ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 1, July 2013

Automatic Speech and Speaker Recognition by MFCC, HMM and Vector Quantization

Deshmukh S.D., Bachute M.R.

Abstract*—The speech and speaker recognition by machine are crucial ingredients for many important applications such as natural and flexible human machine interfaces which are most useful for handicap person to live the better life. The speaker recognition process relies heavily on frequency analysis. This can be done because each person has some very unique characteristics to their voice that can be isolated in the frequency domain. This paper presents an approach to the recognition of speech signal using frequency spectral information with Mel frequency for the improvement of speech feature representation in a HMM based recognition approach. There are two strong reasons why Hidden Morkov Model is used. First the models are very rich in mathematical structure and hence can form the theoretical basis for use in a wide range of applications. Second the models, when applied properly, work very well in practice for several important applications.

***Index Terms*— Hidden Markov Model (HMM), Mel Frequency Cepstral Coefficients (MFCC), Vector Quantization.**

I. INTRODUCTION

This project is designed to ease this communication barrier by helping the computer understand human speech through an speech recognition system. Speech recognition is used to provide access for anybody who has a handicap that prevents use of a keyboard. There is an entire class of people that cannot use a computer at all because they are disabled. Speech recognition could potentially make their lives easier. Computers also need a way to be able to identify who is trying to use them. The most common method of user identification is through the use of passwords. Passwords are not always effective for several reasons. The first reason is that the computer identifies the user purely based on a sequence of characters input by the user. I t is easy to see that anyone knowing this sequence can gain access, even if they are not the intended user. Passwords can also be guessed or broken. There are several characteristics to a person's voice that are unique to the individual. Because of this uniqueness, a person's voice could be a very accurate way to authenticate a user. Speech and speaker Signal Identification consist of the process to convert a speech waveform into features [1] that are useful for further processing. There are many algorithms and techniques are use. It depends on features capability to capture time frequency and energy into set of coefficients for cepstrum analysis. Generally, human voice conveys much information such as gender, emotion and identity of the speaker. Several techniques have been proposed for reducing the mismatch between the testing and training environments. Many of these methods operate either in spectral or in cepstral domain. Firstly, human voice is converted into digital signal form to produce digital data representing each level of signal at every discrete time step. The digitized speech samples are then processed using MFCC to produce voice features. After that, the coefficient of voice features can go through DTW to select the pattern that matches the database and input frame in order to minimize the resulting error between them. The popularly used cepstrum based methods to implemented using MATLAB. This paper reports the findings of the speech as well as speaker recognition study using the MFCC and HMM techniques.

II. LITERATURE SURVEY

ASR have been researched and developed as early as the 1950's at Bell Laboratories .Their ASR system could recognize the numbers 0 - 9 when spoken over a telephone. In the late 1960's, Atal and Itakura independent formulated the fundamental concepts of Linear Predictive Coding (LPC) [10], for estimation of the vocal tract response from speech waveforms. Another technology that was introduced in the late 1980's was the idea of artificial neural networks (ANN). In the 1980s, the cepstrum began to supplant the direct use of the LP parameters as the premiere feature in the important *Hidden Markov Modelling* strategy because of two convenient enhancements that were found to improve recognition rates. [4] The first is the ability to easily smooth the LP-based spectrum using the Liftering and weighting processes described above. This process removes the inherent variability of the LP-based spectrum due to the excitation and apparently improves recognition performance

III. DETAILED METHADOLOGY OF SPEECH RECOGNITION- TRAINING PHASE

Training phase is also known as database creation

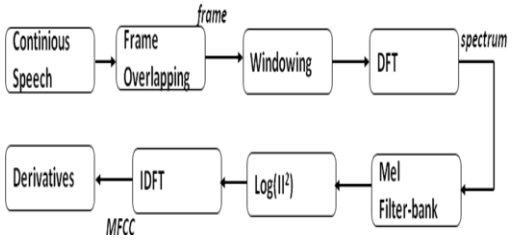

Fig 1: Block diagram of Training phase

Fig1 shows training phase in which MFCC coefficients are extracted through the following steps [4].

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 1, July 2013

A. Read speech signal and Divide into overlapping frames

First block is taking speech input signal then divide signal into overlapping frames. Speech is not a stationary signal, .Frame size is typically 10-25m and frame shift is the length of time between successive frames which is typically, 5-10ms [1].

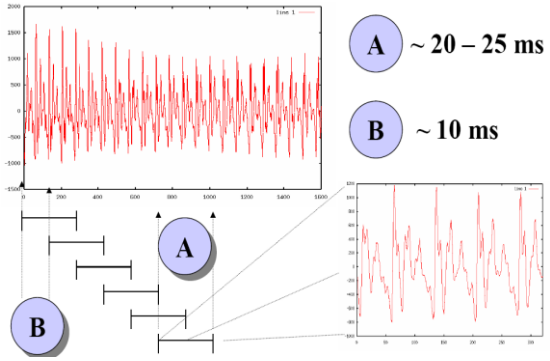

Fig 2: A Is Frame Size & B Is The Frame Shift.

B.Windowing

The third block is windowing .There are different types of windows. In our project we are using hamming window. the windowing is in order to reduce signal discontinuity at either end of the block[3]. A commonly used window is the Hamming window. The equation of Hamming window is

$$\omega[n]=\left\{0.54-0.46\cos\left(\frac{2\pi n}{L-1}\right)\right\}0\le n\le L-1 \quad (1)$$

C. DFT of all frames

The input to DFT is Windowed signal x[n]…x[m] ,Output is For each of N discrete frequency bands. A complex number X[k] representing magnitude and phase of that frequency component in the original signal.

The equation of Discrete Fourier Transform is [4]

$$X[k]=\sum_{n=0}^{N-1}x[n]e^{-j2\frac{\pi}{N}kn} \quad (2)$$

Standard algorithm for computing DFT is Fast Fourier Transform (FFT) with complexity N*log(N) in general, value of N is chosen as N=512 or 1024.

A 24 ms Hamming-windowed signal and its spectrum as computed by DFT .

D. Design Triangular filters on Mel scale

A **Mel** is a unit of pitch. Pairs of sounds perceptually equidistant in pitch are separated by an equal number of mels. Mel-scale is approximately linear below 1 kHz and logarithmic above 1 kHz. Mel scale is calculated by following formula [1,4]

$$Mel(f)=2595\log_{10}\left(1+\frac{f}{700}\right) \quad (3)$$

Human hearing is not equally sensitive to all frequency bands it is less sensitive at higher frequencies, roughly > 1000 Hz. Human perception of frequency is non-linear. As reference for the mel scale, [3] 1000 Hz is usually said to be 1000 mels.

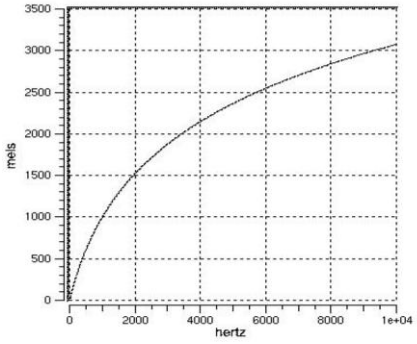

Fig 3: Nonlinear Mel Scale

E. Pass all DFT spectrums through triangular filter.

Mel spaced filter bank is one approach to simulating the subjective spectrum is to use a filter bank, spaced uniformly on the mel scale where the filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval. The modified spectrum of thus consists of the output power of these filters when S is the input. The number of Mel spectrum coefficients, K, is typically chosen as 20[4].

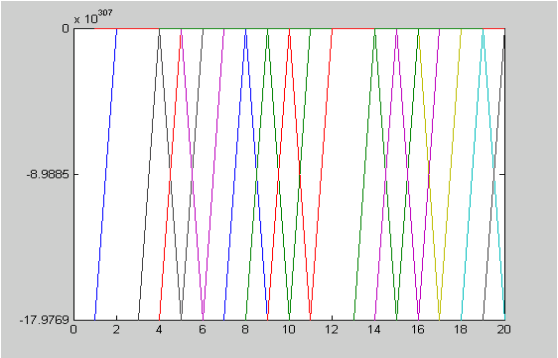

Fig 4 : Mel spaced filter bank

By applying the bank of filters according Mel scale to the spectrum each filter output is the sum of its filtered spectral components [6].

F. Log energy computation

Then Compute the logarithm of the square magnitude of the output of Mel-filter bank Logarithm compresses dynamic range of values Human response to signal level is logarithmic humans less sensitive to slight differences in amplitude at high amplitudes than low amplitudes Makes frequency estimates less sensitive to slight variations in input (power variation due to speaker's mouth moving closer to mike) [3,6] Phase information not helpful in speech.

G. Converting the output to time domain

Convert the log Mel spectrum back to Time domain and at this stage we get MFCC Mel Frequency cepstrum coefficient.The cepstrum is the spectrum of a spectrum. A spectrum gives you information about the frequency components of a signal. A cepstrum gives you information about how those frequencies change.The cepstrum requires

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 1, July 2013

Fourier analysis But we're going from frequency space back to time. So we actually apply inverse DFT [1,4].

$$y_t[k]=\sum_{m=1}^{M}\log\left(\left|y_t(m)\right|\right)\cos\left(k\left(m-.05\right)\frac{\pi}{M}\right),k=0,.....j \quad (4)$$

Details for signal processing gurus: Since the log power spectrum is real and symmetric, inverse DFT reduces to a Discrete Cosine Transform (DCT) then Repeat above process for all words in data base.

IV. DETAILED METHADOLOGY OF SPEECH RECOGNITION- TESTING PHASE

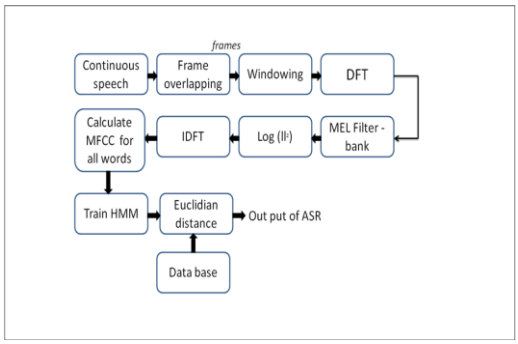

Fig5: Block Diagram of Testing Phase

In testing phase steps are repeat till MFCC coefficient. Comparing the database prepared in training phase by taking the Euclidean distance and HMM algorithm [8].

HIDDEN MARKOV MODEL

To understand the concepts of HMM the following elements are defined as[5,7]:

1. N is the no of states in HMM module i.e.{1,2,3..N } and the state at time t is q_t.
2. M is the different symbols per state
3. .Initial State Distribution $\pi=\{\pi_i\}_{i=1}^{N}$ in which π_i IS defined as π_i =P($q_1=i$) (5)
4. State transition Probability distribution A =[a_{ij}] where

$$a_{ij}=P(q_{t+1}=j\,|\,q_t=i),1\le i,\ j\le N \quad (6)$$

5. Observation symbol probability distribution

$$B=b_j(o_t)_{j=1}^{N}$$

$$b_j(o_t)\sum_{j=1}^{N}\hat{\alpha}t-1(j)a_{ij} \quad (7)$$

in which the probabilistic function for each state j is $b_j(o_t)=P(o_t\,|\,q_t=j)$ the calculation of $b_j(o_t)$ is discrete or continuous observation densities.

The three sets of probability measures are π, A and B and these probability measures use the notation λ. λ=(A,B,π). This is known as HMM model in which the states are hidden so called Hidden Markov Model.

Three basic problems with HMM [6]

Problem 1:
Given the observation sequence O = (o_1 , o_2 , .. ,o_T) , and the model λ = (A, B, π), how the probability of the observation sequence given in the model, computed? That is, how P (O|λ) computed efficiently?

Problem 2:
Given the observation sequence O = (o_1 , o_2 , .. ,o_T) , and the model λ = (A, B, π), how corresponding state sequence, q = (q_1 , q_2 , . . . , q_T),chosen to be optimal ?

Problem 3:
How the probability measures, λ = (A, B, π), adjusted to maximize P (O|λ)? The third problem is seen as the training problem. That is given the training sequences create a model for each word.

A. Solution to Problem 1 - Probability Evaluation

The aim of this problem is to find the probability of the observation sequence = (o_1 , o_2 , . . . , o_T ,), given the model λ, i.e., P (O|λ). Because the observations produced by states are assumed to be independent of each other and at time t, the probability of observation sequence, O = (o_1 , o_2 , . . . , o_T ,) being generated by a certain state sequence q is calculated by a product [7] :

$$P(O|q,B)=bq_1(o_1)\cdot bq_2(o_2)\cdot\ldots\cdot bq_T(o_T) \quad (8)$$

And the probability of the state sequence, q is found as

$$P(q|A,\pi)=\pi q_1\cdot aq_1q_2\cdot aq_2q_3\cdot\ldots\cdot aq_{T-1}q_T \quad (9)$$

The joint probability of O and q, i.e., the probability that O and q occur simultaneously, is simply the product of the above two terms,

i.e.: P (O, q|λ) = P (O|q, B) · P (q|A, π)

$$=\pi q_1 bq_1(o_1)aq_1q_2 bq_2(o_2)\cdot\ldots\cdot aq_{T-1}q_T bq_T(o_T) \quad (10)$$

The aim is to find P (O|λ), and this porobability of O (given the model λ) is obtained by summing the joint probability over all possible state sequences q [5].

Initially at time t = 1 the process starts by jumping to state q_1 with probability πq_1 , and generate the observation symbol o_1 with probability bq_1 (o_1). The clock changes from t to t + 1 and a transition from q1 to q2 will occur with probability aq1 q2 , and the symbol o2 will be generated with probability bq2 (o2). The process continues in this manner until the last transition is made (at time T), i.e., a transition from q_{T-1} to q_T will occur with probability $aq_{T-1}q_T$, and the symbol o_T will be generated with probability bq_T (o_T).

The forward algorithm is An excellent tool which cuts the computational requirements to linear, relative to time.

1 The Forward Algorithm

Consider a forward variable αt (i), defined as:

$$\alpha t(i)=P(o_1\,o_2\ldots ot,\ qt=i|\lambda) \quad (11)$$

Where t represents time and i is the state, this gives that α_t (i) will be the probability of the partial observation sequence, $o_1\,o_2\ldots o_t$, (until time t) when being in state i at

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 1, July 2013

time t. The forward variable can be calculated inductively, $\alpha_{t+1}(i)$ is found by summing the forward variable for all N states at time t multiplied with their corresponding state transition probability, aij , and by the emission probability bj (ot + 1). This is done with the following procedure [5 , 7],:

. Initialization

Set t = 1;

$$\alpha 1\ (i) = \pi i\ bi\ (o_1),\ 1 \le i \le N \qquad (12)$$

2. Induction

$$\alpha t+1\ (j) = bj\ (ot+1)\ \alpha t\ (i) aij\ ,\ 1 \le j \le N \qquad (13)$$

3.Update time

Set t = t + 1;

Return to step 2 if t < T ;

Otherwise, terminate the algorithm (go to step 4).

4. Termination

$$\sum_{i=1}^{N} \alpha T\ (i) \qquad (14)$$

2 The Backward Algorithm

The recursion described in the forward algorithm, is done in the reverse time. By defining the backward variable

βt (i) as: βt (i) = P ($_{ot+1}$ o_{t+2} . . . o_T |q_t = i, λ)

That is, the probability of the partial observation sequence from t + 1 to the end, given state i at time t and the model λ. Notice that the defination for the forward variable is a joint probability whereas the backward probability is a conditional probability [5,8]. In a similar manner (according to the forward algorithm), can the backward be calculated inductively,

The backward algorithm includes the following steps:

1.Initialization:

Set t = T − 1;

$$\beta T\ (i) = 1,\ 1 \le i \le N \qquad (15)$$

2. Induction

$$\beta t+1\ (i) aij\ bj\ (ot+1\ ),\ 1 \le i \le N$$

(16)

3.Update time

Set t = t − 1;

Return to step 2 if t > 0;

Otherwise, terminate the algorithm.

3 Scaling the Forward and Backward Variables

The calculation of αt (i) and βt (i) involves multiplication with probabilities [8,9]. All these probabilities have a value less than 1 (generally significantly less than 1), and as t starts to grow large, each term of αt (i) or βt (i) starts to head exponentially to zero. For sufficiently large t (e.g., 100 or more) the dynamic range of αt (i) and βt (i) computation will exceed

the precision range of any machine (even in double precision) The basic scaling procedure multiplies αt (i) by a scaling coefficients dependent only of the time t and independent of the state i. The scaling factor for the forward variable is denoted ct (scaling is done every time t for all states i - $1 \le i \le N$) . This factor will also be used for scaling the backward variable, βt (i). Scaling αt (i) and βt (i) with the same scale factor will show useful in problem 3 (parameter estimation) . Consider the computation of the forward variable, αt (i). In the scaled variant of the forward algorithm some extra notations will be used. αt (i) denote the unscaled forward variable, αt (i) denote the scaled and iterated variant of αt (i),ˆαt (i) denote the local version of αt (i) before scaling and ct will represent the scaling coefficient at each time [9]. Here follows the scaled forward algorithm:

Initialization: set t=2;

$$\alpha_1\ (i) = \pi i\ bi\ (o_1),\quad 1 \le i \le N \qquad (17)$$

$$\alpha 1\ \hat{}(i) = \alpha 1\ (i),\quad 1 \le i \le N$$

$$c1 = \frac{1}{\sum_{i=1}^{N} \alpha_1(i)}$$

(18)

$$\alpha_1\ (i) = c_1\ \alpha_1\ (i) \qquad (19)$$

2. Induction

$$\alpha_1\ (i) = b_j(o_t) \sum_{j=1}^{N} \alpha_{t-1}(j) a_{ij} \quad 1 \le i \le N \qquad (20)$$

$$c_t = \frac{1}{\sum_{i=1}^{N} \hat{\alpha}_t(i)} \qquad (21)$$

$$\alpha_t(i) = C_t \hat{\alpha}_t(i) \quad 1 \le i \le N2 \qquad (22)$$

3. Update time: Set t = t+1;

Return to step 2 if t<=T;

Otherwise terminate the algorithm (go to step 4)

L = 4;

4. Termination

$$\log p(o|\lambda) = \sum_{t=1}^{T} \log c_t \qquad (23)$$

B. Solution to Problem 2 - "Optimal" State Sequence

When some state transitions have zero probability (aij = 0)? $\gamma t(i) = p(q_t = i|o, \lambda)$ This means that the found optimal path may not be valid. Such a method exist, based on dynamic programming, namely the **Viterb**i algorithm is used [9].

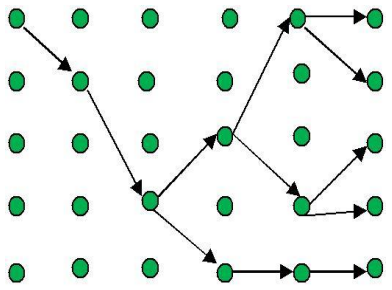

Fig 6: Trellis Diagram

The speaker recognition is carried out to obtain the maximum a posterior (MAP) estimate of the underlying state sequence. In a trellis diagram, such as Figure6, the number of paths diverging from each state of a trellis can grow exponentially by a factor of N at successive time instants. The Viterbi method prunes the trellis by selecting the most likely path to each state. At each time instant t, for each state i, the

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 1, July 2013

algorithm selects the most probable path to state i and prunes out the less likely branches. This procedure ensures that at any time instant, only a single path survives into each state of the trellis. For each time instant t and for each state i, the algorithm keeps a record of the state j from which the maximum-likelihood path branched into i, and also records the cumulative probability of the most likely path into state i at time t.

C. Solution to Problem 3 - Parameter Estimation

The third problem is concerned with the estimation of the model parameters,

$\lambda = (A, B, \pi)$. The problem can be formulated as:

$\lambda* = \arg\max [P(O|\lambda)]$

Given an observation O, find the model $\lambda*$ from all possible λ that maximizes $P(O|\lambda)$. This problem is the most difficult of the three problems. This because there is no known way to analytically find the model parameters that maximizes the probability of the observation sequence in a closed form. However can the model parameters be chosen to locally maximize the likelihood $P(O|\lambda)$. Some common used methods for solving this problem is Baum-Welch method (also known as expectation-maxima ion method) or gradient technics. Both of these methods uses iterations to improve the likelihood $P(O|\lambda)$, however there are some advantages with the Baum-Welch method compared to the gradient techniques [4,9].

V. EXPERIMENTAL RESULTS

As shown in table 1 , the 97.4 % successful detection probability is achieved. it is desirable to have the values as close to the desired values and get a practical using this implementation the words "lights on" (W1) and "open Door"(W2) spoken by different 5 speaker was used for test and result is 97.4 % including speech and speaker recognition. This project achieved same recognition probability for any word instead of "light on" and "open door". Also it gives good results for the speaker's difference voice at various emotions. The acoustical environment where recognizers are used to introduce another layer of corruption in speech signals. This is because of background noise, reverberation, microphones, and transmission channels. The system is robust for such a kind of noise.

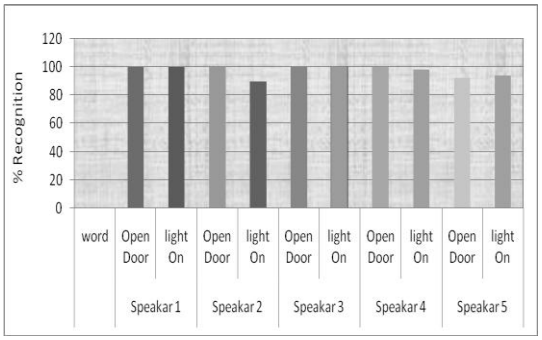

Fig 7: Graph of % Recognition

VI. CONCLUSION

In this project we successfully identify the person speaking and what they were saying. We successfully differentiated between two different speakers saying two different words. The word recognition approach yielded really good results. In the tests performed, the words "open door" and "light on" were used by five different speakers. The percentage of success and failure rate does not vary much if a different speaker speaks the same word. In addition, the speech signals between words and speakers were recorded on the same environment, using the same microphone. For noisy background the recognition rate reduces from 97 % to 92%. In this project the speech and speaker both are Recognized by taking the MFCC co-efficient for vocal cord and vocal track information. The hidden Markov model has been studied thoroughly together with the signal processing of speech signals. A speech recognizer has been implemented In recent years there has been a steady movement towards the development of speech technologies to replace or enhance text input called as Mobile Search Applications. Recently both Yahoo! and Microsoft have launched voice- based mobile search applications. Future work can include improving the recognition accuracy by combining the multiple classifiers.

VII. ACKNOWLEDGMENT

First, I would like to thank Head of Dept. Prof. A.D.Bhoi their guidance and interest. I also thank my Project Guide Prof M.R.Bachute . Their guidance reflects expertise we certainly do not master ourselves. I also thank them for all their patience throughout, in the cross-reviewing process which constitutes a rather difficult balancing act. Secondly, I would like to thank all the Staff Members of E&TC Dept. for providing me their admirable feedback and invaluable insights whenever I discussed my project with them.

REFERENCES

[1] Speech Recognition using Mel cepstrum, delta cepstrum and delta-delta Features submitted to fulfill the requirements for ECE 8993: Fundamentals of Speech Recognition submitted to Dr. Joseph Picone.

[2] Dynamic Time Warping Eamonn J. Keogh† and Michael J. Pazzani

[3] Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter Sandipan Chakroborty* and Goutam Saha.

[4] SPEECH RECOGNITION using HMM with MFCC-An analysis using frequency spectral Decomposition technique Ibrahim Patel1 Dr. Y. Srinivas Rao.

[5] Assoc. Prof., Department of BME,Padmasri.Dr.B.V.Raju Institute of Technology.Narsapur s Ptlibrahim@gmail.com Assoc. Prof., Department of Instrument Technology, Andhra University, Vizag, A.P. srinniwasarau@gmail.com.

[6] Remote Speaker and Speech Recognition A senior design project Department of Electrical Engineering University of California, Riverside Prepared by: Isaac Saldana David Ginsberg Faculty advisor: Yingbo Hua .

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 1, July 2013

[7] Text Dependent Speaker Identification System using Discrete HMM in Noise.

[8] APPLICATIONS OF SPEECH RECOGNITION in areas of telecommunication Lawrence R. Rabiner AT&T Labs.

[9] Speech Recognition Using Hidden Markov Model, Performance evalution in noisy enviorment.MEE-01-27 by Mikael Nilson, Marcus Ejnarsson, Bleking Institute of Technology March 2002.

[10] Itakura F Minimum Prediction Residual Applied to Speech.

[11] Recognti tion IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-23 (1):67-72, 1975.

[12] Rabiner L. R., Levinson S. E., Rosenberg A. E. and Wilpon J. G.Speaker dependent Recognition of isolated word Recognition using clustering techniques, IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-27:336-349, 1979.

[13] Wang Yutai, Li Bo, Jiang Xiaoqing, Liu Feng, Wang Lihao Speaker Recognition Based on Dynamic MFCC Parameters. IEEE conference 2009.

[14] Nitin Trivedi, Dr. Vikesh Kumar, Saurabh Singh, Sachin Ahuja & Raman Chadha. Speech Recognition by Wavelet Analysis International Journal of Computer Applications (0975 – 8887) Volume 15– No.8, February 2011.

AUTHOR BIOGRAPHY

Deshmukh Sharmila Dnyaneshwar
Department of Electronics and Telecommunication,
G.H.Raisoni Institute of Engineering and Technology
E-mail sharmilad80@gmail.com

PROF.BACHUTE M.R.
Department of Electronics and Telecommunication,
G.H.Raisoni Institute of Engineering and Technology
E-mail: mrinal.bachute@raisoni.net