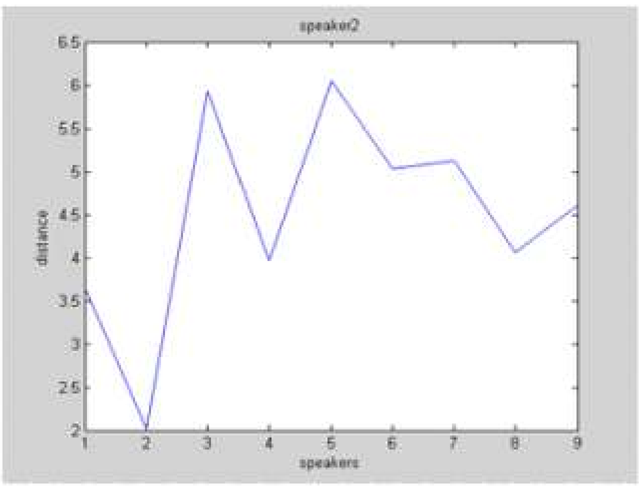
Speaker 1 matches with speaker 1
Speaker 2 matches with speaker 2
Speaker 3 matches with speaker 3

WHEN THERE IS AN IMPOSTER IN PLACE OF SPEAKER 3

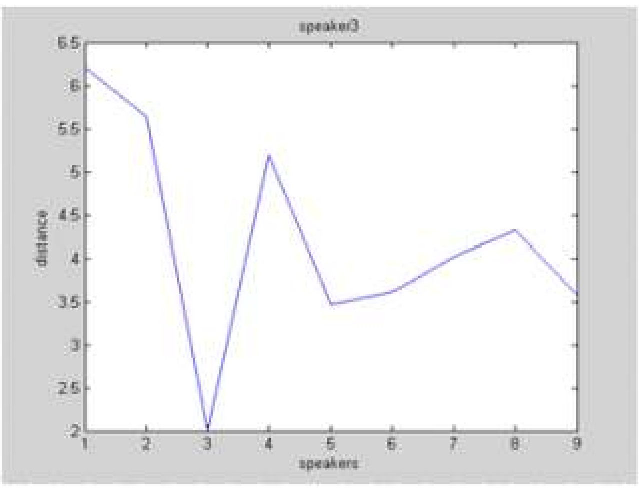
Speaker 1 matches with speaker 1
Speaker 2 matches with speaker 2
Speaker 3 is an imposter and corresponding distance is 1.06040



Plot for the Euclidean distance between the speaker 1 and all speakers.



Plot for the Euclidean distance between the speaker 2 and all speakers.



Plot for the Euclidean distance between the speaker 3 and all speakers.

IV CONCLUSION

The result obtained in this project using MFCC and VQ are applauding. We have computed MFCC corresponding to each speaker and these are vector

V REFERENCES

1.Campbell, J.P., Jr.; “Speaker recognition: a tutorial” Proceedings of the IEEE Volume85, Issue 9, Sept. 1997 Page(s):1437 – 1462.
2.Seddik, H.; Rahmouni, A.; Sayadi, M.; “Text independent speaker recognition using the Mel frequency cepstral coefficients and a neural network classifier” First International Symposium on Control, Communications and Signal Processing, Proceedings of IEEE 2004 Page(s):631 – 634.
3.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
4.Roucos, S. Berouti, M. Bolt, Beranek and Newman, Inc., Cambridge, MA; “Theapplication of probability density estimation to text-independent speaker identification” IEEE International Conference on Acoustics, Speech, and Signal Processing,ICASSP '82. Volume: 7, On page(s): 1649-1652. Publication Date: May 1982
5.Moureaux, J.M., Gauthier P, Barlaud, M and Bellemain P.”Vector quantization of rawSAR data”, IEEE International Conference on Acoustics, Speech, and Signal Processing Volume 5, Page(s):189 - 192, April 1994