ADVANCED SPEAKER RECOGNITION

Amruta Anantrao Malode and Shashikant Sahare
[1]Department of Electronics & Telecommunication, Pune University, Pune, India

ABSTRACT

The domain area of this topic is Bio-metric. Speaker Recognition is biometric system. This paper deals with speaker recognition by HMM (Hidden Markov Model) method. The recorded speech signal contains background noise. This noise badly affects the accuracy of speaker recognition. Discrete Wavelet Transforms (DWT) greatly reduces the noise present in input speech signal. DWT often outperforms as compared to Fourier Transform, due to its capability to represent the signal precisely, in both frequency & time domain. Wavelet thresholding is applied to separate the speech and noise, enhancing the speech consequently.The system is able to recognize the speaker by translating the speech waveform into a set of feature vectors using Mel Frequency Cepstral Coefficients (MFCC) technique. But, input speech signals at different time may contain variations. Same speaker may utter the same word at different speed which gives us variation in total number of MFCC coefficients. Vector Quantization (VQ) is used to make same number of MFCC coefficients. Hidden Markov Model (HMM) provides a highly reliable way for recognizing a speaker. Hidden Markov Models have been widely used, which are usually considered as a set of states with Markovian properties and observations generated independently by those states. With the help of Viterbi decoding most likely state sequence is obtained. This state sequence is used for speaker recognition. For a database of size 50 in normal environment, obtained result is 98% which is better than previous methods used for speaker recognition.

KEYWORDS: *Digital Circuits, Codebook, Discrete Wavelet Transform (DWT), Hidden Markov Model (HMM), Mel Frequency Cepstral Coefficients (MFCC), Vector Quantization (VQ), Viterbi Decoding.*

I. INTRODUCTION

Speaker recognition is the process of automatically extracting the features and recognizing speaker using computers or electronic circuits [2]. All of our voices are uniquely different (including twins) and cannot be exactly duplicated. Speaker recognition uses the acoustic features of speech that are different in all of us. These acoustic patterns reflect both anatomy (size and shape of mouth & throat) and learned behavior patterns (voice pitch & speaking style).

If a speaker claims to be of a certain identity and their speech is used to verify this claim. This is called verification or authentication. Identification is the task of determining an unknown speaker's identity. Speech recognition can be divided into two methods i.e. text dependent and text independent methods. Text dependent relies on a person saying a pre determined phrase whereas text independent can be any text or phrase. A speech recognition system has two phases, Enrolment and verification. During enrolment, the speaker's voice is recorded and typically a number of features are extracted to form a voiceprint. In the verification phase, a speech sample or utterance is compared against a previously created voiceprint. For identification systems, the utterance is compared against multiple voiceprints in order to determine the best match or matches, while verification systems compare an utterance against a single voiceprint. Because of this process, verification is faster than identification.

In many speech processing applications, speech has to be processed in the presence of undesirable background noise, leading to a need to a front-end speech enhancement. In 1995, Donoho introduced wavelet thresholding as a powerful tool in denoising signals degraded by additive white noise [3]. It has the advantage of using variable size time-windows for different frequency bands. This results in a high frequency-resolution (and low time-resolution) in low bands and low frequency-resolution in high bands. Consequently, wavelet transform is a powerful tool for modelling non-stationary signals like speech that exhibit slow temporal variations in low frequency and abrupt temporal changes in high frequency.

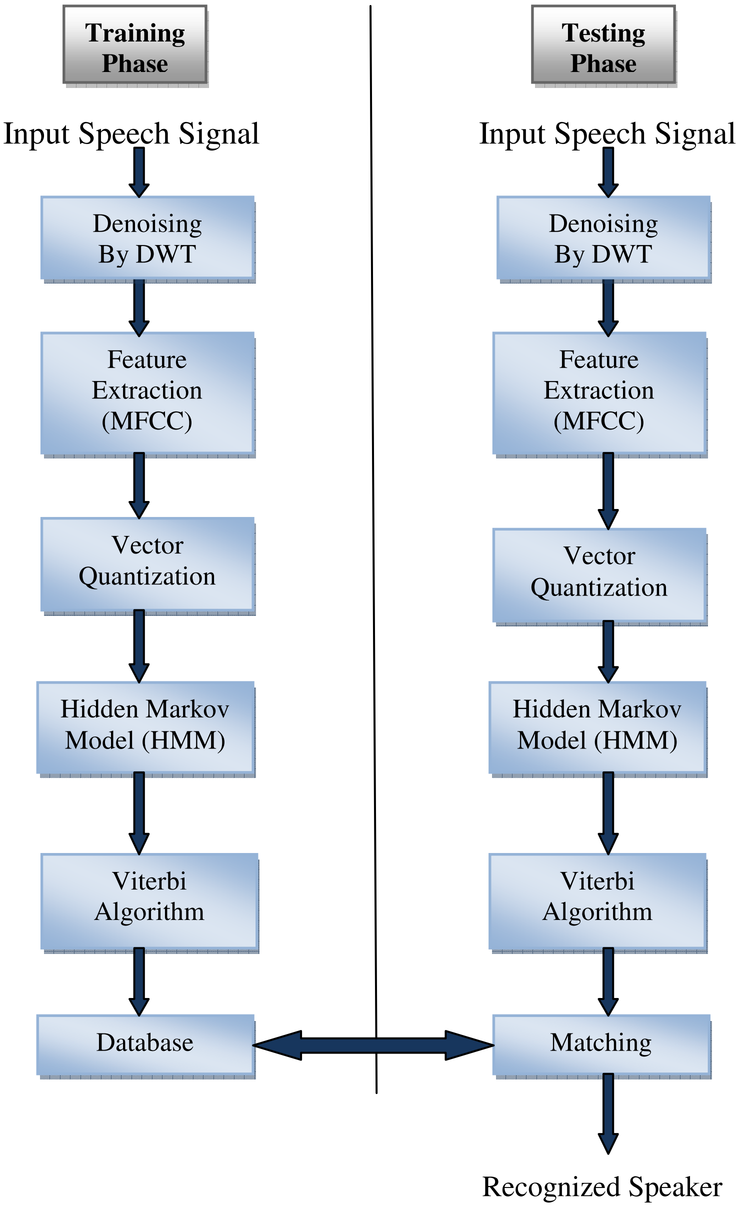


Figure 1. Speaker Recognition System

Figure 1 shows the block diagram of the Speaker Recognition System. In the research of speaker recognition, the characteristic parameter of speech, which can efficiently represent the speaker's specific feature, plays an important role in the whole recognition process. The most frequently used parameters are pitch, formant frequency and bandwidth, Linear Predictive Coefficients (LPC), Linear Predictive Cepstrum Coefficients (LPCC), Mel-Frequency Cepstrum Coefficients (MFCC) and so on. The formant, LPC and LPCC are related to vocal tract, and are good speaker identification characteristics with high SNR (signal to noise ratio). However, when the SNR is low, the differences between the vocal tract parameters estimated from noisy speech signal and those of the real vocal tract model are big. Thus, these characteristic parameters cannot correctly reflect speaker's vocal tract features. [1]

The MFCC parameter, mainly describes the speech signal's energy distribution in a frequency field. This method, which is based on the Mel frequency scale and accords with human hearing characteristics, has better anti-noise ability than other vocal tract parameters, such as LPC. Because of the preferable simulation of the human hearing system's perception ability, it is considered as an important characteristic parameter by the researcher of speech and speaker recognition [1]. The size of MFCC parameters is not fixed & hence VQ can be used to fix the size of MFCC parameters.

Hidden Markov Models which widely used in various fields of speech signal processing is a statistical model of speech signals. To the smooth and time-invariant signals, we can describe by the traditional linear model. But to the nonstationary and time-varying speech signal, we can only make linear processing in the short time. In doing so, the linear model parameter of speech signal is time-variant in a period of time, but in short time it can be regarded as stable and time-invariant. Under the precondition, the simple solving idea of dealing with speech signal is markov chain which made these linear model parameter connect and record the whole speech signal .But it has a problem that how long a period of time as a linear processing unit. It is hard to accurately choose the period of time because of the complex of the speech signal. So, this method is feasible but not the most effective means [4]. Hidden markov models can solve the foresaid problem. It can not only solve the describing stationary signal, but also solve the smooth transition in a short time. On the basis of probability and mathematical statistical theory, HMM can identify any temporary smooth process of different parameters and trace the conversion process.

This paper is organized as follows. The section II deals with Discrete Wavelet Transform. The section III deals with MFCC parameter extraction. Section IV deals with Vector Quantization. In Section V HMM model is presented. The section VI, deals with Viterbi decoding for speaker recognition. At the last VII section shows experimental results & section VIII gives conclusion & Future Scope.

II. DISCRETE WAVELET TRANSFORM

The wavelet denoising is based on the observation that in many signals (like speech), energy is mostly concentrated in a small number of wavelet dimensions. The coefficients of these dimensions are relatively large compared to other dimensions or to any other signal (specially noise) that has its energy spread over a large number of coefficients. Hence, by setting smaller coefficients to zero, one can nearly optimally eliminate noise while preserving the important information of the original signal. Let be a finite length observation sequence of the signal that is corrupted by zero-mean, white Gaussian noise with variance σ^2.

$$y(n) = x(n) + \text{Noise}(n) \qquad (1)$$

The goal is to recover the signal x from the noisy observation $y(n)$. If W denotes a discrete wavelet transform (DWT) matrix, equation (1) (which is in time domain) can be written in the wavelet domain as

$$Y(n) = X(n) + N(n) \qquad (2)$$

Where

$$Y(n) = W_y, \quad X(n) = W_x, \quad N(n) = W_n$$

Let X_{est} be an estimate of the clean signal x based on the noisy observation Y in the wavelet domain. The clean signal x can be estimated by

$$x = W^{-1}X_{est} \quad = W^{-1}Y_{thr} \qquad (3)$$

Where Y_{thr} denotes the wavelet coefficients after thresholding. The proper value of the threshold can be determined in many ways. Donoho has suggested the following formula for this purpose

$$T = \sigma\sqrt{2\log(N)} \qquad (4)$$

Where T is the threshold value and N is the length of the noisy signal (y). Thresholding can be performed as Hard or Soft thresholding that are defined as follows, respectively:

$$\mathrm{THR_H}\,(\mathrm{Y},\mathrm{T}) = \begin{cases} \mathrm{Y}, & |\mathrm{Y}| > T \\ 0, & |\mathrm{Y}| < T \end{cases} \qquad (5)$$

And

$$\mathrm{THR_S}\,(\mathrm{Y},\mathrm{T}) = \begin{cases} \mathrm{Sgn(Y)}\,(|\mathrm{Y}| - \mathrm{T}), & |\mathrm{Y}| > T \\ 0, & |\mathrm{Y}| < T \end{cases} \qquad (6)$$

Soft thresholding gives better result than hard thresolding. Hence, soft thresholding is used [3].

III. Mel Frequency Cepstral Coefficient (MFCC)

Mel-frequency Cepstrum (MFC) is the representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum.

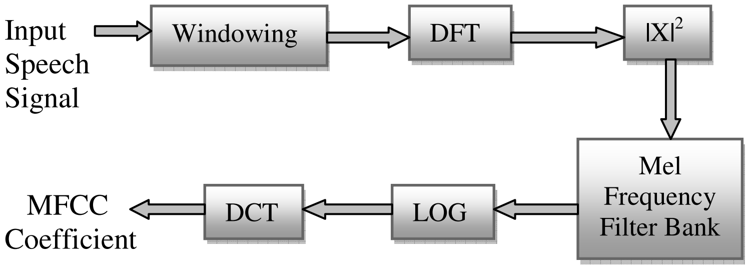


Figure 2.MFCC Parameter Extraction

Let x[n] be a speech signal with a sampling frequency of f_s, and is divided into P frames each of length N samples with an overlap of N/2 samples such that $\{\mathrm{x_1[n]}, \mathrm{x_2[n]} \ldots \mathrm{x_p[n]} \ldots \mathrm{x_P[n]}\}$, where $\mathrm{x_p}$ denotes the $\mathrm{p^{th}}$ frame of the speech signal x[n]. The size of matrix X is N x P. The MFCC frames are computed for each frame [6].

3.1 Windowing, Discrete Fourier Transform & Magnitude Spectrum

In speech signal processing, in order to compute the MFCCs of the $\mathrm{p^{th}}$ frame, $\mathrm{x_p}$ is multiplied with a hamming window

$$\mathrm{w[n]} = 0.54 - 0.46\cos\left(\frac{2\pi\mathrm{n}}{\mathrm{N}}\right) \qquad (7)$$

Followed by the Discrete Fourier transform (DFT) as shown below:

$$\mathrm{X_{p(k)}} = \sum_{\mathrm{n}=0}^{\mathrm{N}-1} \mathrm{x_p}\,[\mathrm{n}]\,\mathrm{w[n]}\,\mathrm{e}^{-\mathrm{j}\frac{2\pi\mathrm{nk}}{\mathrm{N}}} \qquad (8)$$

For k = 0, 1, ⋯ , N -1.

If f_s is the sampling rate of the speech signal x[n] then k corresponds to the frequency $l_f(k) = \frac{kf_s}{N}$. Let $X_P = [X_0(0), X_1(1), \dots, X_p(p), \dots, X_P(P)]^T$ represent the DFT of the windowed p^{th} frame of the speech signal x[n], namely x_p. Accordingly, let $X = [X_0, X_1, \dots X_P]$ represent the DFT of the matrix X. Note that the size of X is N x P and is known as STFT (Short Time Fourier Transform) matrix. The modulus of Fourier transform is extracted and the magnitude spectrum is obtained as $|X|^2$ which is a matrix of size N x P.

3.2 Mel Frequency Filter Banks

For each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'Mel' scale. Mel frequency is given by

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \qquad (9)$$

Next, the filter bank which has linearly spaced filters in the Mel scale, are imposed on the spectrum. The filter response $\psi_i(k)$ of the i^{th} filter in the bank is defined in [5].

$$\psi_i(k) = \begin{cases} 0, & \text{for } k < k_{b_{i-1}} \\ \dfrac{k - k_{b_{i-1}}}{k_{b_i} - k_{b_{i-1}}}, & \text{for } k_{b_{i-1}} \le k \le k_{b_i} \\ \dfrac{k_{b_i} - k}{k_{b_{i+1}} - k_{b_i}}, & \text{for } k_{b_i} \le k \le k_{b_{i+1}} \\ 0, & \text{for } k_{b_{i+1}} < k \end{cases} \qquad (10)$$

If Q denotes the number of filters in the filter bank, then

$$\{ k_{b_i} \} \quad \text{for } i = 0,1,2,\dots.,Q+1 \qquad (11)$$

are the boundary points of the filters and k denotes the coefficient index in the N-point DFT. The boundary points for each filter i (i=1,2,...,Q) are calculated as equally spaced points in the Mel scale using [5].

$$K_{b_i} = \left(\frac{N}{f_s}\right) f_{mel}^{-1}\left[f_{mel}(f_{low}) + \frac{i\{f_{mel}(f_{high}) - f_{mel}(f_{low})\}}{Q+1}\right] \qquad (12)$$

Where, f_s is the sampling frequency in Hz and f_{low} and f_{high} are the low and high frequency boundaries of the filter bank, respectively. f_{mel}^{-1} is the inverse of the transformation and is defined in [5].

$$f_{mel}^{-1}(f_{mel}) = 700\left[10^{\frac{f_{mel}}{2595}} - 1\right] \qquad (13)$$

The mel filter bank M(m,k) is a matrix of size Q X N. Where, m = 1,2,…Q & k = 1,2,…N.

3.3 Mel Frequency Cepstral Coefficient

The logarithm of the filter bank outputs (Mel spectrum) is given by

$$L_p(m,k) = \ln\left\{\sum_{k=0}^{N-1} M(m,k) * |X_p(k)|\right\} \qquad (14)$$

where m = 1,2,⋯ , Q and p = 1,2,⋯ , P. The filter bank output, which is the product of the Mel filter bank, M and the magnitude spectrum, |X| is a Q x P matrix. A discrete cosine transforms of $L_p(m, k)$ results in the MFCC parameters.

$$\phi_p^r\{x[n]\} = \sum_{m=1}^{Q} L_p(m, k) \cos\left\{\frac{r(2m-1)\pi}{2Q}\right\} \qquad (15)$$

where r = 1,2,· .. , F and $\phi_p^r\{x[n]\}$ represents the r^{th} MFCC of the p^{th} frame of the speech signal x[n]. The MFCC of all the P frames of the speech signal are obtained as a matrix Φ.

$$\Phi\{X\} = [\Phi_1, \Phi_2, \dots, \Phi_{p,\dots}\Phi_P] \qquad (16)$$

The p^{th} column of the matrix Φ, namely Φ_P represents the MFCC of the speech signal, x[n], corresponding to the p^{th}frame, $x_p[n]$. [6]

IV. VECTOR QUANTIZATION (VQ)

MFCC parameter matrix Φ is of size Q X P. In this Q is number of Mel filters which is fixed. But, P is the total number of overlapping frames in speech signal. Each frame contains speech samples. At different time same speaker can speak the same word slowly or fast which results in variation in number of samples in input speech signal. Hence P may be different for different speech signal. Hidden Markov Model requires fixed number of states & number of samples in observation sequence. It is required that input to HMM should be of fixed size. Hence Vector Quantization is used to convert MFCC parameters of variable size into fixed size codebook. Codebook contains coefficients of Vector Quantization.

For generating the codebooks, the LBG algorithm is used. The LBG algorithm steps are as follows [16]:

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

$$y_n = y_n(1+\varepsilon) \qquad (17)$$
$$y_n = y_n(1-\varepsilon) \qquad (18)$$

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

3. Nearest neighbour search: for each training vector, find the codeword in the current codebook that is closest & assign that vector to the corresponding cell.
4. Update the codeword in each cell using the centroid of the training vectors assigned to that cell.
5. Repeat steps 3 & 4 until the average distance falls below a present threshold.
6. Repeat steps 2, 3 & 4 until a codebook size of M is designed.

This VQ algorithm gives fixed size codebook of size Q X T. Here T is any number which satisfies following condition:

$$T = 2^i \qquad i = 1,2,3, \dots..$$

V. HIDDEN MARKOV MODEL (HMM)

A hidden Markov model (HMM) is a double-layered finite state process, with a hidden Markovian process that controls the selection of the states of an observable process. In general, a hidden Markov model has N sates, with each state trained to model a distinct segment of a signal process. A hidden Markov model can be used to model a time-varying random process as a probabilistic Markovian chain of N stationary, or quasi-stationary processes [ADSPNR book by Saeed Vaseghi chapter 5].

The Hidden Markov Model (HMM) is a variant of a finite state machine having a set of hidden states, Q, an output alphabet (observations), O, transition probabilities, A, output (emission) probabilities, B, and initial state probabilities, Π. The current state is not observable. Instead, each state produces an output with a certain probability (B). Usually the states, Q, and outputs, O, are understood, so an HMM is said to be a triple, (A, B, Π).[15]

5.1 Formal Definitions

Hidden states Q ={q_i}, i = 1. . . N.
Transition probabilities A = { a_{ij} = P(q_j at t +1 | q_i at t)}, where P(a | b) is the conditional probability of a given b, t = 1, . . . , T is time, and q_i in Q. Informally, A is the probability that the next state is q_j given that the current state is q_i.
Observations (symbols) O = { o_k}, k = 1, . . . , M .
Emission probabilities B = { b_{ik} = b_i (o_k) = P(o_k | q_i)}, where o_k in O. Informally, B is the probability that the output is o_k given that the current state is q_i.
Initial state probabilities Π = {p_i = P(q_i at t = 1)}.
The model is characterized by the complete set of parameters: Λ = {A, B, Π}.

5.2 Forward Algorithm

At first the model parameters are consider as random signals because speech is random signal. To compute the probability of a particular output sequence Forward & Backward algorithms are used.
Let $\alpha_t(i)$ be the probability of the partial observation sequence $O_t = \{(o(1), o(2), \dots, o(t)\}$ to be produced by all possible state sequences that end at the i^{th} state.

$$\alpha_t(i) = P(o(1), o(2), \dots, o(t) \,| q(t) = q_i \qquad (19)$$

Then the unconditional probability of the partial observation sequence is the sum of $\alpha_t(i)$ over all N states.

The Forward Algorithm is a recursive algorithm for calculating $\alpha_t(i)$ for the observation sequence of increasing length t. First, the probabilities for the single-symbol sequence are calculated as a product of initial i^{th} state probability and emission probability of the given symbol $o(1)$ in the i^{th}state. Then the recursive formula is applied. Assume we have calculated $\alpha_t(i)$ for some t. To calculate $\alpha_{t+1}(j)$ we multiply every $\alpha_t(i)$ by the corresponding transition probability from the i^{th} state to the j^{th}state, sum the products over all states, and then multiply the result by the emission probability of the symbol $o(t+1)$. Iterating the process, we can eventually calculate $\alpha_T(i)$, and then summing them over all states, we can obtain the required probability.

Initialization:

$$\alpha_1(i) = p_i b_i\big(o(1)\big) \qquad i = 1,2, \dots, N \qquad (20)$$

Recursion:

$$\alpha_{t+1}(i) = \left[\sum_{j=1}^{N} \alpha_t(j) a_{ji}\right] b_i\big(o(t+1)\big) \qquad (21)$$

$$Here \qquad i = 1, \dots, N \qquad t = 1, \dots, T-1$$

Termination:

$$P\big(o(1)o(2) \dots . o(T)\big) = \sum_{j=1}^{N} \alpha_T(j) \qquad (22)$$

5.3 Backward Algorithm

In a similar manner, we can introduce a symmetrical backward variable $\beta_t(i)$ as the conditional probability of the partial observation sequence from $o(t+1)$ to the end to be produced by all state sequences that start at i^{th} state.

$$\beta_t(i) = P(o(t+1), o(t+2), \ldots, o(T) \mid q(t) = q_i \qquad (23)$$

The Backward Algorithm calculates recursively backward variables going backward along the observation sequence.

Initialization:

$$\beta_t(i) = 1 \qquad i = 1, \ldots, N \qquad (24)$$

Recursion:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j\big(o(t+1)\big)\beta_{t+1}(j) \qquad (25)$$

$$Here\ \ i = 1, \ldots, N \qquad t = T-1, T-2, \ldots, 1$$

Termination:

$$P\big(o(1)o(2) \ldots . o(T)\big) = \sum_{j=1}^{N} p_j b_j\big(o(1)\big)\beta_1(j) \qquad (26)$$

Both Forward and Backward algorithms gives the same results for total probabilities P(O) = P(o(1), o(2), ... , o(T)).

5.4 Baum Welch Algorithms

To find the parameters (A, B, Π) that maximize the likelihood of the observations Baum Welch Algorithm is used. It is used to train the hidden Markov model with speech signals. The Baum-Welch algorithm is an iterative expectation-maximization (EM) algorithm that converges to a locally optimal solution from the initialization values.

Let us define $\xi_t(i,j)$, the joint probability of being in state q_i at time t and state q_j at time t+1 , given the model and the observed sequence:

$$\xi_t(i,j) = P(q(t) = q_i, q(t+1) = q_j | O, \Lambda \qquad (27)$$

$\xi_t(i,j)$ is also *given* by,

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j\big(o(t+1)\big)\beta_{t+1}(j)}{P(O|\Lambda)} \qquad (28)$$

The probability of output sequence can be expressed as

$$P(O|\Lambda) = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j\big(o(t+1)\big)\beta_{t+1}(j) \qquad (29)$$

$$P(O|\Lambda) = \sum_{\mathrm{i=1}}^{\mathrm{N}} \alpha_t(i)\beta_t(i) \qquad (30)$$

The probability of being in state $\mathrm{q_i}$ at time t:

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\Lambda)} \qquad (31)$$

Estimates:

Initial probabilities:

$$\overline{p_\iota} = \gamma_1(i) \qquad (32)$$

Transition probabilities:

$$\overline{a_{\iota\jmath}} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \qquad (33)$$

Emission probabilities:

$$\overline{b_{\jmath k}} = \frac{\sum_{\mathrm{t}}^{*} \gamma_{\mathrm{t}}(\mathrm{j})}{\sum_{\mathrm{t=1}}^{\mathrm{T}} \gamma_{\mathrm{t}}(\mathrm{j})} \qquad (34)$$

In the above equation Σ* denotes the sum over t so that $\mathrm{o(t) = o_k}$.

VI. VITERBI DECODING

From HMM parameters & observation sequence Viterbi decoding finds the most likely sequence of (hidden) states.

Let δ_{t}(i) be the maximal probability of state sequences of the length t that end in state i and produce the t first observations for the given model.

$$\delta_t(i) = \max\{P(q(1), \dots, q(t-1); o(1), \dots, o(t)|q(t) = q_i \qquad (35)$$

The Viterbi algorithm uses maximization at the recursion and termination steps. It keeps track of the arguments that maximize δ_{t}(i) for each t and i, storing them in the N by T matrix ψ. This matrix is used to retrieve the optimal state sequence at the backtracking step.[15]

Initialization:

$$\delta_1(i) = p_i b_i(o(1))$$

$$\psi_1(i) = 0 \qquad for \;\; i = 1, \dots, N \qquad (36)$$

Recursion:

$$\delta_t(j) = max_i[\delta_{t-1}(i)a_{ij}]\; b_j(o(t))$$

$$\psi_t(j) = arg\; max_i\big[\delta_{t-1}(i)a_{ij}\big] \qquad for \;\; j = 1, \dots, N \;\; (37)$$

Termination:

$$p^* = max_i[\delta_T(i)]$$

$$q_T^* = arg\ max_i[\delta_T(i)] \quad (38)$$

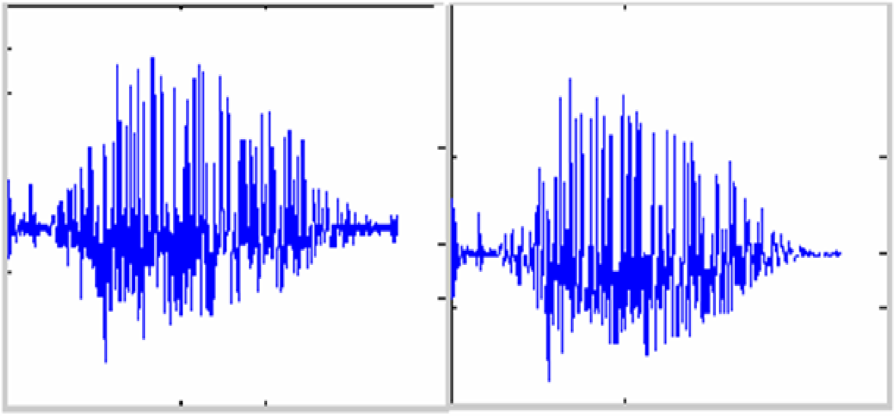
Path (state sequence) backtracking

$$q_t^* = \psi_{t+1}(q_{t+1}^*) \quad t = T-1, T-2, \ldots, 1 \quad (39)$$

VII. RESULTS & DISCUSSION

Speech signal of speaker is recorded by Audacity software at sampling frequency 44.1 KHz, stereo mode & it is saved as .WAV file. Speech is recorded in a noisy environment. Database consists of 5 speech samples of each 10 individuals. Speech signal is the “Hello” word. This is Text dependent Speaker Recognition.

Speech signal is denoised by Discrete Wavelet Transform (DWT). Noisy signal is decomposed by Daubechies family at db10. Result of denoising is given by figure 3.



(a) (b)

Figure 3. (a) Noisy Signal (b) Denoised Signal

MFCC coefficients are find out from input Speech Signal. Mel Filter Banks for 256 point DFT are shown in figure 4. Vector Quantization Coefficients of one filter bank are given figure 5. Output of MFCC is given to VQ to generate fixed size Codebook for different speech signal.

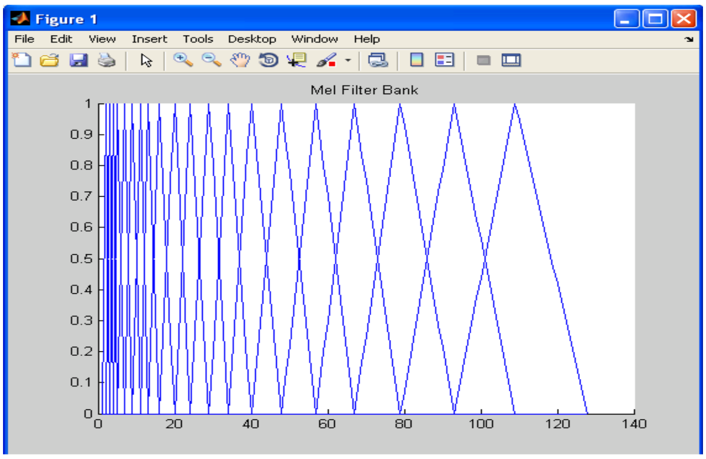


Figure 4. Mel Filter Bank

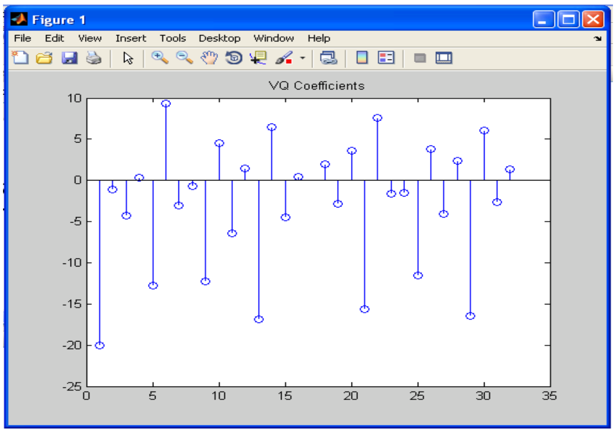
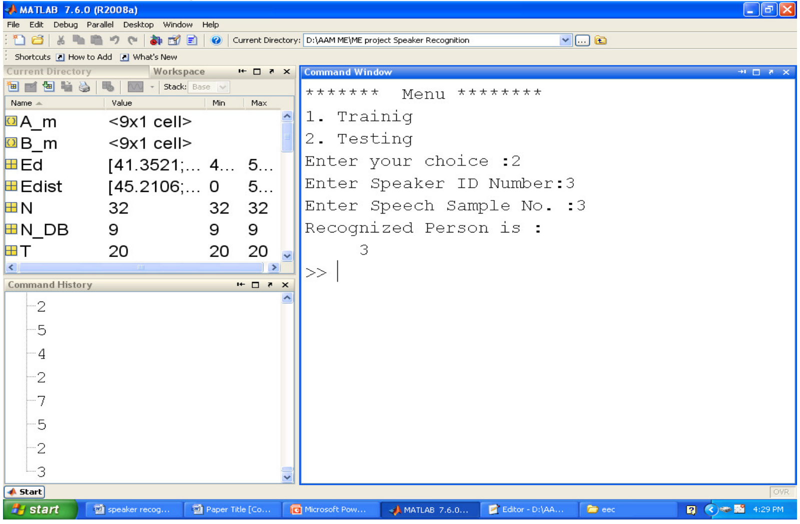


Figure 5. Vector Quantization Coefficients of one Mel Filter

At the time of Training, 3 speech samples of each individual are used. In training phase, HMM parameters & its corresponding best state sequence is find out by Viterbi Algorithm & this data is saved as database.

In testing phase, input speech is denoised at first then its MFCC coefficients are find out. At the last, with the help of HMM parameters & observation sequence new state sequence is find out by using viterbi decoding. This new state sequence is matched with database. If Match founds, person will get recognized otherwise it is not recognized.



Name	Value	Min	Max
A_m	<9x1 cell>		
B_m	<9x1 cell>		
Ed	[41.3521;...	4...	5...
Edist	[45.2106;...	0	5...
N	32	32	32
N_DB	9	9	9
T	20	20	20

```
*******  Menu ********
1. Trainig
2. Testing
Enter your choice :2
Enter Speaker ID Number:3
Enter Speech Sample No. :3
Recognized Person is :
     3
>> |
```

```
2
5
4
2
7
5
2
3
```

Figure 6. Speaker Recognition Output on MATLAB

Speaker Recognition result is shown in following table.

Table 1. Speaker Recognition Result

Sr. No.	Speaker Recognition Method	Result
1	By using HMM	92 %
2	By using DWT, VQ & HMM	98%

VIII. CONCLUSION

This paper proposes the idea that the Speaker Recognition system performance can be improved by VQ and HMM. Although it is believed that the recognition rates achieved in this research are comparable with other systems and researches of the same domain, however, more improvements need to be made specially increasing the training and testing speech data. Also input speech signal contains noise. Denoising can be used at the start to clean the speech signal. More training data & good denoising method can improve accuracy of Speaker Recognition system up to 99.99%. Such system can be implemented on DSP processor for real time Speaker Recognition.

ACKNOWLEDGMENT

First, we would like to thank Head of Dept. Prof. S. Kulkarni for their guidance and interest. Their guidance reflects expertise we certainly do not master ourselves. We also thank them for all their patience throughout, in the cross-reviewing process which constitutes a rather difficult balancing act. Secondly, we would like to thank all the Staff Members of E&TC Department for providing us their admirable feedback and invaluable insights. At the last but not least, we would like to thank our family who always supported & encouraged us.

REFERENCES

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

[2] 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.

[3] Hamid Sheikhzadeh and Hamid Reza Abutalebi, *An improved wavelet-based speech enhancement system.*

[4] ZHOU Dexiang & WANG Xianrong, *The Improvement of HMM Algorithm using wavelet de-noising in speech Recognition,* 2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE)

[5] Md. Afzal Hossan, Sheeraz Memon, Mark A Gregory, *A Novel Approach for MFCC Feature Extraction,* IEEE conference 2010.

[6] Sunil Kumar Kopparapu and M Laxminarayana, *Choice Of Mel Filter Bank In Computing Mfcc Of A Resampled Speech,* 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010).

[7] Ahmed A. M. Abushariah, Tedi S. Gunawan, Othman O. Khalifa., *English Digit speech recognition system based on Hidden Markov Models",* International Conference on Computer and Communication Engineering (ICCCE 2010), 11-13 May 2010, Kuala Lumpur, Malaysia.

[8] ZHAO Yanling ZHENG Xiaoshi GAO Huixian LI Na Shandong, *A Speaker Recognition System Based on VQ,* Computer Science Center Jinan, Shandong, 250014, China

[9] Suping Li, *Speech Denoising Based on Improved Discrete Wavelet Packet Decomposition"* 2011 International Conference on Network Computing and Information Security

[10] Wang Chen Miao Zhenjiang & Meng Xiao, *Differential MFCC and Vector Quantization used for Real-Time Speaker Recognition System,* Congress on Image and Signal Processing 2008.

[11] J.Manikandan, B.Venkataramani, K.Girish, H.Karthic and V.Siddharth, *Hardware Implementation of Real-Time Speech Recognition System using TMS320C6713 DSP,* 24th Annual Conference on VLSI Design 2011.

[12] Yariv Ephraim and Neri Merhav, *Hidden Markov Processes*, IEEE transactions on information theory, vol. 48, no.6,June,2002.

[13] L. H. Zhang, G. F. Rong, *A Kind Of Modified Speech Enhancement Algorithm Based On Wavelet Package Transformation,* Proceedings of the 2008 International Conference on Wavelet Analysis and Pattern Recognition, Hong Kong, 30-31 Aug2008.

[14] H_akon Sandsmark, *Isolated-Word Speech Recognition Using Hidden Markov Models,* December 18, 2010

[15] Lawrence R. Rabiner, Fellow IEEE, *A Tutoral on Hidden Markov Models & Selected Application in Speech Recognition,* Proceeding of the IEEE vol.77, No.2, Feb 1989.

[16] Dr. H. B. Kekre, Ms. Vaishali Kulkarni, *Speaker Identification by using Vector Quantization,* Dr. H. B. Kekre et. al. / International Journal of Engineering Science and Technology Vol. 2(5), 2010, 1325-1331.

[17] A. Srinivasan, *Speaker Identification and Verification using Vector Quantization and Mel Frequency Cepstral Coefficients,* Research Journal of Applied Sciences, Engineering and Technology 4(1): 33-40, 2012 ISSN: 2040-7467 © Maxwell Scientific Organization, 2012.

Authors

Amruta Anantrao Malode was born in Pune, India on 14 January 1986. She received Bachelor in E & TC degree from the University of Pune in 2008. She is currently pursuing ME in E & TC (Signal Processing Branch) from the University of Pune. Her research interest includes signal processing, image processing & embedded systems.

Shashikant Sahare is working in MKSSS's Cummins College of Engineering for Women, Pune in Pune University, India. He has completed his M-Tech in Electronics Design Technology. His research interest includes signal processing & Electronic design.