DOI: 10.21917/ijsc.2019.0255

SPEAKER IDENTIFICATION FOR ISOLATED GUJARATI DIGITS USING MFCC AND VQ

Pooja Prajapati and Miral Patel
Department of Information Technology, G H Patel College of Engineering and Technology, India

Abstract

The research presented in this paper is part of an ongoing investigation of speaker identification for Gujarati isolated digit. In our previous work, we evaluated feature extraction method of Gujarati isolated digit for speaker identification using Mel-Frequency Cepstral Coefficient (MFCC). Here we establish on our previous research work for speaker identification of Gujarati isolated digits. The aim of our proposed work is to provide best security or identification to the authentication system which will identify Gujarati speakers. For the projected work we have used MFCC and VQ computation scheme for the Feature extraction and pattern matching techniques. The study also explains about using the technique in brief. For the proposed approach dataset of Gujarati numeral (0 to 10) was recorded from different speakers from different age groups which serve to train and test each speaker speech sample. The distance between each test codeword and each codeword in master codebook is computed. That difference helps in making recognition decision. An experimental evaluation is done using MATLAB simulations. The outcomes indicate that our proposed work of speaker identification system for Gujarati isolated digit achieves good amount of results on our Gujarati digit database with the combination of the proposed technique.

Keywords:

Feature Extraction, Isolated Gujarati Digit, MFCC, Speaker Identification, VQ

1. INTRODUCTION

Speech is a common and one of the most significant ways of human computer interaction. Voice is a natural way of communication and non-intrusive as a biometric [5]. Like fingerprints, it conveys the identity of the speaker as voice print. The Human voice is a unitary type of signal that comprises a miscellaneous type of information, including words, beliefs, language and identity of the speaker [6] [7]. There are a number of situations in which correct recognition of person is required. The use of biometric based, forensics, Voiceprint is a safe and a secure method for authenticating an individual's identity that unlike passwords or token, these features of voice cannot be stolen, duplicated or forgotten [10] [11] [18]. Basically speech recognition problem can focus on identifying the speech or speaker who uttered the speech or language in which the speech is uttered. In recent years, a serious amount of work has been contributed to English and other Indian languages; out of these Gujarati languages has the least amount of work [14]. Thus, it's taking to produce a speaker identification system for Gujarati language. A state of the art speaker recognition system has three fundamental sections, a feature extraction unit for representing speech signal in a compact manner, a modeling scheme to characterize those features using statistical approach [12] [13], and lastly a classification scheme for characterizing the unknown utterance. A significant amount of research has been performed on Speaker identification from digitized speech for applications such as verification of identity. These systems use features which are used in speech recognition and speaker recognition. Systems are trained on data without background noise and performance tends to degrade in noisy environments. Here some research and development related to different languages for isolated digit and word recognition are discussed,

Chuahan and Tanawala (2015) have compared both MFCC and LPC method under vector quantization (VQ) method at comparative study of MFCC and LPC Algorithms for Gujarati isolated word recognition in which they use database of both male and female voices where each word is repeated at 5 times by speakers where results shows that using LPC, accuracy is above 85%, MFCC is more above 95%. So author concluded here that MFCC performs better for feature extractor [1].

Elouahabi et al. (2016) have described amazigh isolated word speech recognition system using hidden Markov model Toolkit (HTK), where amazigh language is known as Berber or Tamazight that is so vast in Africa. HTK tool that uses HMM model for to develop the system and MFCC for feature extraction. Proposed model works on both 10 digits and 33 alphabets collected from 60 both male and female speakers. The Overall accuracy of the proposed model achieves 80% [23].

Therese and Lingam, (2015) have described speaker based language independent isolated speech recognition system proposed a model in which most widely used MFCC is used for feature extraction also k means algorithm is used for specific feature extraction. Here this proposed system is not only used for recognizing the speech but also for language in which speech is uttered. Vocabulary contains digits from 1 to 10 of seven different languages. Which achieves 97.14% accuracy except digit three and seven [9].

Kumari et al. (2017) have described Singer Identification using MFCC and LPC and its comparison for ANN and Naïve Bayes Classifier, where author focuses on the singer identification using MFCC and LPC coefficients from Indian audio songs. In which the audio songs used are divided into segments each of 10 seconds and for each segments 13 Mel-Frequency Cepstral Coefficients (MFCC) and 13 linear predictive coding (LPC) coefficients are computed. Classifier models are trained using Naive Bayes classifier and back propagation algorithm using neural network. The results shows that MFCC features proved to provide a better result as compared to LPC for both the classifiers with the identification percentage of 77% [19].

Mallikarjunan, et al. (2018) have described Text-Independent Speaker Recognition in Clean and Noisy Backgrounds Using Modified VQ-LBG Algorithm where the Mel frequency cepstrum coefficient (MFCC) technique for extracting the features from the speaker speech sample. These cepstrum coefficients are named as extracted features. The extracted MFCC features are given as input to the modified vector quantization via Linde–Buzo–Gray (modified VQ-LBG) process and expectation maximization (EM) algorithm. Vector quantization technique is mainly used for feature matching where a separate codebook will be generated for each speaker. Also The EM algorithm is utilized to develop the Gaussian mixture model–universal background model (GMM–UBM). In GMM–UBM model, *k* means cluster is summed up to consolidate data about the covariance structure of the information and the focuses of the inert Gaussians. From the comparative analysis it proves that VQ-LBG algorithm gives better performance compared to the GMM-UBM model [20].

Apart from this survey it's observed that the isolated digit recognition system is implemented in English language. Little work is done in other similar languages like Gujarati, Hindi, Bengali, Tamil etc. So we conclude with the decision for isolated Gujarati digit recognition, apart from those languages, requires more attention. Also it is shown that MFCC is used for feature extraction and VQ-LBG is for pattern matching achieves more accurate results in identification or recognition system. Where, the Speaker Identification for Gujarati language is quite difficult because there are many highly confusable forms of digits spoken by different speakers of different dialects and different regions and also necessary too, for the illiterates for Security purpose, transaction authentication, facilities on computer access control, monitoring, telephone voice authentication for long distance calling and banking access etc. Where our proposed approach works in a way to identify or recognize the different patterns of the Gujarati font like frequency, pitch etc. When user speaks any digit from the microphone the different patterns of the digits will be identified and it will be compared with the corresponding pattern stored in the standard phoneme databases and corresponding highest matching digit of Gujarati language will be return in form of text on the screen. So, The objective of our proposed model is to provide better security authentication through speaker identification that help in achieving better improvements in identification of Gujarati speakers.

The paper is organized into four sections. Section 1 gives an introduction, section 2 indicates the Proposed Model of Speaker Identification. For Gujarati Isolated Digit section 3 discussed Experimental Setup section 4 focuses on the methodology of technique used for the proposed approach and section 5 shows Comparative Analysis and Discussion of Results and section 6 followed by Conclusion and future work.

2. THE PROPOSED MODEL OF SPEAKER IDENTIFICATION FOR GUJARATI ISOLATED DIGIT

The Fig.1 shows the steps to perform identification of speaker in the Speaker Identification process. The process of speaker identification is divided into two main stages, the enrollment phase and the identification phase [9]. During the registration phase (also known as speaker training), speech samples are accumulated from the loudspeakers, and they are utilized to train their models. The collection of enrolled models is also called a speaker database. In the second phase (identification phase), a test sample of speaker is compared against the speaker database. Both the phases involve a common first step, feature extraction, where the speaker dependent features are extracted from the speech sample [2] [3]. Then in the enrollment phase, these characteristics are modeled and stored in the speaker database. In the identification step, the extracted features are compared against the stored models present in the speaker database. Based on results obtained from these comparisons the final decision about speaker identity is made [15].

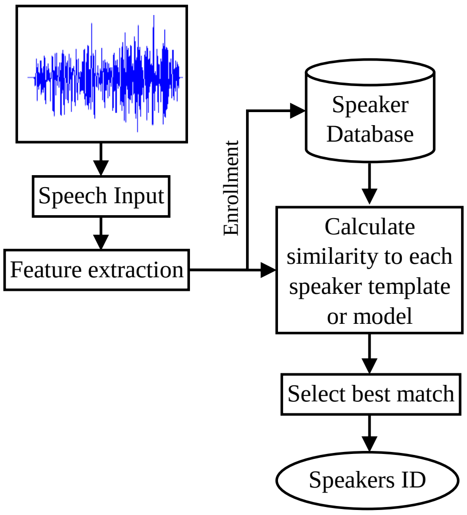


Fig.1. Proposed Work

In our proposed approach for speaker identification system for Isolated Gujarati Digit we have applied feature extraction technique of MFCC to extract the characteristic from voice samples. After extracting features from voice signals using feature extraction technique. The next measure is performed for the pattern matching using VQ method. Vector quantization (VQ) is employed for comparing the trained data with input data. VQ is applied for calculating the similarity by calculating minimum distance between feature vectors of each sample and all those samples which is already stored in the database. Based on results obtained from these comparisons the final decision about speaker identity is made. That nearest match helps in identifying each speaker. In order to train and test the recognizer, here we are concentrating on Database development and methodology used in our proposed work that is discussed in above sections.

3. EXPERIMENTAL SETUP

For the projected work, as there was no standardized dataset available for Gujarati numerals, the database is prepared for Gujarati numerals spoken by different speakers.

3.1 DATASET PREPARATION

Here, Speech samples are collected of Gujarati 0 to 10 digits from 40 speakers of different age groups. There are ten speakers in each group. Each speaker pronounced from 0 to 10 in Gujarati language. Here, we consider some factors while collecting speech samples like speaking condition, pronunciation of different speakers, Environment etc. that may pretend to train the dataset. Speakers are consisting of different age and gender. There are four age groups, a first age group having ranged between 5 to 15 years, in this group speech sample is accumulated from the minors. The second age group having ranged between 16 to 30 years, in this group speech sample is collected from tanagers as well as from younger. The third age group having ranged between 31 to 50 years, in this group speech sample is gathered up from five males and five females. Forth age group having ranged between 51 to 80 years, in this group speech sample is collected from aged people. Speech samples were broken by a disturbance. Each speech samples were collected within free sound recorder software in .wav format. This input speech data are passed to the feature extraction module. The feature extraction module extracts the unique characteristics of spoken data using mfcc() in MATLAB Environment. The pronunciation of each Gujarati numeral is given below in Table.1.

Table.1. Pronunciation of Gujarati Numerals

Gujarati Numerals	Pronunciation
૦	Shunya
૧	Ek
૨	Be
૩	Tran
૪	Char
૫	Panch
૬	Chha
૭	Saat
૮	Aath
૯	Nav
૧૦	Dash

4. METHODOLOGY OF PROPOSED WORK

Mel-Frequency Cepstral Coefficient (MFCC) is one of the best technique for feature extraction, especially for automatic speech and speaker recognition system. Which is most widely practiced technique for feature extraction in ASR system as well as speaker recognition because it is a standard method as well as less complex in implementation and more effective and robust under various conditions [4] [18] [19].

MFCC gives more efficient and accurate result, than other feature extraction techniques in the voice identification system [3] [16] [17]. The MFCC coefficients can be used as audio classification features to improve the classification accuracy of the speaker there is a computation for extracting the Cepstral feature parameters from the Mel scaling frequency domain. The procedural steps for obtaining that feature vectors using MFCC are given to lower place. The results of MFCC process are shown and described in following sections.

4.1 FEATURE EXTRACTION USING MFCC

Mel Frequency Cepstral Coefficients (MFCCs) are coefficients that represent audio. They derive from a type of coastal representation of the audio file. The difference between the cepstrum and the Mel-frequency cepstrum is in the MFCC, These cepstrum coefficients are the result of cosine transform of the real logarithm of the short time energy spectrum expressed on a Mel-frequency scale. In MFCC, the main advantage is that it uses Mel frequency scaling which is approximate to the human ear [9]. A block diagram of the structure of an MFCC computation process is shown below in Fig.2.

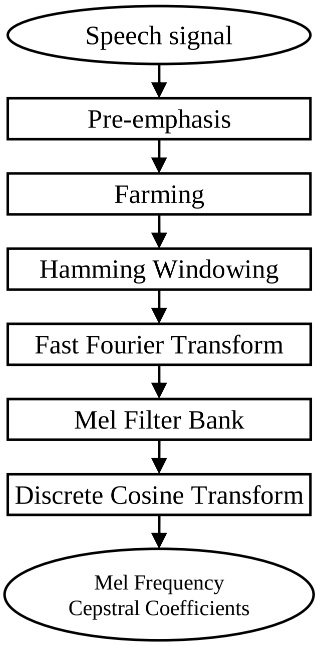


Fig.2. Block diagram of MFCC computation process [3]

As shown in Fig.1, the basic steps in the computation process of extracting Cepstral coefficients from the speech signal are positioned below.

- Pre-emphasize input signal.
- Perform short-time Fourier analysis to get a magnitude spectrum.
- Wrap the magnitude spectrum into Mel-spectrum.
- Take the log operation on the power spectrum (i.e. Square of Mel-spectrum).
- Apply the discrete cosine transform (DCT) on the log-Mel power spectrum to derive Cepstral features and perform Cepstral [3] [8].

The signal is segmented into shorter frames where each frame comprises of the signal information of 20 to 30 milliseconds for which the vocal tract shape is assumed to be stationary of each one when a person utters any speech. The frames are allowed to overlap with each other to avoid any loss in the signal while segmenting. Each frame is multiplied by a windowing function. In this model hamming window function is applied. A vector of acoustical coefficients is extracted from each windowed short frame. Since the frequency bands are placed logarithmically these coefficients allow better processing of speech data. In the melfrequency Cepstral coefficient scale the calculation of the Cepstral coefficients is same as the real cepstrum except Mel cepstrum's frequency scale is warped to keep up a correspondence to the mel scale. The mel scale is mainly based on the study of observing the pitch or frequency perceived by the human. The scale is divided into the units referred as mel. Mel (melody) is a unit of pitch. The mel scale is normally a linear mapping below 1000Hz and logarithmically spaced above 1000Hz. The Fig.6 below shows the example of normal frequency mapped into the

Mel frequency. Thus the mel-scale used in this work is to map between linear frequency scale of the speech signal to logarithmic scale for frequencies higher than 1kHz. This makes the spectral frequency characteristics of signal closely corresponding to the human auditory perception [13][21].

The mel scale mapping is formulated as,

$$Mel(f)= 2595*\text{Log}_{10}(1+f/700) \qquad (1)$$

where, f is the linear frequency in speech signal.

For the present work, related to the MFCC computation scheme is addressed carefully along with an experimental evaluation.

4.1.1 Pre-Emphasis:

This step processes the passing of a signal through a filter which emphasizes higher frequencies. This process will increase the energy of the signal at higher frequency. The speech generated from mouth will loss the information at high frequency. Thus, it needs the pre emphasis process in order to compensate the high frequency loss. Each frame needs to be emphasized by the high frequency filter. And for speech signal spectrum, the higher the frequency is, the more serious loss will be, where requires doing some computation of high frequency information that is known as pre emphasis. In speech there is 1st order high pass filter.

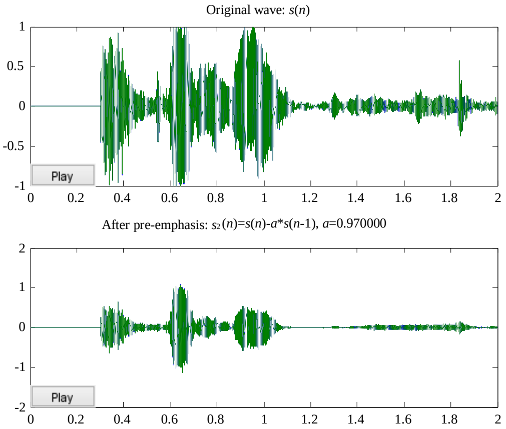


Fig.3. The pre-emphasis wav for digit ‘O’ collected from the minors

Speech signal in time domain after pre emphasis can be defined as,

$$S_1(n) = S(n) – \alpha S(n\text{-}1) \qquad (3)$$

where $s(n)$ is the speech signal and parameter α is usually between 0.94 and 0.97. Pre-emphasis is needed for high frequency in order to improve phone recognition performance. The simulation of Pre-emphasis wav for digit ‘O’ collected from the minors is shown in below Fig.3.

4.1.2 Framing:

The pre-emphasized speech signal is segmented into small duration blocks of frames. The width of the frames is generally about 30ms with an overlap of about 20ms (10ms shift). The first frame contains N sample points of the speech signal. The second frame begins M samples after the first frame, and overlaps it by N×M samples. Similarly, the third frame begins 2M samples after the first frame (or M samples after the second frame) and overlaps it by N×2M samples. This process continues until all the speech is accounted for within one or more frames. Typical values of N and M are N = 256 (which is equivalent to ~30msec windowing and facilitate the fast radix-2 FFT) and M = 100. For each frame 20mfcc were calculated.

4.1.3 Windowing:

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as, where N is the number of samples in each frame, then the result of windowing is the signal. Typically the Hamming Window is used. In Matlab “hamming” function is used to find the hamming window. The objective is to reduce the spectral effects. The coefficients of a Hamming window are computed from the following equation,

$$W\left[k+1\right]=0.54-0.46\cos\left(2\pi\frac{k}{n-1}\right),k=0,1,...,n-1. \qquad (3)$$

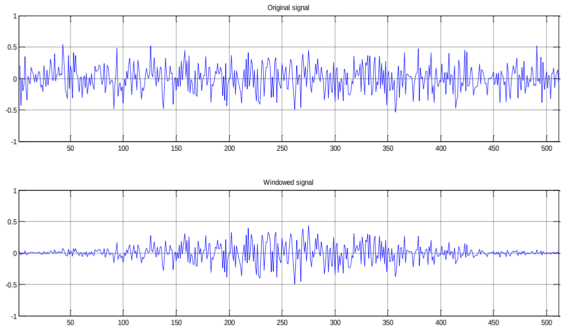


Fig.4. Windowed Frame on audio digit ‘O’ collected from the minors

The simulation of Windowed Frame on audio digit ‘O’ collected from the minors is shown in below Fig.4.

4.1.4 Fast Fourier Transform:

The next step is FFT performed to obtain the magnitude frequency response of each frame which is assumed of periodic within the frame.

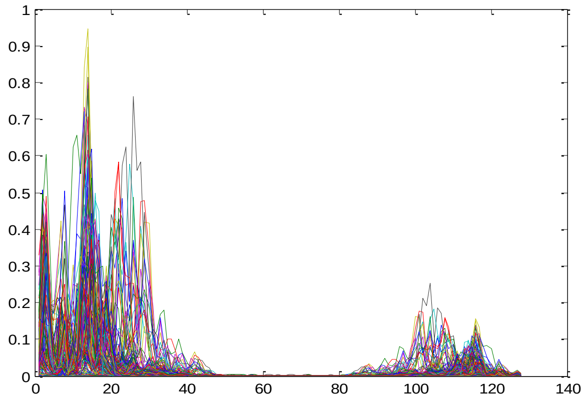


Fig.5. Detailed FFT magnitude of audio digit ‘O’ collected from the minors

The Fast Fourier Transform which converts each frame of *N* samples in time domain to frequency domain. The frame blocking step that was previously done was to enable the ease of performing of the FFT. The triangular bandpass filters are used to extract an envelope like features. The multiple the magnitude frequency response from a set of triangular bandpass filters to get the log energy of each triangular bandpass filter which will give the nonlinear perception for different tones or pitch of voice signal. The amplitude spectrum of the signal passed through the window is calculated by FFT. FFT size can be 512, 1024 or 2048.

The simulation of Detailed FFT magnitude of audio digit 'ο' collected from the minors is shown in below Fig.5.

4.1.5 Mel Filter Bank

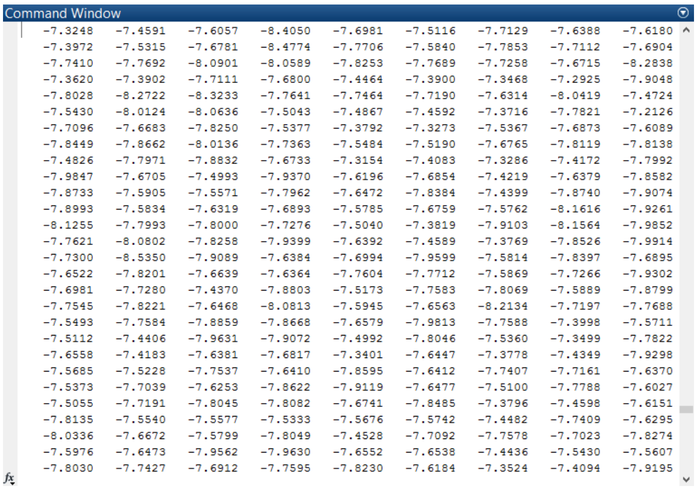
Mel (melody) is a unit of pitch. The spectrum obtained from the above step is Mel Frequency Wrapped. The Mel-frequency scale is a linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. As a reference point, the pitch of a 1kHz tone, 40dB above the perceptual hearing threshold, is defined as 1000 Mels. Therefore, we can use the following approximate formula to compute the Mels for a given frequency *f* in Hz [22].

$$Mel(f) = 2595*\mathrm{Log}_{10}(1+f/700) \qquad (4)$$

The major work done in this process is to convert the frequency spectrum to Mel spectrum. As research work shown that speech signal does not follow linear scale. So, for each tone with actual frequency, *f* is measured in Hz, pitch is measured on a scale called the 'Mel scale'.

4.1.6 Discrete Cosine Transform and Mel Cepstrum Coefficients:

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT).



```
   -7.3248   -7.4591   -7.6057   -8.4050   -7.6981   -7.5116   -7.7129   -7.6388   -7.6180
   -7.3972   -7.5315   -7.6781   -8.4774   -7.7706   -7.5840   -7.7853   -7.7112   -7.6904
   -7.7410   -7.7692   -8.0901   -8.0589   -7.8253   -7.7689   -7.7258   -7.6715   -8.2838
   -7.3620   -7.3902   -7.7111   -7.6800   -7.4464   -7.3900   -7.3468   -7.2925   -7.9048
   -7.8028   -8.2722   -8.3233   -7.7641   -7.7464   -7.7190   -7.6314   -8.0419   -7.4724
   -7.5430   -8.0124   -8.0636   -7.5043   -7.4867   -7.4592   -7.3716   -7.7821   -7.2126
   -7.7096   -7.6683   -7.8250   -7.5377   -7.3792   -7.3273   -7.5367   -7.6873   -7.6089
   -7.8449   -7.8662   -8.0136   -7.7363   -7.5484   -7.5190   -7.6765   -7.8119   -7.8138
   -7.4826   -7.7971   -7.8832   -7.6733   -7.3154   -7.4083   -7.3286   -7.4172   -7.7992
   -7.9847   -7.6705   -7.4993   -7.9370   -7.6196   -7.6854   -7.4219   -7.6379   -7.8582
   -7.8733   -7.5905   -7.5571   -7.7962   -7.6472   -7.8384   -7.4399   -7.8740   -7.9074
   -7.8993   -7.5834   -7.6319   -7.6893   -7.5785   -7.6759   -7.5762   -8.1616   -7.9261
   -8.1255   -7.7993   -7.8000   -7.7276   -7.5040   -7.3819   -7.9103   -8.1564   -7.9852
   -7.7621   -8.0802   -7.8258   -7.9399   -7.6392   -7.4589   -7.3769   -7.8526   -7.9914
   -7.7300   -8.5350   -7.9089   -7.6384   -7.6994   -7.9599   -7.5814   -7.8397   -7.6895
   -7.6522   -7.8201   -7.6639   -7.6364   -7.7604   -7.7712   -7.5869   -7.7266   -7.9302
   -7.6981   -7.7280   -7.4370   -7.8803   -7.5173   -7.7583   -7.8069   -7.5889   -7.8799
   -7.7545   -7.8221   -7.6468   -8.0813   -7.5945   -7.6563   -8.2134   -7.7197   -7.7688
   -7.5493   -7.7584   -7.8859   -7.8668   -7.6579   -7.9813   -7.7588   -7.3998   -7.5711
   -7.5112   -7.4406   -7.9631   -7.9072   -7.4992   -7.8046   -7.5360   -7.3499   -7.7822
   -7.6558   -7.4183   -7.6381   -7.6817   -7.3401   -7.6447   -7.3778   -7.4349   -7.9298
   -7.5685   -7.5228   -7.7537   -7.6410   -7.8595   -7.6412   -7.7407   -7.7161   -7.6370
   -7.5373   -7.7039   -7.6253   -7.8622   -7.9119   -7.6477   -7.5100   -7.7788   -7.6027
   -7.5055   -7.7191   -7.8045   -7.8082   -7.6741   -7.8485   -7.3796   -7.4598   -7.6151
   -7.8135   -7.5540   -7.5577   -7.5333   -7.5676   -7.5742   -7.4482   -7.7409   -7.6295
   -8.0336   -7.6672   -7.5799   -7.8049   -7.4528   -7.7092   -7.7578   -7.7023   -7.8274
   -7.5976   -7.6473   -7.9562   -7.9630   -7.6552   -7.6538   -7.4436   -7.5430   -7.5607
   -7.8030   -7.7427   -7.6912   -7.7595   -7.8230   -7.6184   -7.3524   -7.4094   -7.9195
```

Fig.6. The vector of feature values obtained through the discrete cosine transform (DCT)

Discrete Cosine Transform (DCT) is applied on the log energy to have different Mel-scale cepstral coefficients. The DCT converts the signal from the frequency domain into a time domain. Because, the features are similar to cepstrum, it is referred to as the Mel-scale cepstral coefficients. MFCC can be used as the feature for speech recognition. For better performance can generated by adding the log energy and perform delta operations. As new features in MFCC, Delta cepstrum can be generated which has advantages in the time derivatives of the energy of the signal. It can use for finding the velocity and acceleration of energy with MFCC. Mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT). Therefore, we can calculate the MFCC's.

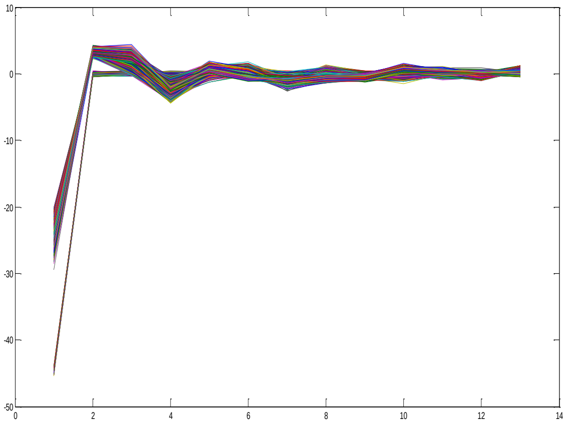


Fig.7. Mel-frequency cepstral coefficient of audio digit 'ο' collected from the minors

4.2 VECTOR QUANTIZATION

Vector quantization (VQ) is used for comparing the trained data with new entered input data. It is a classical Quantization technique that allows the modeling of probability density functions by the distribution of vectors. It divides a large set of points called vectors into groups having approximately the same number of points closest to them. The density matching property of VQ is powerful for identifying the density of large and highdimensioned data [15] [20]. All information points are mapped by the indicator of their closest centroid [15] [16]. The training datasets for the VQ are obtained by recording Gujarati digits. The recorded data were compared with already stored data sets. VQ is an operation of mapping vectors from a large vector space into a finite number of regions in that space. Each region is called a cluster and can be represented by its center called codeword. The collection of all code words is called a codebook. The Fig.7 presents a conceptual diagram to illustrate this recognition process, In Fig.7, Only two speakers and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from the speaker 1 while the triangles are from the speaker 2 [16]. After the enrollment session, the acoustic vectors extracted from input speech of a speaker provide a set of training vectors. As described above, the next important step is to build a speaker-specific VQ codebook for this speaker using those training vectors. There is a well-know algorithm, namely LBG algorithm [15][16], for clustering a set of L training vectors into a set of M codebook vectors.

The algorithm is formally implemented by the following recursive procedure that is illustrated in Fig.8 which shows the detailed steps of the LBG algorithm.

- *Design a 1-vector codebook:* this is the centroid of the entire set of training vectors (hence, no iteration is required here).
- Double the size of the codebook by splitting each current codebook y_n according to the rule where *n* varies from 1 to

the current size of the codebook, and ε is a splitting parameter (we choose $\varepsilon = 0.01$).

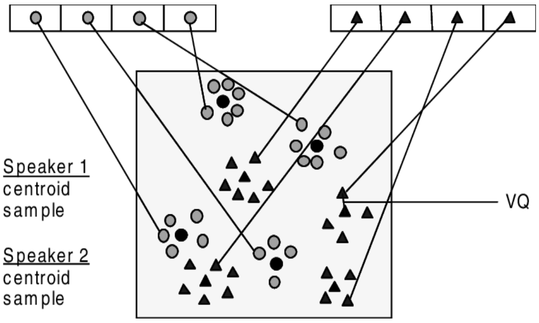


Fig.8. Vector quantization codebook formation [15] [16]

- *Nearest-Neighbor Search*: for each training vector, find the codeword in the current codebook that is closest (in terms of similarity measurement), and assign that vector to the corresponding cell (associated with the closest codeword) [15] [16].

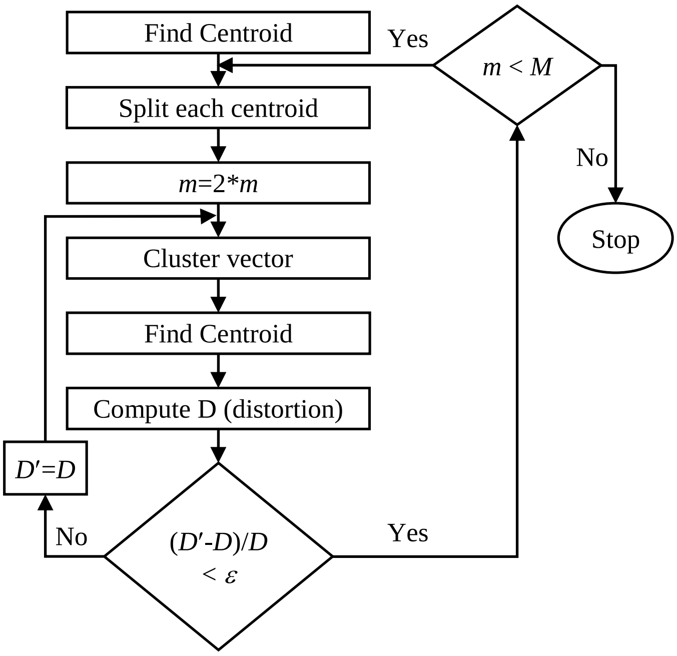
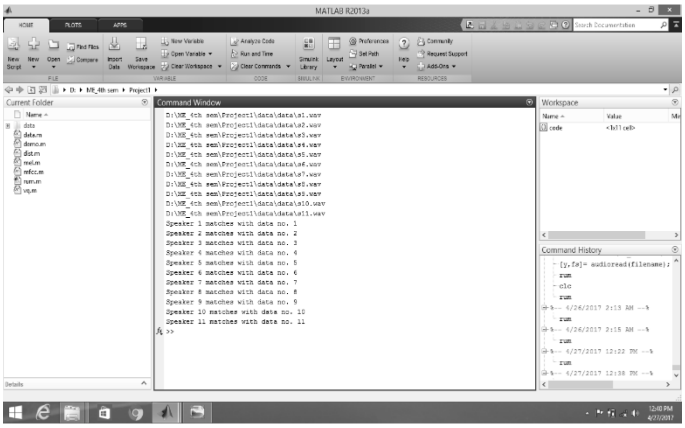


Fig.9. Flow diagram of the LBG algorithm [15][16]

- *Centroids Update*: update the codeword in each cell using the centroids of the training vectors assigned to that cell.
- *Iteration 1*: repeat steps 3 and 4 until the average distance falls below a preset threshold 6. Iteration 2: repeat steps 2, 3 and 4 until a codebook size of *M* is designed.

Intuitively, the LBG algorithm designs an M-vector codebook in stages. It starts first by designing a 1-vector codebook, then uses a splitting technique on the code words to initialize the search for a 2-vector codebook, and continues the splitting process until the desired M-vector codebook is obtained [15] [16]. Cluster vectors is the nearest-neighbor search procedure which assigns each training vector to a cluster associated with the closest codeword. “Find centroids” is the centroid update procedure. Compute D (distortion) sums the distances of all the training vectors in nearest-neighbor search so as to determine whether the procedure has converged. The tested results are shown at a lower place in Fig.8, where new input speech signals matches with the stored trained speech data signal.



```
D:\ME_4th sem\Project1\data\data\s1.wav
D:\ME_4th sem\Project1\data\data\s2.wav
D:\ME_4th sem\Project1\data\data\s3.wav
D:\ME_4th sem\Project1\data\data\s4.wav
D:\ME_4th sem\Project1\data\data\s5.wav
D:\ME_4th sem\Project1\data\data\s6.wav
D:\ME_4th sem\Project1\data\data\s7.wav
D:\ME_4th sem\Project1\data\data\s8.wav
D:\ME_4th sem\Project1\data\data\s9.wav
D:\ME_4th sem\Project1\data\data\s10.wav
D:\ME_4th sem\Project1\data\data\s11.wav
Speaker 1 matches with data no. 1
Speaker 2 matches with data no. 2
Speaker 3 matches with data no. 3
Speaker 4 matches with data no. 4
Speaker 5 matches with data no. 5
Speaker 6 matches with data no. 6
Speaker 7 matches with data no. 7
Speaker 8 matches with data no. 8
Speaker 9 matches with data no. 9
Speaker 10 matches with data no. 10
Speaker 11 matches with data no. 11
>>
```

```
[y,fs]= audioread(filename);
run
clc
run
%-- 4/26/2017 3:13 AM --%
run
%-- 4/26/2017 3:18 AM --%
run
%-- 4/27/2017 12:22 PM --%
run
%-- 4/27/2017 12:38 PM --%
```

Fig.10. Testing of Input with Database of 0 to 10 Gujarati digits using VQ

5. COMPARATIVE ANALYSIS AND DISCUSSION OF RESULTS

The graphical representation of Fig.11 shows the comparative results between Gujarati digits with different speakers using VQ technique. From this comparative analysis part we observe that the speaker identification system for Gujarati isolated digit achieve higher recognition rate for identification of speaker for Gujarati digits, where speakers are from different age group that achieves higher accuracy rate, all these four speakers and their results are compared in Fig.9 with Gujarati digit, which achieve 90-95% higher recognition rate for the identification of the speaker for Gujarati digit 1, 5, 6 and 8 among each speaker.

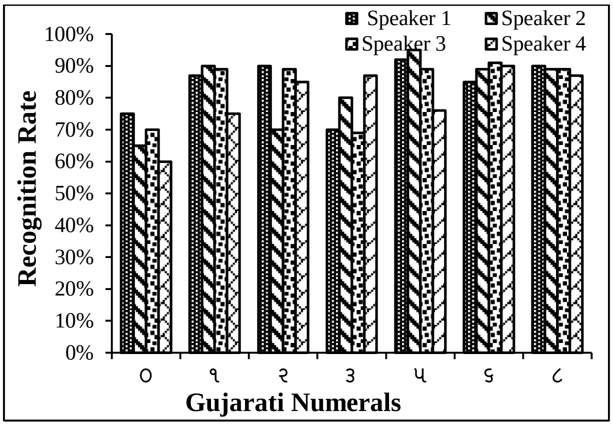


Fig.11. Comparative Results

The graphical representation of Fig.12 shows the overall accuracy of MFCC, DTW, HMM, VQ, KNN and the proposed technique with the combination of MFCC and VQ. Which also compare the accuracy results achieved in English and other similar languages with the comparison of the proposed database of Gujarati digits. Which achieve 90-95% higher recognition rate for the identification of the speaker with the Gujarati language as compared to the other languages. Based on this analysis and discussion, it is observed that the combination of MFCC and VQ achieves good amount of results as compared with other techniques.

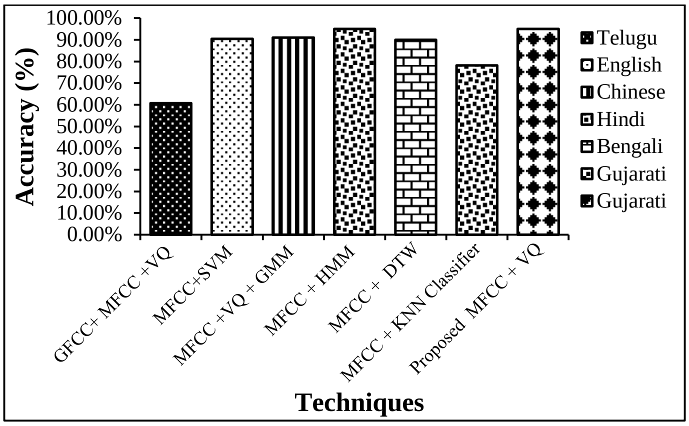


Fig.12. Comparison of Proposed Technique with Other Techniques and Languages

6. CONCLUSION AND FUTURE WORK

In this research, an approach is implemented for the speaker identification system for Gujarati isolated digit. Which is applied to some Gujarati Digit's datasets of different age group with the identification of the speaker. Which achieves 90% of accurate higher recognition rate for the identification of the speaker with the Gujarati digit 1, 5, 6 and 8 among each speaker. MFCC here used to extract features from speech signals. Here, It is implemented as feature extraction technique and applied on some Gujarati digits like '0', '1', '2' etc. As the results show MFCC achieves no of Mel-coefficients value per each frame of the wav file. VQ method is employed as a pattern matching technique for testing of input data of Gujarati 0 to 10 digits that data were compared with already stored datasets of each speaker. As a result, there is need to develop more approaches. Because there are different types of problems arise in speech recognition like variability, including speaker-generated variability and variability in channel and recording conditions. It is very important to investigate feature parameters that are stable over time, insensitive to the variation of speaking manner, including the speaking rate and level, and robust against variations in voice quality due to causes such as voice disguises or colds and much more. So, the future endeavors will be to apply to the database with a large vocabulary of Gujarati digit with speech to text recognition with the combination of VQ and other pattern matching techniques. The MFCC can be used for best feature extraction to improve the hardiness of the speaker identification system later on.

REFERENCES

[1] H.B. Chauhan and B.A. Tanawala, "Comparative Study of MFCC & LPC Algorithms for Gujarati Isolated Word Recognition", *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 3, No. 2, pp. 1-4, 2012.

[2] Parwinder Pal Singh and, Pushpa Rani, "An Approach to Extract Feature using MFCC", *IOSR Journal of Engineering*, Vol. 4, No. 8, pp. 21-25, 2014.

[3] A. Ghadee Ganesh, B. Jonvale and Ratnadeep.R. Deshmukh, "Speech Feature Extraction Using Mel-Frequency Cepstral Coefficient (MFCC)", *Proceedings of International Conference on Emerging Trends in Computer Science, Communication and Information Technology*, pp. 23-28, 2010.

[4] A. K. Kumbharana, "Speech Pattern Recognition for Speech to Text Conversion", Ph.D. Dissertation, Department of Computer Science, Saurashtra University, 2007.

[5] M.A. Anusuya and S.K. Katti, "Speech Recognition by Machine: A Review", *International Journal of Computer Science and Information Security*, Vol. 6, No. 3, pp. 1-25, 2009.

[6] Shall Gujral, Monika Tuteja and Baljit Kaur, "Various Issues In Computerized Speech Recognition Systems", *International Journal of Engineering Research And General Science*, Vol. 2, No. 4, pp. 1-6, 2014.

[7] Neha Chadha, R.C. Gangwar and Rajeev Bedi, "Current Challenges and Application of Speech Recognition Process using Natural Language Processing: A Survey", *International Journal of Computer Applications*, Vol. 131, No. 1, pp. 1-7, 2015.

[8] Maruti Limkar, Rama Rao and Vidya Sagvekar, "Isolated Digit Recognition using MFCC & DTW", *International Journal of Advanced Electrical and Electronics Engineering*, Vol. 1, No. 1, pp. 59-64, 2012.

[9] Therese S Shanthi and S.C. Lingam, "Speaker Based Language Independent Isolated Speech Recognition System", *Proceedings of IEEE International Conference on Communication, Information and Computing Technology*, pp. 15-17, 2015.

[10] Sarika S. Admuthe and Shubhada Ghugardare, "Survey Paper on Automatic Speaker Recognition Systems", *International Journal of Engineering and Computer Science*, Vol. 4, No. 3, pp. 10895-10898, 2015.

[11] A. Revathi and Y. Venkataramani, "Speaker Independent Continuous Speech and Isolated Digit Recognition Using VQ & HMM", *Proceedings of International Conference on Communications and Signal Processing*, pp. 198-202, 2011.

[12] A. Choudhary, R.S. Chauhan and G. Gupta, "Automatic Speech Recognition System for Isolated and Connected Words of Hindi Language By using Hidden Markov Model Toolkit (HTK)", *Proceedings of International Conference on Emerging Trends in Engineering and Technology*, pp. 1-7, 2013.

[13] S.V. Chapaneri and D.J. Jayaswal, "Efficient Speech Recognition System for Isolated Digits", *International Journal of Computer Science and Engineering Technology*, Vol. 4, No. 3, pp. 228-236, 2013.

[14] C. Patel Bharat and A. Desai Apurva, "Recognition of Spoken Gujarati Numeral and Its Conversion into Electronic Form", *International Journal of Engineering Research and Technology*, Vol. 3, No. 9, pp. 1-5, 2014.

[15] D. Gupta, M. Radha, M. Navya and P.B. Manoj. "Isolated Word Speech Recognition using Vector Quantization (VQ)", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 2, No. 5, pp. 12-18, 2012.

[16] Ch Srinivasa Kumar and P. Mallikarjuna Rao, "Design of an Automatic Speaker Recognition System using MFCC,

Vector Quantization and LBG Algorithm", *International Journal on Computer Science and Engineering*, Vol. 3, No. 8, pp. 29-42, 2011.

[17] Tushar Ratanpara and Narendra Patel, "Singer Identification using Perceptual Features and Cepstral Coefficients of an Audio Signal from Indian Video Songs", *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 16, pp. 1-15, 2015.

[18] Lei Lei and She Kun, "Speaker Recognition using Wavelet Cepstral Coefficient, I-Vector, and Cosine Distance Scoring and Its Application for Forensics", *Journal of Electrical and Computer Engineering*, Vol. 2016, pp. 1-16, 2016.

[19] Kumari Rambha Ranjan, Kartik Mahto, Dipti Kumari and S.S. Solanki, "Singer Identification using MFCC and LPC and its Comparison for ANN and Naive Bayes Classifier", *International Journal of Latest Engineering Research and Applications*, Vol. 2, No. 4, pp. 25-30, 2017.

[20] M. Mallikarjunan, P.Karmali Radha, K.P. Bharath and Rajesh Kumar Muthu, "Text-Independent Speaker Recognition in Clean and Noisy Backgrounds using Modified VQ-LBG Algorithm", *Circuits, Systems and Signal Processing*, pp. 1-18, 2018

[21] Kritagya Bhattarai et al., "Experiments on the MFCC Application in Speaker Recognition using Matlab", *Proceedings of IEEE 7th International Conference on Information Science and Technology*, pp. 1-7, 2017.

[22] Pooja Prajapati and Miral Patel, "Feature Extraction of Isolated Gujarati Digits with Mel Frequency Cepstral Coefficients (MFCCs)", *International Journal of Computer Applications*, Vol. 163, No. 6, pp. 29-34, 2017.

[23] Safaa Elouahabi, Mohamed Atounti and Mohamed Bellouki, "Amazigh Isolated Word Speech Recognition System using Hidden Markov Model Toolkit (HTK)", *Proceedings of International Conference on Information Technology for Organizations Development*, pp. 1-5, 2016.