ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Real Time Speaker Recognition using Mel-Frequency Cepstral Coefficients (MFCC) ,VQLBG & GMM Techniques

Nisha N. Nichat[1], P. C. Latane[2]

M. E, Department of Electronics & Telecommunication, G. S. Moze College of Engineering, Balewadi Pune , India

Professor, Department of Electronics & Telecommunication, Sinhgad Institute of Technology, Lonawala, India

ABSTRACT: Speaker Identification is a technique for authenticating user's identity using characteristics extracted from their voices. This technique is one of the most useful biometric identification techniques associated to areas in which security is a most important. This is also used for authentication, surveillance, forensic speaker identification and a number of related events. Speaker recognition can be categorized into identification and verification. Speaker identification is the process to detect which registered speaker provides a given sample. Speakerverification, is the process of accepting or rejecting the identity of the speaker

The process of Speaker identificationis divided into two modules: - feature extraction and feature matching. Featureextraction is the process in which a small amount of data from the voice signalextracted that can be used to identify each speaker. Featurematching is a process which provides identification of the unknown speaker by comparing the extracted features from his/her voice input with the ones from a set of known speakers.

Proposed work comprisestrimming a recorded voice signal, framing it, passing it through a window function, calculating the Short Term FFT, extracting its features and matching it with a stored template. Mel frequencyCepstral Coefficients (MFCC) are applied for feature extraction purpose. VQLBG (Vector Quantization via Linde-Buzo-Gray), and GMM (Gaussian Mixture Modelling) algorithms are used to generate template and feature matching purpose.

KEYWORDS: Speaker Identification, feature mapping, feature extraction, MFCC, VQLBG , GMM, Real Time, MATLAB

I. INTRODUCTION

Speaker recognition is the identification of a person from characteristics of voices (*voice biometrics*).It is also called voice recognition. Speaker recognition isa technique of recognizing automatically who is speaking on the basis of individual information included in speech waves. This technique uses the speaker's voice to verify their identity and provides control access to services such as database access services, voice dialling, voice mail, information services, security control for confidential information areas, remote access to computers and several other applications where security is the main issue.

Speech is a complicated signal produced as a result of several transformations occurring at several different levels: semantic, linguistic, articulatory, and acoustic. Changes in these transformations are reflected in the differences in the acoustic properties of the speech signal. Also theirdifferences related to speakers which are a result of a combination of anatomical differences intrinsic in the vocal tract and the speaking habits of different individuals. In speaker identification, all these differences are considered and used to distinguish between speakers.

Figure shows block diagram of biometric system. First block represents sensor which is an interface between the real world and the system i.e. it has to captures all the necessary data from the real world. The second block performs all the necessary pre-processing i.e. it has to remove all background noise from the sensor to enhance the input. Third block represents features extraction in which necessary features are extracted. The next feature matching step generates templates and performs matching operation.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

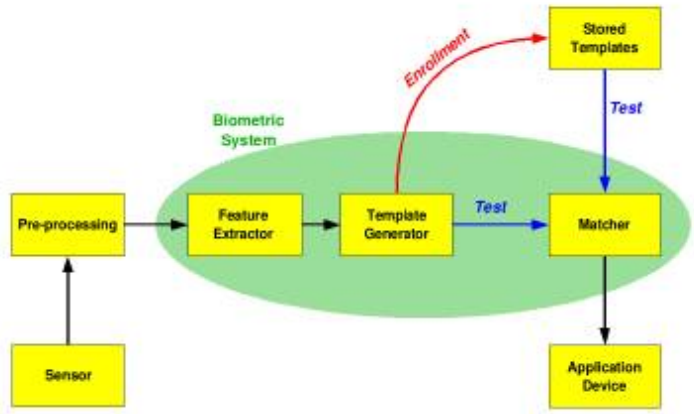

Fig 1: Block Diagram of Biometric system

Following figure shows block diagram of speaker recognition technique.In speaker recognition we will use microphone instead of sensor.

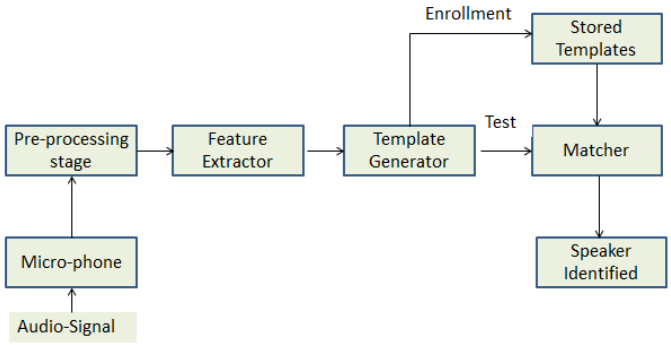

Fig 2- Block Diagram of Speaker identification

Speaker identification is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. Similar to a biometric system, it has two phases: -

Enrollment or Training phase
In the training phase, each speaker has to provide their voice samples so that the system can build a reference model for that speaker.
Operation or Testing phase
During the testing phase, the input speech signal is matched with stored reference model(s) and identification decision is made.
In real time speaker recognition technique we create two folders: one for training data and one for testing data. In training folder each speaker has to provide their voice sample then features are extracted using feature extracting algorithm as explained below. Then the system create database for all samples.During testing all speakers provide their voice sample which is automatically stored in testing folder then features are extracted and matched with stored database of training folder similarly this matching function is performed by one of feature matching algorithm as explained below, all these functions are performed by using MATLABand identification decision is made.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

II. FEATURE EXTRACTION

In the feature extraction stage we extract useful features from the speech signals like pitch variation, different style of speaking same word by different people, some speakers put stress on specific part of the word and other do not .
The speech signal is a called a quasi-stationary signal i.e. a slowly timed varying signals. An example of a speech signal is shown in Fig

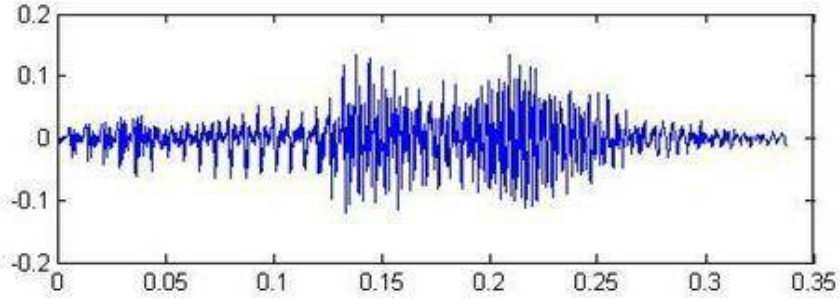

Fig 3: Speech signal

Various pre-processing tasks must be performed before extracting the features. The speech signal needs to undergo various signal conditioning steps before being subjected to the feature extraction methods. These tasks contain: -

- Truncation
- Frame blocking
- Windowing
- Short term Fourier Transform

MFCC (Mel FrequencyCepstral Coefficients)
MFCC is most popular feature extraction technique used for speaker identification. Figure below gives the block diagram of a MFCC processor.

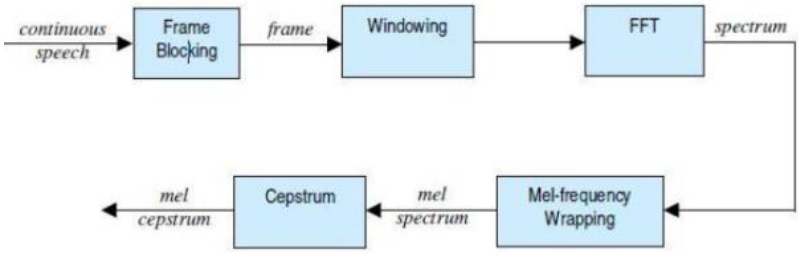

Fig 3 : Block Diagram of MFCC processor

As given in the block diagram the continuous speech signalare subjected to frame blocking, windowing usually hamming window is used and FFT in the pre-processing step,Generally FFT is used to increase the speed of processing. The result of the later step is the spectrum of the signal.The logarithmic mel-scaled filter bank is applied to the Fourier transform frame. This is expressed in mel-frequency scale, which is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz have been used to capture the phonetically important characteristics of speech. .MFCCs use Mel-scale filter bank .In final step we calculate Discrete Cosine Transform (DCT) of the output from the filter bank. For each Speech Frame, A set of MFCC is calculated.This set of coefficient is called an acoustic vector which represent the important characteristics of speech and very important for further processing in speech identification.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

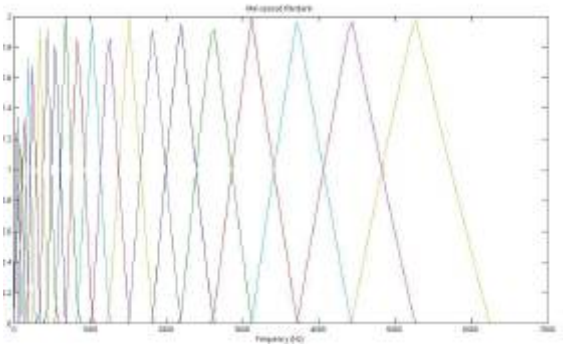

Fig 4: Continuous speech signal

III. FEATURE MATCHING

After extraction of the feature vectors from the speech signal, the next step is the feature matching. Matching of the patterns is required to be carried out for speaker identification. F3eature matching can be carried out usingalgorithms like VQLBG, DTW and stochastic models such as GMM, Here we focused only on VQLBG and GMM.

VQLBG(Vector Quantization)
Vector quantization (VQ in short) is one of feature matching algorithm. It takes a large set of feature vectors of a registereduser and produce a smaller set of feature vectors that represent the centroids of the distribution, i.e. points spaced so as to minimize the average distance to every other point. During the training phase code books are generated for each speaker using VQ method. For each code wordyi, there is a "nearest neighbor" region associated with it called Voronoiregion and is defined by

$$V_i = \left\{ x \in R^k : \|x - y_i\| \le \|x - y_j\|, \text{for all } j \ne i \right\}$$

The Voronoi region associated with the given centroid is cluster region for the vector. The Euclidean distance is defined by:

$$d(x, y_i) = \sqrt{\sum_{j=1}^{k} (x_j - y_{ij})^2}$$

wherexj is the jth component of the input vector, and yij is the *j*th component of the codeword*yi*.
Below is the block diagram of a speaker identification model using VQ:-

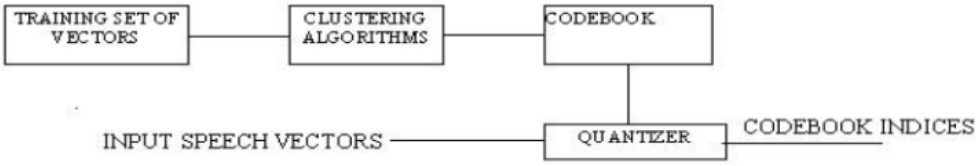

Fig 5 : Block Diagram of the basic VQ Training and classification structure

Speaker identification can be done using the code book generated for each speaker using VQ method. Figure below best defines the process:-

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

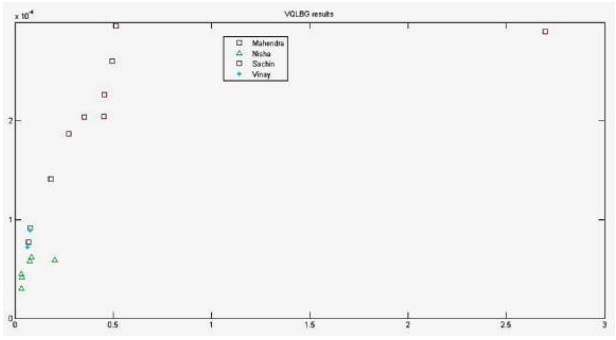

Fig 6: VQ for all samples

As shown in the figure, there are four users for which features are extracted and VQ has been performed. For each speakercodebooks are generated. During the testing phase, feature vectors of unknown speaker are mapped to feature space assuming the unknown speaker will be one of the register speakers. Then for each feature vector Euclidian distance is calculated for a given code book to find the nearest code word. The least distance found is termed as VQ distortion. Similarly distortions are calculated for the remaining feature vectors and summed up. Same procedure is repeated for rest of the speaker. The least sum of the VQ distortions will give the desired user.

GMM (Gaussian Mixture Modelling)

For speaker identification, GMM is one of the non-parametric methods. Feature vectors are displayed in D-dimensional feature space after clustering, they some-how look like Gaussian distribution. It means each matching cluster can be viewed as a Gaussian probability distribution and features fitting to the clusters can be best characterised by their probability values. The only difficulty lies in efficient classification of feature vectors. The use of Gaussian mixture density for speaker identification is driven by two facts.
They are:
1- Individual Gaussian classes are interpreted to represents set of acoustic classes. These acoustic classes represent vocal tract information.

2- Gaussian mixture density provides smooth approximation to distribution of feature vectors in multi-dimensional feature space. A Gaussian mixture density is weighted sum of M component densities and given by the equation:-

Figure 7 gives a better understanding of what GMM really is:-

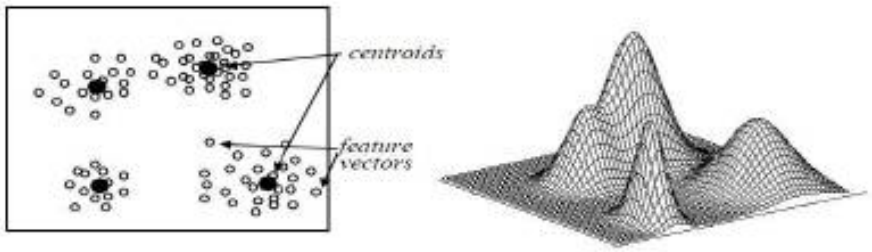

Fig 7: Gaussian Mixture Modelling

A Gaussian mixture density is weighted sum of M component densities and given by the equation:-

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x})$$

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

where x refers to a feature vector, pi stands for mixture weight of ith component and bi (x) is the probability distribution of the ith component in the feature space. As the feature space is D-dimensional, the probability density function bi (x) is a D-variate distribution [13]. It is given by the expression:-

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)' \Sigma_i^{-1} (\vec{x}-\vec{\mu}_i)\right\}$$

Whereμi is the mean of ith component and Σ i is the co-variance matrix [13].
The complete Gaussian mixture density is represented by mixture weights, mean and co-variance of corresponding component and denoted as:-

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\} \quad i = 1, \ldots, M.$$

IV RESULTS

The results of feature extraction and matching methods are shown in this section.
Feature Extraction
In this we extract features from these sampled values. Below figures shows the feature vectors and their MFCC

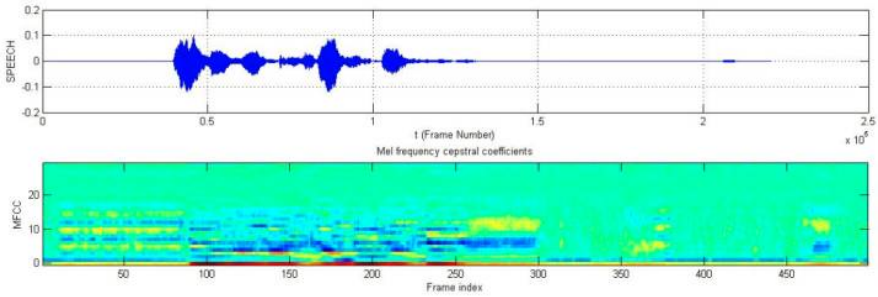

Fig8:-Output ssof MFCC

Feature Matching:

A. VQ Using LBG Algorithm

In VQ we convert the large set of feature vector into small set of feature vector and then cluster them to “Voronoi” region and each region can be represented by code book which is a template for a given speaker. When an unknown speaker is encountered, feature vectors are calculated and VQ distortion for each feature is calculated and summed up. Below figure shows the confusion matrix and GUI for VQ.

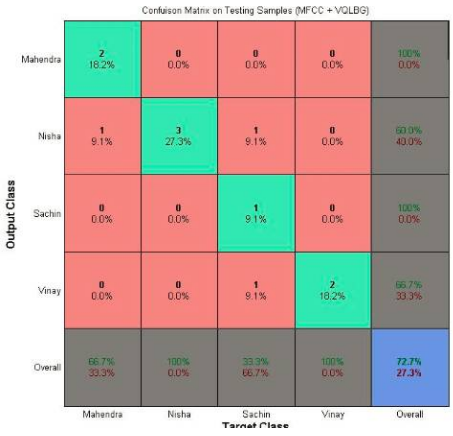

Fig 9 : VQ with MFCC Confusion matrix

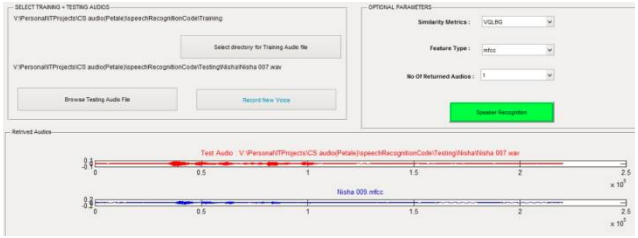

Fig 10 : GUI Result for VQ & MFCC

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Each row & column refers to unknown speaker and register speaker respectively. Each element of a single row is the total VQ distortion obtained by using the code book of corresponding registered speaker. To explain the process we can just focus on row5. Here1st speaker (Mahendra) has overall accuracy of 66.7%. Hence the algorithm assigns speaker 1 as identity for 5throw . This is same for all unknown speakers. Overall accuracy for this algorithm obtained 72.5%.

B.*GMM (Gaussian Mixture Modelling)*

GMM assumes vector space to be divided into specific components depending on clustering of feature vectors and frames the feature vector distribution in each component to be Gaussian. As initially we have no idea about which vector belongs to which component a likelihood maximization algorithm is followed for optimal classification. For testing purpose we calculated posteriori probability of test utterance and the reference speaker maximizing Gaussian distribution is termed as identity of unknown speaker. The confusion matrix and GUI of this procedure is given below:-

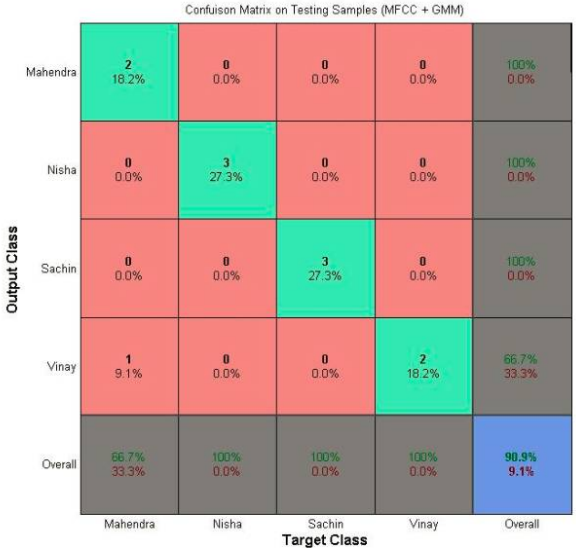

Fig11 : GMM with MFCC confusion matrix

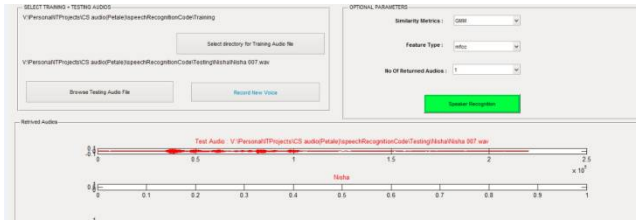

Fig 12 : GUI Result for GMM & MFCC

Overall accuracy for this algorithm was 90.9%

V. CONCLUSION

The results obtained using MFCC and VQ are considerable. Accuracy obtained using VQLBG algorithm was 72.5%. It can be improved by taking voice sample using high quality audio devices in a noise free environment. Use of more number of centroids increases the performance factor but degrades the computational efficiency. Hence an economical trade-off between code vectors and number of computation is required for optimized performance of VQLBG algorithm. Then we went for another method for speaker identification known as GMM. Accuracy of 90.9% was obtained using GMM algorithm for same data which clearly indicates its high efficiency. When number of components used becomes high its computational efficiency degrades a bit. But when we go for its high accuracy, these gaps can be compromised.

REFERENCES

[1] Seddik, H.; Rahmouni, A.; Samadhi, M.; "Text independent speaker identification using the Mel frequency cepstral coefficients and a neural network classifier" First International Symposium on Control, Communications and Signal Processing, Proceedings of IEEE 2004 Page(s):631 – 634.
[2] Roucos, S. Berouti, M. Bolt, Beranek and Newman, Inc., Cambridge, MA; "The application of probability density estimation to text-independent speaker identification" IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '82. Volume: 7, On page(s): 1649- 1652. Publication Date: May 1982.
[3] Childers, D.G.; Skinner, D.P.; Kemerait, R.C.; "The cepstrum: A guide to processing" Proceedings of the IEEE Volume 65, Issue 10, Oct. 1977 Page(s):1428 – 1443.
[4] Campbell, J.P., Jr.; "Speaker identification: a tutorial" Proceedings of the IEEE Volume 85, Issue 9, Sept. 1997 Page(s):1437 – 1462.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

[5] Castellano, P.J.; Slomka, S.; Sridharan, S.; "Telephone based speaker identification using multiple binary classifier and Gaussian mixture models" IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 Volume 2, Page(s) :1075 – 1078 April 1997.
[6] Zilovic, M.S.; Ramachandran, R.P.; Mammone, R.J "Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions".; IEEE Transactions on Speech and Audio Processing, Volume 6, May 1998 Page(s):260 – 267
[7] Davis, S.; Mermelstein, P, "Comparison of parametric representations for monosyllabic word identification in continuously spoken sentences" , IEEE Transactions on Acoustics, Speech, and Signal Processing Volume 28, Issue 4, Aug 1980 Page(s):357 – 366
[8] Y. Linde, A. Buzo& R. Gray, "An algorithm for vector quantizer design", IEEE Transactions on Communications, Vol. 28, issue 1, Jan 1980 pp.84-95.
[9] S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum", IEEE Transactions on Acoustic, Speech, Signal Processing, Vol.34, issue 1, Feb 1986, pp. 52-59.
[10] Fu Zhonghua; Zhao Rongchun; "An overview of modeling technology of speaker recognition", IEEE Proceedings of the International Conference on Neural Networks and Signal Processing Volume 2, Page(s):887 – 891, Dec. 2003.
[11] PRADEEP. CH," TEXT DEPENDENT SPEAKER RECOGNITION USING MFCC AND LBG VQ", National Institute of Technology, Rourkela, 2007
[12] Eamonn J. Keogh† and Michael J. Pazzani‡ ," Derivative Dynamic Time Warping", Department of Information and Computer Science University of California, Irvine, California 92697 USA
[13] Douglas A. Reynolds and Richard C. Rose, "Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models", IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 3, NO. 1, JANUARY 1995
[14] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from incomplete data via the EM algorithm, "J.Royal Stat. Soc., vol 39, pp. 1-38, 1977.
[15] Reynolds D.A.: "A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification", Ph.D. thesis, Georgia Institute of Technology, September 1992.