ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 9, March 2014

Development of Automatic Speech Recognition of Marathi Numerals - A Review

Yogesh K. Gedam, Sujata S. Magare, Amrapali C. Dabhade, Ratnadeep R. Deshmukh
Dept. of Computer Science and IT, Dr.B.A.M. University, Aurangabad

Abstract*— Speech is the way of communication between human beings. This paper describes the development of automatic speech recognition of Marathi numerals i.e. Shunya (Zero) to Nau (Nine). Feature extraction and feature matching technique plays an important role for speech recognition. We used Mel Frequency Mel cestrum (MFCC) technique for feature extraction and for dynamic time wrapping (DTW) technique for feature matching. Vector Quantization is used to minimize the data of the extracted feature. Total database collection is about 100 speakers with the help of high performance Headsets and PRAAT software for data recording. Data is divided into male and female speakers. It is recorded in noisy environment. Noise removal technique is used then feature extraction and feature matching technique is applied. To deals with the different speaking speed, DTW technique is used.

Index Terms—**Automatic Speech Recognition, DTW, Mel Frequency Mel Cestrum (MFCC), Vector Quantization**

I. INTRODUCTION

Speech is the way of communication between human beings. Human are interact or communicate with each other with the help of speech. A little work has been done for Indian languages compared to non-Indian languages. Many researchers around the world are trying to develop new interface system for communication between human and computer with maximum accuracy. Speech has potential of being important mode of interaction with computer. Automatic Speech recognition is the way of processing a speech signals into sequence of words or into text form. With the speech recognition system the voice of speech data spoken by human being are converted into electrical signals. These signalsare then transfers into a coding pattern and desirable meaning is obtained. Communication of human being with environment is obtained by sending out the signals or information in the form of sight, audio etc. Speech recognition is basically process of automatically identifying the individual speaker with the help of information in speech waves [1]. From last sixty years, researchers work for speech recognition. Now ASR system today finds an application that requires human machine interface and can speak and recognize the speech in its native languages [2].

II. ABOUT MARATHI LANGUAGE

Marathi language belongs to the group of Indo-Aryan language which is a part of the largest of group of Indo-European languages,all of which can be traced back to a common root which is official language of Maharashtra state in India. All Indo-Aryan language originates from Sanskrit. Three Prakrit languages i.e., Saurseni, Magadhi and Maharashtri are simpler in structure, emerged from Sanskrit. Marathi language uses Devanagari script.The number of Marathi speakers all over the world is close to 72 million [4]. A character represents one vowel and zero or more consonant Marathi language contains total 12 vowels and 36 consonants [5]. The universities like Maharaja Sayajirao University of Baroda (Gujarat), Osmania University (Andhra Pradesh), Gulbarga University (Karnataka), Devi Ahilya University of Indore and Goa University (Panaji) have separate departments for Marathi language. Hindi is written in Devanagari script while other17 languages recognized by the constitution of India are: 1) Assamese 2) Tamil 3) Malayalam 4) Gujarati 5) Telugu 6) Oriya 7) Urdu 8) Bengali 9) Sanskrit 10) Kashmiri 11) Sindhi 12) Punjabi 13) Konkani 14) Marathi 15) Manipuri 16) Kannada and 17) Nepali [6]. Marathi is said to be a descendant of Maharashtri which was Prakrit spoken by people residing in region of Maharashtra. Other than Sanskrit, Marathi has also been influenced by the languages of its neighboring states which are Kannad (state of Karnataka) and Telugu (state of Andhra Pradesh). The script currently used in Marathi is called 'Balbodh' which is a modified version of Devanagari script. Earlier, another script called 'Modi' was in use till the time of the Peshwas (18th century). This script was introduced by Hemadpanta, a minister in the court of the Yadava kings of Devgiri (13th century). This script looked more like today's Dravidian scripts and offered the advantage of greater writing speed because the letters could be joined together. Today only the Devanagari script is used which is easier to read but does not have the advantage of faster writing.

III. DATABASE COLLECTION

The basic requirement for developing a speech database is of correct Text corpus which would be recorded from various speakers from the native place of Aurangabad region.The text corpus designed should be grammatically correct. Marathi digits are selected from Zero (Shunya) to Nine (Nau). Text corpus designed should be grammatically correct. The number of speaker should be collected from native speaker of Marathi language from Marathwada region. For data collection, ‘PRAAT’ software and ‘Sennheiser PC360’ and ‘Sennheiser PC350’ headset, used for recording the speech samples. The sampling frequency was set to 24000 Hz with Mono sound type and the recorded sample was saved as ‘.wav’ file. The PC360 and PC350 headsets are having noise cancellation facility and the signal to noise ratio (SNR) is less. PRAAT is a very flexible tool to do speech analysis. It offers a wide range of standard and non-standard procedures, including spectrographic analysis, articulator synthesis, and neural networks [7]. Number of speakers selected are 100. The selected speakers were 50 males and 50 females. Each speaker is asked to speak the numbers Shunya (zero) up to Nau (nine) i.e. ten numbers with three utterances of each number. Total data collected is about 3000.The data collected is divided into male and female speakers.

IV. APPROACHES TO SPEECH RECOGNITION

It contains following approaches

- Acoustic Phonetic Approach
- Pattern Recognition Approach
- Knowledge Based Approach
- Connectionist Approach
- Support Vector Machine (SVM)

Acoustic phonetic approaches

It is basically depends upon theory of acoustic phonetics and postulate [8]. The basis of acoustic phonetic approach is basically on finding the speech sound and providing appropriate labels to it (Hemdal and Hughes, 1967). It also postulates that there exist finite, distinctive phonetic units in spoken language. These units are broadly characterized by a set of acoustic properties.

Pattern Recognition Approach

It involves two essential steps namely pattern training and pattern comparison [9] [10] [11]. It uses well formulated mathematical framework and established consistent speech pattern representation from set of labeled training samples via formal training algorithm. Speech pattern recognition can be in the form of speech template o statistical model (HMM) and can be applied to a sound, a word, a phrase etc.

Knowledge Based Approach

Artificial intelligence approach a hybrid of the acoustic phonetic approach and pattern recognition approach and it attempts to mechanize the recognition procedure according to way of person applies intelligence in visualizing and characterizing speech based on a set of measured acoustic features.Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram

Connectionist Approach (Artificial Neural Networks)

The artificial intelligence approach (Lesser et al. 1975; Lippmann 1987) attempts to mechanize the recognition procedure according to the way a person applies intelligence in visualizing, analyzing, and characterizing speech based on a set of measured acoustic features.

Support Vector Machine (SVM)

Support Vector Machines were first introduced by Vapnik [12]. It represents a new approach to pattern classification. It has great ability to generalized often resulting in better performance than traditional technique like Artificial Neural Network (ANN). Support vector machine can be used as a regularized radial basis function classifier.

V. PROPOSED APPROACH

This paper is used to recognize the Marathi digits from different speakers starting from Shunya (zero) to Nau (nine). For speech recognition, feature extraction (MFCC), feature matching technique (DTW) is used.

FEATURE EXTRACTION

Feature extraction technique is process of removing redundant and unwanted information from the speech signals and retaining only useful information [14]. Sometimes, important data may be loss, while removing unwanted information.It involves analysis of the speech signals [15] and may involve transforming the signal into a form appropriate for the models used for classification. The most commonly used features in speech nowadays are the Mel Frequency Cepstral Coefficients (MFCC), in which the frequency bands are positioned logarithmically along the Mel scale, andthe Perceptual Linear Predictive (PLP), in which the bark scale and an all-pole model is used[16][17].

Mel Frequency Coefficient Ceptrum (MFCC)

MFCC is based on human peripheral auditory system and cannot perceive frequencies over 1KHz. MFCC are based on the short-term analysis, and thus a MFCC vector is computed from each frame. MFCC have two types of filter which are spaced at low frequency below 1000 Hz linearly and logarithmic spacing above 1000 Hz. As reference point, Pitch of 1 KHz tone, above 40 dB the perceptual hearing threshold is defined as 1000 Mels [18].

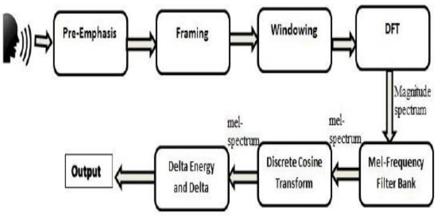

Fig 1: Steps for MFCC technique [19]

A. *Preprocessing*

Speech signals are normally preprocessed before feature extraction to enhance the accuracy and efficiency of feature extraction. Speech signal in the form of wave has suffer from noise hence it is necessary to remove it. Hence to reduce spectrally flatten speech signal, pre-emphasis is applied. It uses first order high pass FIR filter to preemphasize the higher frequency component.

B. *Framing and windowing*

To analyze a signal at one time is difficult task; hence speech signal is split into frames of size range 0-20 ms in time domain and analyses in short time. After that overlapping is applied to it because on each individual frame, hamming window is applied. In speech recognition, the mostcommonly used window shape is the hamming windowshown in “(1)”.

Hamming window:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{L-1}\right) \qquad 0 \geq n \geq L-1$$

$$0 \qquad\qquad\qquad\qquad otherwise$$

…… (1)

Hamming window gets rid of some of the information at the beginning and end of each frame.To avoid the risk of losing the information from the speech signal, the frame is shifted 10 ms so that the overlapping between two adjacent frames is 50%.The windowing function is applied after dividing the signal into frames that contain nearly stationary signal blocks [19].

C. *Fourier Transform*

DFT is used to convert each frame of N samples from time domain to frequency domain. To obtain a good frequency resolution, a 512 point Fast Fourier Transform (FFT) is used. It is used to convert the convolution of glottal pulse and vocal track in time domain into frequency domain.

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-j2kn\frac{\mu}{n}} \qquad ……. (2)$$

Spectral analysis signifies that different timber in speech signals corresponds to different energy distribution over frequencies. So FFT is executing to obtain magnitude frequency response of each frame. It also prepares the signal for next stage [20].

D. *Mel-Frequency Filter Bank*

A filter bank is created by calculating a number of peaks, uniformly spaced in the Mel-scale and then transforming the back to the normal frequency scale where they are used a peaks for the filter banks. Equation as shown in “(3)” is used to compute the Mel for given frequency f in Hz.

$$F(mel) = 2595 * \log 10\left(1 + \frac{f}{700}\right) \qquad ……. (3)$$

E. *Discrete Cosine Transform*

It express the finite sequence of the data points in terms of sum of cosine function oscillating with different frequencies. It converts the log Mel spectrum into time domain and results into Mel Frequency Cepstral Coefficient and set is called as acoustic vectors. The Cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis [21]. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT).

F. *Delta Energy and Delta Spectrum*

The voice signal and the frames changes, so there is a need to add features related to the change in Cepstral features over time. The energy in a frame for a signal x in a window from time sample t1 to time sample t2, is represented as shown below in “(4)” here x[t] is signal,

$$Energy = \sum x^2[t] \qquad ……. (4)$$

FEATURE MATCHING

There are many feature-matching techniques used in speaker recognition such as Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization.

A. *Dynamic Time Warping*

This algorithm is based on dynamic programming and used for measuring the similarity between two time series which may vary in time. This technique also used to find the optimal alignment between two times series if one time series may be “warped” non-linearly by stretching or shrinking it along its time axis. This warping can then be used to find corresponding regions between the two time series or to determine the similarity between the two time series. It also aims at aligning two sequences of feature vectors by warping the time axis repetitively until an optimal match between the two sequences is found.

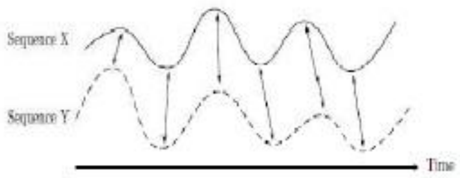

Fig 2: Time Alignment of two time independent sequences

For given an n- by- m matrix where the (i^{th},j^{th}) element of the matrix contains the distance $d(q_i, c_j)$between the two points q_i and c_j is constructed .The absolute distance between the values of two sequences is calculated using the Euclidean distance shown in “(5)”.

$$d(q_i, c_j) = (q_i - c_j)^2 \qquad ……. (5)$$

Each matrix element (i, j) corresponds to the alignment between the points and then, accumulated distance is measured by shown by “(6)”

.

$$D(i,j) = \min[D(i-1, j-1), D(i-1, j), D(i, j-1)] + D(i,j)$$

……. (6)

VECTOR QUANTIZATION (VQ)

The term vector quantization, which is also called "block quantization" or "pattern matching quantization" is often used in data compression. This technique is based on principle of block coding [23]. VQ is used for number of application like speech recognition, pattern recognition face detection, speech data compression, iris recognition, Content Based Image Retrieval (CBIR) etc. In vector quantization technique, the unique representation of each speaker is done in efficient way. Vector quantization is the process in which mapping of vectors is done from a large vector space to a finite number of regions in that space. Each region is known as a cluster and can be represented by its center known as a codeword. The collection of all code words is known as a codebook [24]. Each cluster or centroid represents a different class of speech signal in which data significantly compressed. If the feature vectors are not specifically quantized, the system would be too large and computationally complex. Vector quantization is works by encoding values from a multidimensional vector space into a finite set of values from a discrete subspace of lower dimension. A lower-space vector requires less storage space, so the data is compressed. The compressed data has errors that are inversely proportional to density. This is due to the density matching property of vector quantization. It is a fixed-to-fixed length algorithm and may be thought as an aproximator [25].

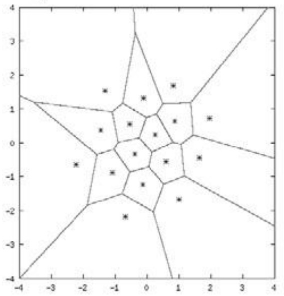

Fig 3: An example of a 2-dimensional VQ [25]

Here, every pair of numbers falling in a particularregion is approximated by a star associated with that region. In figure 3, the star showscodevectors and set of all codevectors is called the codebook. Regions are defined by the bordersare called encoding regions where as set of all encoding regions is called the partition of the space [26].

A. LBG Algorithm

It is an iterative algorithm which is proposed by Y. Linde, A. Buzo & R. Gray. It alternatively solves optimality criteria [27]. It requires an initial codebook which is obtained by the splitting method. The initial codevector is set as the average of the entire training sequence. This codevector is then split into two parts. The iterative algorithm is run with these two vectors as the initial codebook. Again resulting final two codevectors are split into four. This process is repeated until the desired number of codevectors is obtained. The algorithm is summarized in the flowchart of Figure 4. [27] The LBG algorithm steps are as follows [11]

1. Design a 1-vector codebook; this is the centroid of the entire set of training vectors.

2. Double the size of the codebook by splitting each current codebook y_n according to the rule

$$y_n += y_n(1+\epsilon)$$

$$y_n += y_n(1-\epsilon)$$

Where n varies from 1 to the current size of the codebook, and ε is a splitting parameter.

3. Find the centroid for the split codebook. (i.e., the codebook of twice the size)

4. Iterate steps 2 and 3 until a codebook of size M is designed.

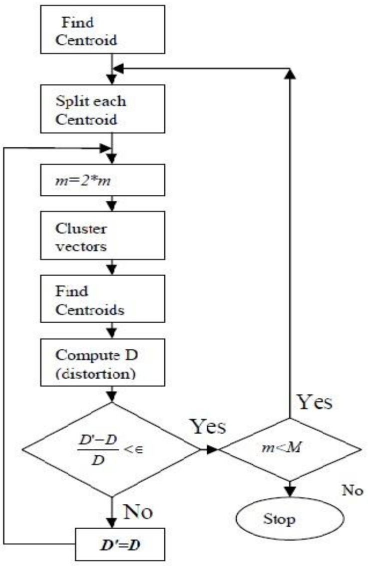

Fig 4: Flowchart of VQ-LBG algorithm

VI. ANALYSIS FROM THE REVIEW

In this paper we describe the feature extraction technique i.e. MFCC for speech recognition and its analysis. First the speech data will be collected from the native speakers of Marathi Language. The selected speakers where from different geographical region of Marathwada region of Maharashtra state. The speakers were comfortable with reading and speaking the Marathi Language. The speakers are classified on the basis of gender and were in between the age group of 18 to 30 years. The selected speakers were 50 males and 50 females. Speech recognition system contains feature extraction and feature matching. Feature extraction technique removes the noise and unnecessary information from the speech signals and recognized the pattern. Feature matching technique it approximate the general extension of the classes within the feature space from training set. Mel-Frequency Cepstral Coefficients (MFCC) is one amongst the most normally used feature extraction methodology in speech recognition. The use of Mel-Frequency Cepstral Coefficients can be considered as one of the standard method for feature extraction.

VII. CONCLUSION AND FUTURE WORK

The main aim of our project is to recognize isolated speech using MFCC and DTW techniques. In this paper we describe the speech recognition of Marathi digits using MFCC and DTW technique. In MFCC, the main advantage is that it uses Mel-frequency scaling which is very approximate to the human auditory system. Feature matching was done with the help of Dynamic Time Warping (DTW). DTW is the best nonlinear feature matching technique in speech identification, with minimal error rates and fast computing speed.With the help of VQ, the unique representation of each speaker isdone in an efficient manner.Many problemsarise during the data collection for approaching the peoples to data recording. Further work is done for increasing the accuracy and also performance of speech recognition for Marathi digits.

VIII. ACKNOWLEDGMENT

The authors would like to thank the University Authorities for providing the infrastructure to carry out the research. This work is supported by University Grants Commission.

REFERENCES

[1] Priyanka Mishra, Suyash Agrawal, “Recognition of Speaker Using Mel Frequency Cepstral Coefficient & Vector Quantization”, International Journal of Science, Engineering and Technology Research (IJSETR),Volume 1, Issue 6, December 2012.

[2] Omprakash Prabhakar, Navneet Kumar Sahu, “A Survey On: Voice Command Recognition Technique”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013.

[3] K. Samudravijaya,“Multilingual Telephony Speech Corpora of Indian Languages”, In Proceeding Computer Processing of Asian Spoken Languages. Eds, S. ItahashiAns C. Tseng, Consideration Books Los Angeles, pp.189-193, 2010.

[4] Agrawal S. S., K.K. Arora,S. Arora,Samudravijaya K, “Text and Speech Corpus Development in Indian Languages”, ibid, pp. 94-97.

[5] Santosh K. Gaikwad, Bharti W. Gawali, PravinYannawar, “A Review on Speech Recognition Technique”, International Journal of Computer Applications (0975 –887).Volume 10–No.3, Nov 2010.

[6] GopalakrishnaAnumanchipalli, Rahul Chitturi, Sachin Joshi, Rohit Kumar, Satinder Pal Singh, R.N.V. Sitaram, S. P Kishore, “Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems”.

[7] PRAAT Software.

[8] Source: http://www.fon.hum.uva.nl/praat, cited on 03/08/2013.

[9] IBM (2010) online IBM Research.

[10] Source:http://www.research.ibm.com, cited on 07/12/2013.

[11] F. Itakura, “Minimum Prediction Residula Applied to Speech Recognition”, IEEE Transaction on Acoustics, Speech, Signal Proc., and ASSP-23 (1):67-72, February 1975.

[12] L. R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, Proc. IEEE, 77 (2):257-286, February 1989.

[13] L. R. Rabiner, “Fundamentals of Speech Recognition”, Prentice Hall, Englewood (1993).

[14] Cory S. Myers, D Lawrence R. Rabiner,”A Level Building Dynamic Time Warping Algorithm For Connected Word Recognition”, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. Assp-29, No. 2, April 1981.

[15] Anjali Jain, O.P. Sharma, “A Vector Quantization Approach for Voice Recognition Using Mel Frequency Cepstral Coefficient (MFCC): A Review”, IJECT Vol. 4, Issue Spl - 4, April - June 2013, ISSN: 2230-7109 (Online) | ISSN: 2230-9543 (Print).

[16] 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 , ISSN: 2277 128X , Volume 2, Issue 2, February 2012

[17] Stolcke A., Shriberg E., Ferrer L., Kajarekar S., Sonmez K., Tur G.(2007), “Speech Recognition As Feature Extraction For Speaker Recognition”, SAFE, Washington D.C., USA pp 11-13.

[18] Zheng Fang, Zhang Guoliang, Song Zhanjiang, “Comparison of Different Implementations of MFCC”, Journal of Computer Sci. Technology., 16(6):582–589, 2001

[19] H. Hermansky, “Perceptual Linear Predictive (PLP) Analysis of Speech”,J. Acoustic. Soc. America, 87, 1990

[20] MarutiLimkar, Rama Rao, Vidya Sagvekar, “Isolated Digit Recognition Using MFCC AND DTW”, International Journal on Advanced Electrical and Electronics Engineering, (IJAEEE), ISSN (Print): 2278-8948, Volume-1, Issue-1, 2012

[21] Siddheshwar S. Gangonda, Dr. Prachi Mukherji, “Speech Processing for Marathi Numeral Recognition using MFCC and DTW Features”, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622

[22] Shivanker Dev Dhingra ,Geeta Nijhawan , Poonam Pandit, “Isolated Speech Recognition Using MFCC and DTW”, International Journal of Advanced Research in Electrical,Electronics and Instrumentation Engineering, Vol. 2, Issue 8, August 2013

[23] “MFCC and Its Applications in Speaker Recognition” VibhaTiwari, Deptt. of Electronics Engg., Gyan Ganga Institute of Technology and Management, Bhopal, (MP) INDIA (Received 5 Nov., 2009, Accepted 10 Feb., 2010).

[24] Anjali Bala, Abhijeet Kumar, Nidhika Birla, “Voice Command Recognition System Based on MFCC and DTW”, International Journal of Engineering Science and Technology Vol. 2 (12), 2010, 7335-7342.

[25] Anjali Jain, O.P. Sharma, “A Vector Quantization Approach for Voice Recognition Using Mel Frequency Cepstral Coefficient (MFCC): A Review”, IJECT Vol. 4, Issue Spl - 4, April - June 2013, ISSN: 2230-7109 (Online) | ISSN: 2230-9543 (Print).

[26] Singh Satyanand, Dr. E.G Rajan,"Vector Quantization Approach for Speaker Recognition Using MFCC and InvertedMFCC”, International Journal of Computer Applications,Vol. 17, No. 1, pp. 1-7, March 2011.

[27] Md. RashidulHasan, Mustafa Jamil, Md. GolamRabbani, Md. SaifurRahman, “Speaker Identification Using Mel Frequency Cepstral Coefficients”, 3rd International Conference on Electrical & Computer Engineering ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh.

[28] R. M. Gray, ``Vector Quantization," IEEE ASSP Magazine, pp. 4--29, April 1984.

[29] Y. Linde, A. Buzo & R. Gray, “An algorithm for vector quantizer design”, IEEE Transactions on Communications, Vol. 28, pp.84-95, 1980.

AUTHOR BIOGRAPHY

Yogesh K. Gedam received his Bachelor's Degree in Information Technology from Walchand College of Engineering, Sangli. Currently he is pursuing his Master's in Computer Science & Engineering from Department of CS & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. His research interest includes Speech Recognition, Data Mining and Machine Learning.

Sujata S. Magare received herBachelor's Degree in Information Technology from Hi-Tech Institute of Technology, Aurangabad.Currently she is pursuing her Master's in Computer Science & Engineering from Department of CS & IT, Dr. Babasaheb Ambedkar Marathwada University, and Aurangabad. Her research interest includes Character Recognition, Speech Recognition and Data Mining.

Amrapali C. Dabhade received he rBachelor's Degree in Information Technology from P. E. S. college of Engineering, Aurangabad.Currently,she is pursuing her Master's in Computer Science & Engineering from Department of CS & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. Her research interest includes Speech Recognition, Machine Learning, Remote Sensing and GIS.

Dr. R. R. Deshmukh is a Professor & Member of Management Council, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. He has received M. Sc. (CSE), M. E. (CSE), Ph.D., FIETE degrees. He has 35 National and International Journal Publication. His Research interest includes Human Computer Interfacing, Data Mining, Data Warehouse and Computational Auditory Scene Analysis.