Feature Extraction of Isolated Gujarati Digits with Mel Frequency Cepstral Coefficients (MFCCs)

Pooja Prajapati
PG Student
Information Technology Department,
G H Patel Collage of
Engineering and Technology, India

Miral Patel
Professor
Information Technology Department,
G H Patel Collage of
Engineering and Technology, India

ABSTRACT

The aim of this paper is to present feature extraction method of Gujarati isolated digit for speaker identification using Mel-Frequency Cepstral Coefficient (MFCC). The objective of MFCC is to extract features that are present in speech signal. That produces Mel-coefficients of speech data which helps in representing speaker specific characteristics, thus this technique is one of the best technique for feature extraction especially for automatic speech & speaker recognition system. This can offer better security than keypad input system at the ATM, cashless system, mobile password, etc. The proposed approach helps to implement the speaker identification system. Where dataset of Gujarati numeral (0 to 10) was recorded from different speakers from different age groups. This paper presents approach for extracting features from the speech signal of spoken words using the Mel-Scale Frequency Cepstral Coefficients. All this implementation is built in MATLAB environment. The result describes how it transform the input waveform into a sequence of acoustic feature vectors.

General Terms

Feature extraction, Isolated Gujarati digit, MFCC, Speaker Identification

Keywords

Feature extraction, Isolated Gujarati digit, MFCC, Speaker Identification

1. INTRODUCTION

Speech is a common & one of the most important ways of human computer interaction. Voice is a natural way of communication & non-intrusive as a biometric [5]. Like fingerprints, it carries the identity of the speaker as voice print. The Human voice is a one type of signal that contains a mixed type of information, including words, feelings, language & 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, Voiceprint is a safe & a secure method for authenticating an individual's identity that unlike passwords or token, these characteristics of voice cannot be stolen, duplicated or forgotten [10] [11]. In recent years, a good amount of work has been contributed to English & other Indian languages; out of these Gujarati language has the least amount of work [14]. So, it's choosing to develop 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 as MFCC is a heart of the ASR system [2]. The main objective of MFCC is solving the issue of pattern classification. Most of the feature extraction techniques use low level spectral information which conveys vocal tract characteristics. Out of this all techniques, MFCC (Mel frequency Cepstral coefficient) is a standard & most widely used technique in Speaker identification system [1]. MFCC is a popular also due to efficient computation schemes available for it & its robustness in the presence of different noises.

The objective of our proposed study is to discuss the functioning of a MFCC computation process. Which may help a lot to the researchers to get the overview about the MFCC process and the result analysis achieved by that system helps researchers in choosing & understanding the technique that how MFCC is a better feature extractor for speech signals.

This paper is divided into four sections. Section 1 gives an introduction, Section 2 discussed Experimental Setup Section 3 indicates the methodology of proposed work & Experimental Results with each step and Section 4 followed by Conclusion & future work.

2. 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.

2.1 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 & 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 & 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

3. METHODOLOGY OF PROPOSED WORK

Mel-Frequency Cepstral Coefficient (MFCC) is one of the best technique for feature extraction, especially for automatic speech & 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]. MFCC gives more efficient and accurate result, than other feature extraction techniques in the voice identification system [3]. MFCC is used in speaker identification with speaker information like, contents and channels. 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. This paper performed the feature extraction approach for speaker identification system with isolated Gujarati numerals. The results of each step of this process are shown and described in each of the following sections.

3.1. Feature extraction using MFCC

Mel Frequency Cepstral Coefficients (MFCCs) are coefficients that represent audio. They derive from a type of cepstal representation of the audio file. The difference between the cepstrum & 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 Figure 1.

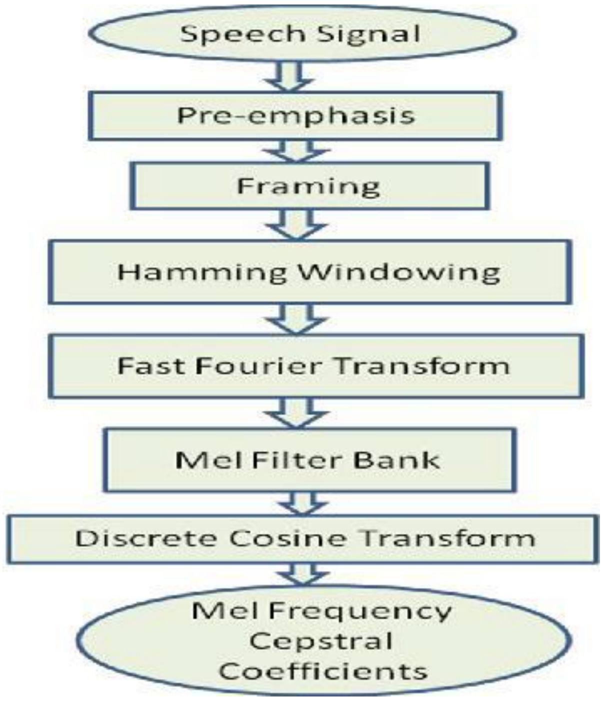


Fig 1: Block diagram of MFCC computation process

As shown in above Figure 1 Basic steps in the computation process of exracting cepstral coefficients from the speech signal are positioned below.

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

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

3.1.1. 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 to do some computation of high frequency information that is known as pre emphasis. In speech there is 1st order high pass filter.

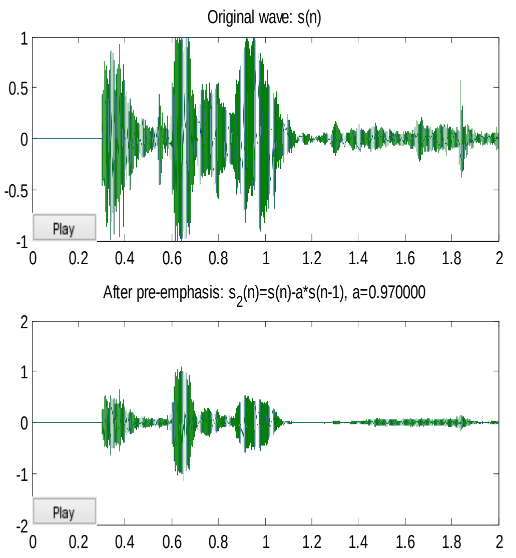


Fig 2: The pre-emphasis wav for digit ‘O’ collected from the minors

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

S1 (n) = S(n) – α S(n - 1)

Where s (n) is the speech signal & 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 figure 2.

3.1.2. 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 ~ 30 msec windowing and facilitate the fast radix-2 FFT) and M = 100. For each frame 20 mfcc were calculated.

3.1.3. 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 it defines 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.

$$W[K+1] = 0.54 - 0.46\cos\left[2\pi\frac{k}{n-1}\right],k = 0,\ldots\ldots,n-1$$

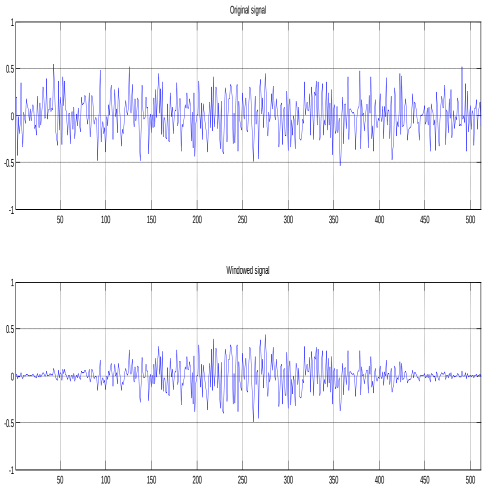


Fig 3: Windowed Frame on audio digit ‘O’ collected from the minors

The simulation of Windowed Frame on audio digit ‘O’collected from the minors is shown in below figure 3.

3.1.4. 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.

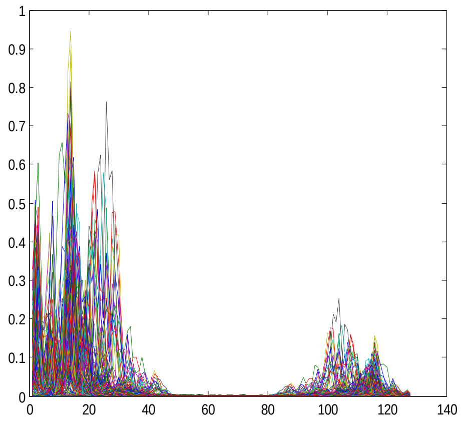


Fig 3: Detailed FFT magnitude of audio digit ‘O’ collected from the minors

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 ‘૦’collected from the minors is shown in below figure 4.

3.1.5. Mel Filter Bank

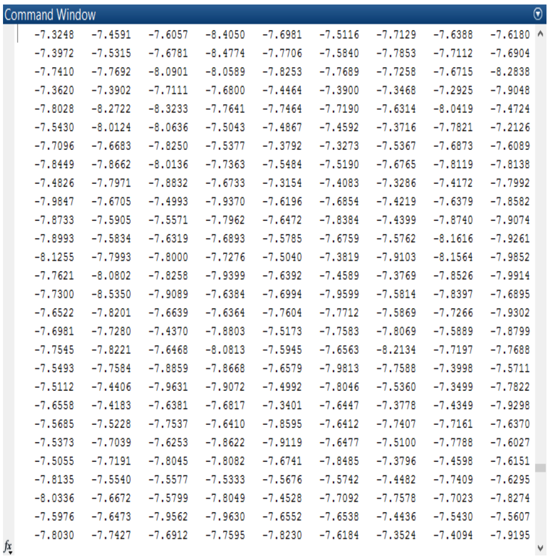
Mel (melody) is a unit of pitch. The spectrum obtained from the above step is Mel Frequency Wrapped. The Mel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold,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

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

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’.

3.1.6. Discrete Cosine Transform & 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 5: The vector of feature values obtained through the discrete cosine transform (DCT)

Discrete Cosine Transform (DCT) is applied to the large 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.

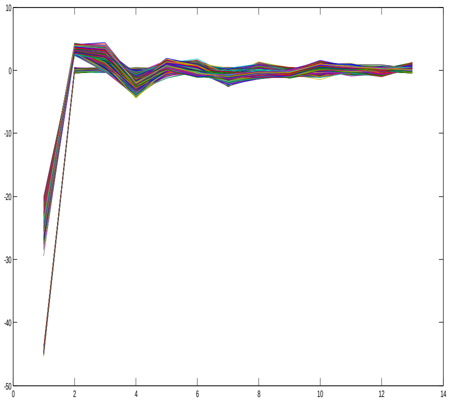


Fig 6: Mel-frequency cepstral coefficient of audio digit ‘૦’ collected from the minors

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, It can convert them to the time domain using the Discrete Cosine Transform (DCT). Thus, It can calculate the MFCC's.

4. CONCLUSION AND FUTURE WORK

In this research, the system implements for the Speaker identification system for Gujarati numerals. In which vector of feature values has been obtained through the discrete cosine transform (DCT). That demonstrates how using MFCC, it generate so many Mel vector per each frame. That helps in proving how the MFCC feature extraction technique is more effective and robust, and with the aid of this technique normalizes of the features as well and it is a quite popular technique for speech & speaker recognition. In future, The combination of these features with MFCCs will be utilized to improve the hardiness of the speaker identification system later on. VQ method will be applied to improve the efficiency & precision of the work.

5. ACKNOWLEDGMENT

Pooja Prajapati remains thankful to Miral Patel (Department Of Information Technology, GCET) for their useful discussions & suggestions during the preparation of this technical paper.

6. REFERENCES

[1] H. B. Chauhan, Prof. B. A. Tanawala, (FEB-2015), Comparative Study of MFCC & LPC Algorithms for Gujarati Isolated Word Recognition, IJIRCCE, vol.3, issue 2

[2] Parwinder Pal Singh, Pushpa Rani, (August. 2014), An Approach to Extract Feature using MFCC, IOSR Journal of Engineering, Vol. 04, Issue 08, PP 21-25

[3] A. Ghadee Ganesh B. Jonvale, and Ratnadeep.R.Deshmukh, (January 2010), Speech Feature Extraction Using Mel- Frequency Cepstral Coefficient (MFCC), Conference Paper of Emerging Treads in computer science, communication & information technology

[4] A. K. Kumbharana, (2007), Speech Pattern Recognition for Speech To Text Conversion, *etheses*

.saurashtrauniversity . edu /337/1/ kumbharana _ck_ thesis _cs .pdf by CK Kumbharana

[5] M.A.Anusuya, S.K.Katti, (2009), Speech Recognition By Machine: A Review, IJCSIS, Vol. 6, No. 3

[6] Shall Gujral , Monika Tuteja , Baljit Kaur , (June-July 2014), Various Issues In Computerized Speech Recognition Systems, International Journal Of Engineering Research And General Science, Volume 2, Issue 4

[7] Neha Chadha, R.C. Gangwar, Rajeev Bedi, (Dec-2015), Current Challenges and Application of Speech Recognition Process using Natural Language Processing: A Survey, IJCA,vol.131

[8] MarutiLimkara, RamaRaob & Vidya Sagvekarc, (2012), Isolated Digit Recognition Using MFCC & DTW, IJAEEE, vol.1, issue 1 , pp (59-64)

[9] Therese S Shanthi, Lingam Chelpa, (IEEE 2015), Speaker Based Language Independent Isolated Speech Recognition System

[10] Miss. Sarika S. Admuthe, Mrs. Shubhada Ghugardare, (March 2015), Survey Paper on Automatic Speaker Recognition Systems, International Journal Of Engineering And Computer Science, Volume 4 Issue 3, Page No. 10895-10898

[11] Revathi A, Venkataramani Y, (IEEE 2011), Speaker Independent Continuous Speech & Isolated Digit Recognition Using VQ & Hmm, pp 198- 202

[12] Choudhary Annu, Chauhan R, S, Gupta Gautam, (ACEEE 2013), Automatic Speech Recognition System for Isolated & Connected Words of Hindi Language By Using Hidden Markov Model Toolkit (HTK), Association of Computer Electronics and Electrical Engineers,

[13] Chapaneri , Santosh V, Jayaswal, Deepak J (2013), Efficient Speech Recognition System For Isolated Digits, IJCSET, vol.4, issue 3, pp 228-236

[14] Patel Bharat C, Desai Apurva A, (2014), Recognition of Spoken Gujarati Numeral and Its Conversion into Electronic Form, IJERT, vol.3, issue 9

7. AUTHOR PROFILE

Pooja Prajapati completed her Bachelor of information technology from Gujarat Technological University, Gujarat, India. She is currently a research student doing her Master of information technology in G H. Patel college of Engineering & Technology, Gujarat Technological University, Gujarat, India. Her area of interest includes speech recognition & Artificial Intelligence

Miral Patel completed her Bachelor of Computer Engineering from Birla Vishwakarma Mahavidyalaya, Sardar Patel University, Gujarat, India and Master of Information Technology from G H. Patel College of Engineering and Tech., Gujarat Technological University (GTU). She is currently working toward the Ph.D. degree from Changa University, Gujarat, India and have more than more than 10 Years of experience in industry and academic institutes. She has visited USA under the Gate government Project, Chicago, IL. Her Area of interest includes natural language processing, Software engineering, Project management, and Artificial intelligence.