International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-4 Issue-1, June 2014

Comparative Analysis of Different Windowing Techniques in MFCC Speaker Recognition

Aamir Eftikhar Bondre, Meenakshi Ananth, Nishu Nandita, Sriragh Karat, Sadashiva V Chakrasali

Abstract—Speaker recognition is the process of automatically recognising the speaker on the basis of individual information included in speech waves. The objective of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity. Speaker recognition technology can be used in many services such as voice dialling, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers. Feature extraction is an important process in speaker recognition. In this paper Mel Frequency Cepstrum Coefficients method is used in order to design a text dependent speaker recognition system. Different types of windowing methods are used during feature extraction. In this paper, a comparative analysis of different windowing techniques is done in order to determine the most effective windowing technique for MFCC speaker recognition.

Keywords: Speaker, MFCC, Mel, Frequency, Cepstrum, Coefficients.

I. INTRODUCTION

The human speech has a number of features based on which speakers can be differentiated. The maximum energy involved in an average long term speech spectrum is within the frequency band of 250Hz-500Hz. The lower frequency bands correspond to the vowel sounds and the higher frequency bands correspond to the consonant sounds. The idea involved in speaker recognition is to extract, characterise and identify the speaker using the individual speech signal. Speaker recognition is very important in today's world where security and privacy is of utmost concern. The traditional password authentication methods are not effective enough to provide the desired security and privacy of the users in today's world, which has led to the advent of bio metric security systems such as voice recognition system, retina recognition etc. Speaker recognition systems have a number of advantages as compared to other biometric systems such as: (1) Voice is ubiquitous. (2) Voice recognition is not intrusive in nature. (3) Voice recognition software is very flexible in nature as the users don't have to remember a particular password to provide access.

Manuscript published on 30 June 2014.
***Correspondence Author(s)
Aamir Eftikhar Bondre, Electronics and Communication, MSRIT, Bangalore, India.
Meenakshi Ananth, Electronics and Communication, MSRIT, Bangalore, India.
Nishu Nandita, Electronics and Communication, MSRIT, Bangalore, India.
Sriragh Karat, Electronics and Communication, MSRIT, Bangalore, India.
Sadashiva V Chakrasali,t, Electronics and Communication, MSRIT, Bangalore, India.

(4) Voice recognition doesn't require any additional infrastructure which makes it cost effective in nature.
(5) Voice recognition systems are robust in nature. In this paper, the Mel Frequency Cepstrum Coefficients (MFCC) method has been used in order to extract features from the voice signals. The code consists of a testing phase and a training phase. The data from the training phase and the testing phase are compared , based on which it is determined whether the two voice samples match.

II. SPEAKER RECOGNITION

Speaker recognition system has become one of the most popular methods in the advent of biometric security. The idea of speaker recognition is essentially derived from modeling the human body. The human body performs feature matching at a very low level pattern classification and processing, which makes it very difficult for the machines to perform feature matching in a similar manner[1]. This low level pattern classification involves many other forms of knowledge such as linguistic and semantic knowledge. Thus, an Automatic Speech Recognition (ASR) essentially balances between the ideal and the practical models. A speaker recognition system contains of two main parts.

- Feature extraction - Feature extraction is the process of extracting the features of a particular voice sample that can be later used to represent a speaker.
- Feature matching - Feature matching is the process of identifying an unknown speaker by matching his/her features with an existing database.

A speaker recognition system comprises of two phases, training phase and testing phase. The training phase involves the process of extracting the features in the voice sample and storing them whereas in the testing phase, the features extracted from the voice samples are matched with the features stored in the database[2][13]. Thus, training is the task of familiarizing and testing is the actual identification process. The level of match by comparing the training phase and the testing phase is used in order to arrive at a result. A speaker recognition system consist of four different modules:

- Front end processing- The process of converting the voice input sample into a set of feature vectors is called front end processing. It is performed in the training phase as well as the testing phase [2].
- Speaker modelling- The reduction of feature data by modelling the distributions is done by this process.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

- Speaker database - The features of the speaker which are modelled is stored in the database.
- Decision logic-The feature matching is done in this module, the features of the unknown speaker is compared with rest of the speaker models in the database[2].

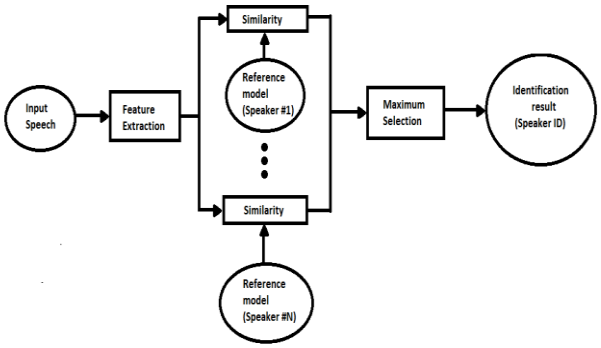

Figure 1. Training Phase

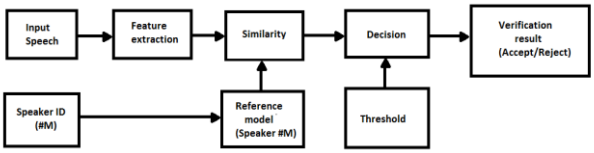

Figure 2. Testing Phase

The task of speaker recognition is made difficult by the highly variant nature of the input. Since the system being designed is text dependent in nature, the word which is recorded and tested must be the same in order to obtain a positive result. The speaker's voice can vary greatly in the training and testing sessions due to various factors such as health condition, amplitude, rate of speech etc [6]. The environmental and background noises also present a challenge to the speaker recognition system. Thus, it is important to develop a robust system that copes with these real world problems and computes results in real time. The computation of these results must be done in real time as many applications such as security of the data in a smartphone must be done in real time[6][14].

III. Feature Extraction Technique

The first step in the implementation of any speech recognition system is extracting the features. The features can be extracted either directly from the time domain signal or from a transformation domain depending upon the choice of signal analysis approach. Some of the audio features that have been successfully used for audio classification include Linear Predictive Coding (LPC) and Mel-Frequency Cepstral Coefficients (MFCC).

A. Linear Predictive Coding (LPC)

LPC (Linear Predictive Coding) analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering and the remaining signal is called the residue[2]. In LPC system, each sample of the signal is expressed as a linear combination of the previous samples. This equation is called a linear predictor and hence it is called as linear predictive coding. The coefficients of the difference equation(the prediction coefficients) characterize the formants.

B. Mel-Frequency Cepstral Coefficients(MFCC)

MFCC (Mel-Frequency Cepstral Coefficients) is based on the human peripheral auditory system. The human perception of the frequency contents of sounds for speech signal does not follow a linear scale. Thus for each tone with an actual frequency t measured in Hz, a subjective pitch is measured on a scale called the Mel scale [2][15]. The Mel frequency scale is linear frequency spacing below 1000 Hz and logarithmic spacing above 1 kHz. As a reference point, the pitch of a 1kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels.

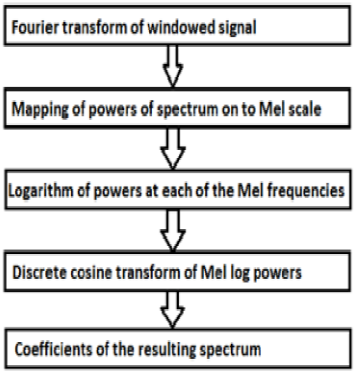

Figure 3. MFCC Flowchart

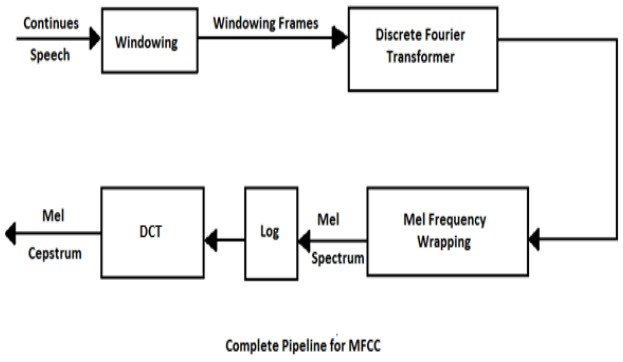

Figure. 4. Complete Pipeline for MFCC

IV. Mel Frequency Cepstral Coefficients

The extraction and selection of the best parametric representation of the acoustic signals is an important task in the design of any speech recognition system; it significantly affects the recognition performance.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

A compact representation would be provided by a set of mel-frequency cepstrum coefficients (MFCC), which are the results of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale[4][8]. The MFCCs are proved more efficient. The calculation of MFCC includes the following steps as shown in the Figure 3.We can use the following approximate formula to compute the mels for a given frequency f in Hz.

$$Mel(f) = 2595 * log_{10}(1 + \frac{f}{700}) \quad (1)$$

MFCCs are calculated by first taking the Fourier transformof a windowed signal and mapping the powers of the spectrums obtained above onto the mel-scale, using triangular overlapping filters. Next, the log of powers at each of the mel frequencies is taken and Discrete Cosine Transform is applied to it. The MFCCs are the amplitudes of the resulting spectrum [12][11]. The Discrete Cosine Transform is done for transforming the mel coefficients back to time domain and for
decorrelation.

$$C_n = \sum_{k=1}^{k} log(S_k)\ cos\{n(\frac{k-1}{2}) * \frac{\pi}{k}\} \quad (2)$$

$n=1,2,3....k$

V. K Means Vector Quantization

Feature matching is the process of identifying or matching the unknown data with the given set of data in the database. In this paper, the feature matching technique used is the k-means vector quantization method. The main aim of k means clustering technique is to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as the prototype for the cluster. As a result of this data space is partitioned into Voronoi cells [2]. Vector quantization is a method in which the modeling of the probability density functions is done by th distribution of vector prototypes. In vector quantization large set of points are divided into groups which have approximately the same set of points as closest to them. Every group in vector quantization is represented by its centroid point. Since data points are represented by the index of their closest centroid commonly occurring data have low error, and rare data high error. It can be used for lossy data correction and density estimation [2].

A. Algorithm for k Means Clustering

1) K random vectors are selected from the training set and are called code-vectors.
2) The squared Euclidean distance of all the training vectors with the selected k vectors is found and k clusters are formed [5].
3) The training vectors X_j is put in i th cluster if the squared Euclidean distance of the X_j with i th code vector is minimum.
4) If the squared Euclidean distance of X_j with code vectors is minimum for more than one code-vector then X_j is put in any one of them.
5) The centroid of each cluster is computed.
6) The centroid in each cluster act as an input for the next cluster.
7) The Mean Square Error is computed for each of the k clusters.
8) The net Mean Square Error is then computed.
9) The above process is repeated till the Mean Square Error converges.

The fundamental advantages of k means clustering are:

- The k means algorithm provides faster computation as compared to hierarchal for huge data provided k is kept small[5].
- The clusters produced by k means algorithm are tighter.

VI. Comparison Of Implementation Using Different Windows

In speaker recognition, the windows are applied to raw speech frames in order to reduce the spectral leakages effect [12]. Windows are basically used in speaker recognition to remove discontinuities in speech. While extracting MFCC the window attenuates both ends of the frame (this is compensated by overlapping the frames at the next stage).This removes the abrupt changes at the ends. These windows have reasonable side lobe and main lobe characteristics which are required for the DFT computation. In practice, selecting the optimal window function for speech processing application is still an open challenge. Thus, in this paper, the main objective is to find out the most efficient window for a text dependent MFCC speaker recognition system. In order to do this we have taken four common windowing techniques into consideration:

1) Blackman window
2) Hamming window
3) Hanning window
4) Kaiser window

The code was tested on both male and female voices. The threshold value for male voice was set higher than female voice to accommodate the variations in male voice. Ten variations were tested for each voice that included duplication of the original sentence, speaking at a slower or faster pace, speaking in a louder or softer voice as well as mimicry.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Table 1. Blackman Window

Speaker	No. of Attempts	Correct Acceptance	Correct Rejection	False Acceptance	False Rejection
S1	10	6	2	0	2
S2	10	3	1	1	5
S3	10	3	2	0	5
S4	10	4	2	0	4
S5	10	4	1	1	4
S6	10	4	2	0	4
S7	10	4	1	1	3
S8	10	8	2	0	0
S9	10	5	2	0	3
Total	**90**	**42**	**15**	**3**	**30**

Table 2. Hamming Window

Speaker	No. of Attempts	Correct Acceptance	Correct Rejection	False Acceptance	False Rejection
S1	10	6	2	0	2
S2	10	3	1	1	5
S3	10	3	2	0	5
S4	10	5	2	0	3
S5	10	3	1	1	5
S6	10	4	2	0	4
S7	10	4	1	1	3
S8	10	8	2	0	0
S9	10	5	2	0	3
Total	**90**	**41**	**15**	**3**	**31**

The observation was made for correct acceptance and rejection and false acceptance and rejection. Pace change was found to be more acceptable than volume change. Also, when a different sentence or a different language was spoken, it was found to be closer to the threshold if it had one or more words of the original sentence than without any. Percentage efficiency shows Kaiser Window to be the most efficient amongst the 4 windows used here. Graphs for percentage efficiency in terms of correct acceptance and mean square error for each of the windows are shown. It is seen that Kaiser window shows highest percentage of efficiency. It gives the least mean square error as well.

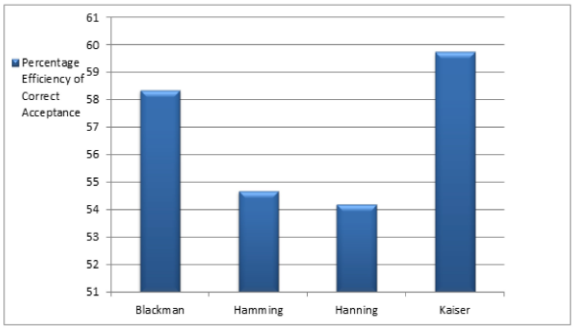

Figure 5. Efficiency of Correct Acceptance

Table 3. Hanning Window

Speaker	No. of Attempts	Correct Acceptance	Correct Rejection	False Acceptance	False Rejection
S1	10	6	2	0	2
S2	10	3	1	1	5
S3	10	3	2	0	5
S4	10	4	2	0	4
S5	10	3	1	1	5

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

S6	10	4	2	0	4
S7	10	3	1	1	5
S8	10	8	1	1	0
S9	10	5	2	0	3
Total	**90**	**39**	**14**	**4**	**33**

Table 4. Keiser Window

Speaker	No. of Attempts	Correct Acceptance	Correct Rejection	False Acceptance	False Rejection
S1	10	6	2	0	2
S2	10	3	1	1	3
S3	10	3	2	0	5
S4	10	5	2	0	3
S5	10	4	1	1	4
S6	10	4	2	0	4
S7	10	3	1	1	5
S8	10	8	2	0	0
S9	10	5	2	0	3
Total	**90**	**43**	**14**	**4**	**29**

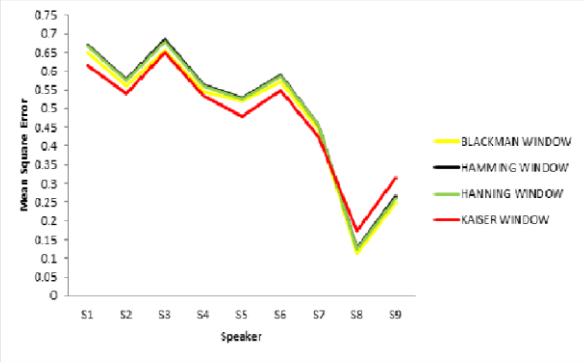

Fig. 6. Mean Square Error

VII. CONCLUSION

Speaker recognition has been recognised as an important biometric security tool in today's world. Thus, it is important to make sure that a standard is maintained in the quality of this system. In this paper, we have thus made a comparative analysis of a text dependent Mel Frequency Cepstrum Coefficients speaker recognition tool using four different commonly used windowing techniques. The implementation was done considering th different types of variation which can occur in the real world scenario. Thus, male and female voices were taken into consideration and different variations of their voices were tested which involved pace change, volume change and mimicry. From the observations it was found that pace change was more acceptable than volume change. After rigorous analysis it was found that Kaiser window provides the most efficient result amongst Hamming, Hanning and Blackman. The most ineffective windowing technique among the four was found to be Hanning. Future Scope In this paper, the analysis of the windowing techniques were performed for a text dependent MFCC speaker recognition. The text dependent MFCC speaker recognition system has limitations, these limitations can be overcome by a text independent speaker recognition system. The text independents peaker recognition can be further improved upon by making a universal language independent speaker recognition system which is solely based on the sound of the speaker.

REFERENCES

1. K.K. Paliwal and B.S. Atal, "Frequency related representation of speech." in Proc. EUROSPEECH,p.p.65-68 Sep. (2003).
2. Vibha Tiwari, "MFCC and its applications in speaker recognition" International Journal on Emerging Technologies ISSN : 0975-8364.
3. T. Fukuda, M. Takigawa and T. Nitta, "Peripheral features for HMM based speech recognition" in Proc.ICASSP,1: 129-132(2001).
4. M. Pandit and J. Kittler, "Feature selection for a dtw-based speaker verification system" Proceedings of IEEE Int.Conf. Acoust. And Signal Processing,2: 769-772 (1998).
5. Dr. H.B. Kekre, Ms. Tanuja K. Sarode, "Vector Quantized Codebook Optimization using K-Means",International Journal on Computer Science and Engineering,Vol.1(3), 2009, 283-290.
6. Darshan Mandalia and Pravin Gareta,"Speaker Recognition Using MFCC and Vector Quantization Model".
7. Atal, B.S. and S.L. Hanauer,"Speech analysis and synthesis by linear prediction of the speech wave",Journal of the acoustical society of America,50: 637-655(1971)
8. Speaker recognition using MFCC by S. Khan, Mohd Rafibul lslam, M. Faizul, D. Doll, IJCSES (International Journal of Computer Science and Engineering System)2(1): 2008.
9. Molau, S, Pitz, M, Schluter, R, and Ney, H., "Computing Mel frequency coefficients on Power Spectrum",Proceedings of IEEE ICASSP-2001,1: 73-76(2001).
10. Lawrence Rabiner and Biing-Hwang Juang, Fundamentals of Speech Recognition Prentice-Hall, Englewood Cliffs, N.J.,(1993).
11. Bhupinder Singh, Rupinder Kaur, Nidhi Devgun, Ramandeep Kaur,"The process of Feature Extraction in Automatic Speech Recognition System for Computer Machine Interaction with Humans: A Review",International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 2, February 2012 ISSN: 2277 128X
12. Leigh D. Alsteris and Kuldip K. Paliwal,"Importance Of Window Shape For Phase-Only Reconstruction Of Speech",presented in International Conference on Acoustics,Speech and Signal Processing
13. J.B. Allen and L.R. Rabiner," A unified approach to short time Fourier analysis and synthesis"Proc. IEEE, Vol. 65, No.11, pp. 1558 1564, 1977

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

14. Premakanthan and W.B. Mikhael, Speaker verification/ recognition and the importance of selective feature extraction: Review,Proceedings of the 44th IEEE 2001,Midwest Symposium, 1:14-17(2001).
15. Goutam Saha and Malyaban Das,On Use of Singular Value Ratio Spectrum as Feature Extraction Tool in Speaker Recognition Application, CIT-2003, pp. 345-350,Bhubaneswar, Orissa, India, (2003).

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)