International Journal of Computer Network and Security(IJCNS)
Vol 4. No 1. Jan-Mar 2012 ISSN: 0975-8283
www.ijcns.com

Performance Analysis of Speaker Identification System Using GMM with VQ

M.G.Sumithra[1],A.K.Devika[2]
Professor, Department of ECE,PG student, Department of ECE,
Bannari Amman Institute of Technology, Sathyamangalam,Tamilnadu,India.
mgsumithra@rediffmail.com
devikaakadavath@yahoo.com

Abstract

Personal identity identification is an important requirement for controlling access to protected resources. Biometric identification by using certain features of a person is a more secured solution for security identification. Advances in speech processing technology and digital signal processors have made possible the design of high-performance and practical speaker recognition systems. A more flexible speaker identification system is able to operate without explicit user cooperation and independency of the spoken utterance (text-independent mode).This paper proposes a system for text independent speaker identification by extracting MFCC features and implementing optimized GMM speaker modeling. Expectation Maximization algorithm is used to compute the GMM parameters. Performance of the proposed system is evaluated based on its identification accuracy.It is compared with the system using VQ speaker modeling technique. A TIMIT database of 100 speakers is used to study the performance of the proposed system.

Key terms: Feature extraction, Speaker modeling, vector quantization, speaker identification, Mel-frequency cepstral coefficients(MFCC), Gaussian mixture model(GMM),Gaussian mixture model-Expectation maximization(GMM-EM)

1. INTRODUCTION

Technology of speaker recognition is widely adopted for the growing needs of entry control and security management. Certain applications include the verification of person's identity whereas some other applications require personal identity identification. Speaker identification is a challenging pattern classification task. Campbell defines speaker identification (SI) process as: "use of a machine to recognize a person from a spoken phrase" [1]. It is used enormously in many applications such as security systems, information retrieved services, etc. Portable identification systems are expected to be widely used in future in many purposes, such as mobile applications. The basic speaker identification system consists of two phases-training phase and testing phase. In training phase, the speech signals of all the speakers are acquired, their respective features are extracted, feature matrix is formed and finally they are stored in a database along with proper labeling while in the testing phase, the speech signal of the unknown speaker is acquired, corresponding feature matrix is generated and finally compared with the matrices in the stored database. The label of best possible match is intimated to the identity of the unknown speaker.The process of extracting these features is known as feature extraction which in turn is converted to feature matrix and the process of comparing the stored database with the unknown speaker database is known as feature matching. Success in speaker identification depends on extracting and modeling the speaker dependent characteristics of the speech signal which can effectively distinguish one talker from another.

Two forms of speaker identification are typically distinguished, namely text-dependent and text-independent identification. In a text-dependent setup, a predetermined group of words or sentences are used to enroll a set of speakers, and these words or sentences are then used to verify the speakers. In a text-independent system, no constraint is placed on what can be said by the speaker[2]. In text-dependent applications, hidden Markov models (HMMs) can have some advantages when incorporating temporal knowledge, at the same time GMMs still show the best performance to date for text-independent speaker recognition with high accuracy [3]. However, to have a detailed description of the acoustic space and also achieve good identification performance, the number of Gaussian mixture components in each model is usually large, especially when diagonal covariance matrices are used. The disadvantage of GMM is that it requires sufficient data to model the speaker well. To overcome this problem, Reynolds et al. introduced GMM–universal background model (UBM) for the speaker recognition task [4]. The disadvantage is that a gender-balanced large speaker set is required for UBM training [5].

In this paper an efficient text independent speaker identification system was designed using GMM speaker modeling technique and its identification accuracy is compared with system using VQ as the speaker model. The

comparison is also done with other feature extraction techniques.

This paper is organized as follows: In section 2 the principle feature extraction method is discussed**.** Section 3 deals with speaker modeling technique. Section 4 discusses the performance evaluation. In section 5 and 6, the experimental results and conclusion are discussed.

2. FEATURE EXTRACTION

From the automatic speaker identification task point of view, it is useful to think about speech signal as a sequence of features that characterize both the speaker as well as the speech. It is an important step in identification process to extract sufficient information for good discrimination in a form and size which is amenable for effective modelling.Feature extraction transforms the raw speech signal into a compact but effective representation that is more stable and discriminative than the original signal.Before extracting the features the speech signal is preprocessed using following steps.

i) Framing ii) Windowing

Initially the continuous speech signal is divided into frames where each frame consists of Msamples. Very often successive frames are overlapping with each other by *M* samples [6]. For our experiments, we have used window size of *M* = 512 and frame size *N*=100.Windowing is carried in order to prevent an abrupt change at the end points of the frame for which it isusually multiplied by a window function and those segments are called windowed frames. Hamming windowis used to multiply each frame.

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{M-1}\right) \quad (1)$$

where M is the number of samples in each frame.

A. Mel frequency cepstral coefficient(MFCC)

MFCC is one of the most popular methods for extracting features from the speech signal developed by Davis and Mermelstain [7]. MFCC's are shown to be less susceptible to the variation of the speaker's voice and surrounding environment.It is based on the known variations of human ears. Critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech.

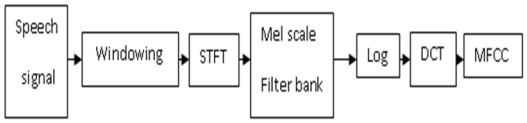
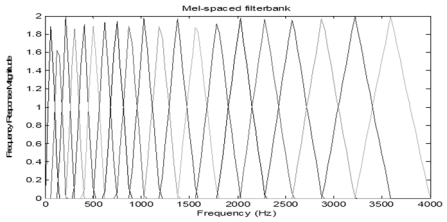


Fig.ure. 1 Extraction of MFCC from speech signals

A Mel is a unit of measure of perceived pitch or frequency of a tone. It does not correspond linearly to the normal frequency,it isapproximately linear below 1 kHz and logarithmic above.One useful way to create mel-spectrum is to use afilter bank, one filter for each desired mel-frequency component. Every filter inthis bank has triangular bandpass frequency response[8]. The following function transforms real (linear frequency) to mel frequency.

$$\text{Mel}(f) = 2595\log\left(1+\frac{f}{700}\right) \quad (2)$$



Figue.2 Mel spaced Filter Banks

The next step is to convert the log mel spectrum back to time. This is done by means of Discrete Cosine Transform.The result is called the mel frequency cepstral coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis.

3. CREATION OF SPEAKER MODEL

A. Vector quantization(VQ)

A speaker identification system must able to estimate probability distributions of the computed feature vectors. Storing every single vector that generate from the training mode is impossible, since these distributions are defined over a high-dimensional space. It is often easier to start by quantizing each feature vector to one of a relatively small number of template vectors, with a process called vector quantization. It is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a codeword [9].

The collection of all code words is called a codebook. The training material is used to estimate the code book**.** Here a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors hence, a vector quantizer *Q* of dimension *k* and size N is a mapping from a vector in the *k*-dimensional space into one of N centroids in the space. Thenext important step is to build a speaker-specific VQ codebook for this speaker using those training vectors.

There is a well-known algorithm, namely the LBG algorithm [10], for clustering a set of L training vectors into a set of M codebook vectors.

B. LBG algorithm

The algorithm is formally implemented by the following recursive steps:

STEP 1

Design a 1-vector codebook; this is the centroid of the entire set of training vectors (hence, no iteration is required here).

STEP 2

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

$y^+(n)=y(n)(1+\varepsilon)$ (3)

$y^-(n)=y(n)(1-\varepsilon)$ (4)

where nvaries from 1 to the current size of the codebook and ε is the splitting parameter (choose ε = 0.01).

STEP 3

Nearest-Neighbor Search: for each training vector, find the codeword in the current codebook that is closest (in terms of similarity measurement), and assign that vector to the corresponding cell (associated with the closest codeword).

STEP 4

Centroid Update: update the codeword in each cell using the centroid of the training vectors assigned to that cell.

STEP 5

Iteration 1: repeat steps 3 and 4 until the average distance falls below a preset threshold.

STEP 6

Iteration 2: repeat steps 2, 3 and 4 until a codebook size of M is designed. Intuitively, the LBG algorithm designs an M-vector codebook in stages. It starts first by designing a 1-vector codebook, then uses a splitting technique on the code words to initialize the search for a 2-vector codebook, and continues the splitting process until the desired M-vector codebook is obtained

C. Gaussian Mixture Model-Expectation Maximization

The distribution of feature vectors extracted from a person's speech is next modeled by a GMM – Gaussian Mixture Model.The output mixture density of GMM is linear combination of M components – normal (Gaussian) distributions, called mixtures. The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities[11]. For speaker identification, each speaker is represented by a GMM and is referred to by his/her model λ .The model λ are collectively represented by the notation

$$\lambda = \{\rho_i, \mu_i, \Sigma_i\} \; i = 1,\ldots, M \qquad (5)$$

Given a collection of training vectors, maximum likelihood model parameters are estimated using the iterative expectation-maximization (EM) algorithm[12]. The EM algorithm iteratively refines the GMM parameters to monotonically increase the likelihood of the estimated model for the observed feature vectors, that is, for iterations k and k+1, $\rho(X|\lambda(k+1)) \geq \rho(X|\lambda(k))$. The basic idea of the EM algorithm is, beginning with an initial model λ, to estimate a new model, such that,

$$\rho(X \setminus \bar{\lambda}) \geq \rho(X \setminus \lambda) \qquad (6)$$

The new model then becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached. This is the same basic technique used for estimating HMM parameters via the Baum-Welch re-estimation algorithm. On each iteration, the following re-estimation formulas are used which guarantee a monotonic increase in the model's likelihood value.

Mixture weights

$$\vec{\rho}_i = \frac{1}{T}\sum_{t=1}^{T} \rho\left(i / \vec{x}_t, \lambda\right) \qquad (7)$$

Variance

$$\overline{\sigma_i^2} = \frac{\sum_{t=1}^{T} \rho\left(i / \vec{x}_t, \lambda\right) x_t^2}{\sum_{t=1}^{T} \rho\left(i / \vec{x}_t, \lambda\right)} - \overline{\mu_i^2} \qquad (8)$$

Mean

$$\overrightarrow{\overline{\mu}_i} = \frac{\sum_{t=1}^{T} \rho\left(i \setminus \overrightarrow{x_t}, \lambda\right) \overrightarrow{x_t}}{\sum_{t=1}^{T} \rho\left(i \setminus \overrightarrow{x_t}, \lambda\right)} \qquad (9)$$

EM computes the mean vector, weights and variance in each iteration and the values are updated using the same formula. The use of expectation maximization (EM) optimization procedure with GMM has established very good

results. Using the mean vector created by EM algorithm, the speaker model was generated by LBG algorithm, a vector quantization technique.

D. Feature Matching

After creating the speaker model, a nearest neighbor search has to be performed for each training vector. This is achieved by calculating the Euclidean distance between the respective model and the training vector, for every training vector. Euclidean Distance(ED) is given as:

$$ED=\sqrt{(x-x_1)^2+(y-y_1)^2} \quad (10)$$

Where, (x, y) is co-ordinates of trained speaker,(x1, y1) is co-ordinates of unknown speaker**.** The vector with the smallest Euclidean distance is assigned with that vector. In speaker identification system, only the distance between the claimed user's speaker model and the identification feature vectors is calculated. Based on the threshold, the user is said to be a true speaker, otherwise the user is said to be a false speaker or imposter.

4. PERFORMANCE ANALYSIS

The performance of a speaker identification system is measured in terms of false acceptance rate (FA %) and false rejection rate (FR %). False acceptance error consists of accepting identity claim from an imposter. False rejection error happens when a true speaker is rejected. It is represented as:

$$F_A=\left(I_A/I_T\right)\times 100 \quad (11)$$

$$F_R=\left(C_A/C_T\right)\times 100 \quad (12)$$

$$T_E=F_A+F_R \quad (13)$$

where, I_A-No of Imposter classified as true speakers, I_T - Total no of speakers, F_A-False Acceptance, F_R- False Rejection,T_E - Total error of identification system, C_A - No of true speakers classified as Imposters,C_T - Total no of speakers

5. RESULTS AND DISCUSSION

In this paper, an enrolled database of 100 speakers is created. To analyze the speech signal, hamming window is used. Each frame contains 512 samples with an overlap of 100 samples into the consecutive frame and the sampling rate used is 8000 Hz. The number of filter banks used for MFCC is 40. 8 samples of a speaker were used in which 5 samples were used for testing and 3 were used for training.

Initially the speech signal of the unknown speaker is acquired and features are extracted using any of the feature extraction techniques. This yields the feature vectors .Feature extraction techniques like MFCC,MMFCC,RPLP,LPCC and BFCC were used for analysis in which MFCC feature extraction technique yields maximum identification accuracy with 128 initial centroids. Fig.3shows the comparison of identification accuracy of proposed system and VQ system for different feature extraction techniques.

The feature vectors of the unknown speaker are combined together to form the feature matrix. The feature matrix thus formed is compared with the vector quantized code book matrices present in the stored data base using VQ technique. In GMM, EM algorithm compute parameters like mean, variance and weights iteratively, are updated until some convergence threshold is reached.

Fig.4 shows the comparison of identification accuracy for GMM system and VQ system respectively with a distance minimum value equal to 4. It is observed that GMM has obtained an identification accuracy of 98.87% with 128 number of gaussians since it exhibits less false rejection rate when compared to VQ modeling technique a with higher number of centroids.

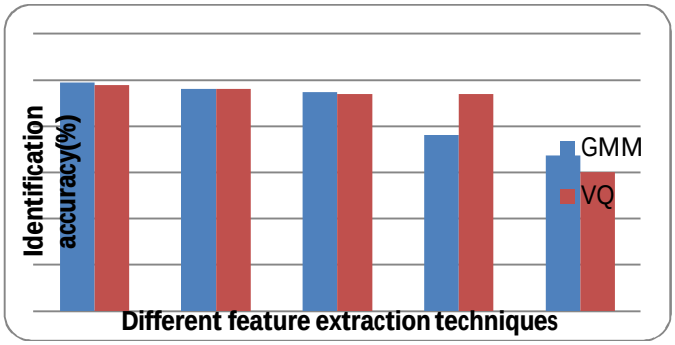


Figure. 3 Comparison of identification accuracy of GMM and VQ system for different feature extraction techniques

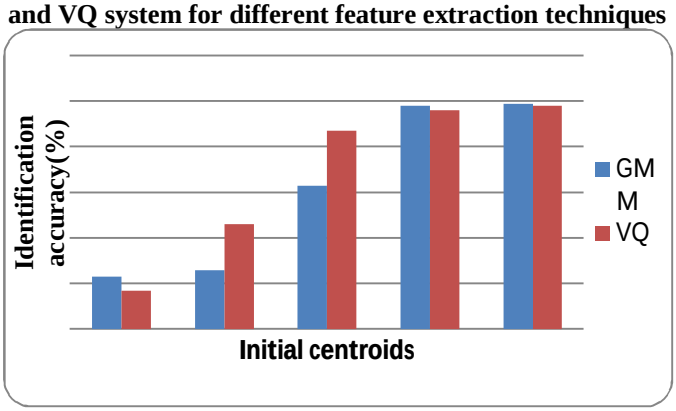


Figure 4Comparison of identification accuracy of GMM

and VQ system using MFCC as the feature extraction technique

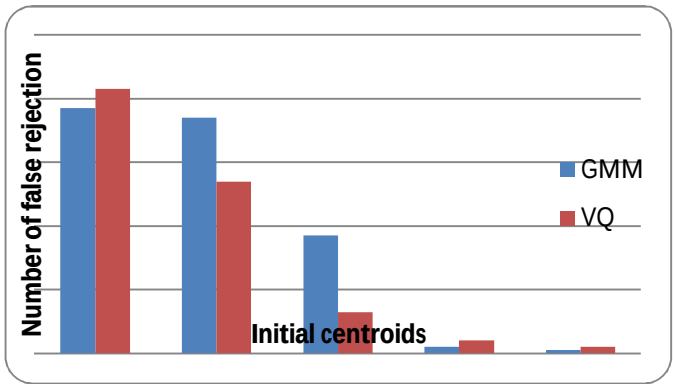


Figure. 5 Comparison of false rejection rate of GMM and VQ system using MFCC as the feature extraction technique

Fig 5 compares the false rejection rate of identification system with GMM andVQ speaker modeling technique.But it is observed that the false acceptance rate is lesser for VQ system on increasing the number of centroids as shown in Fig.6.

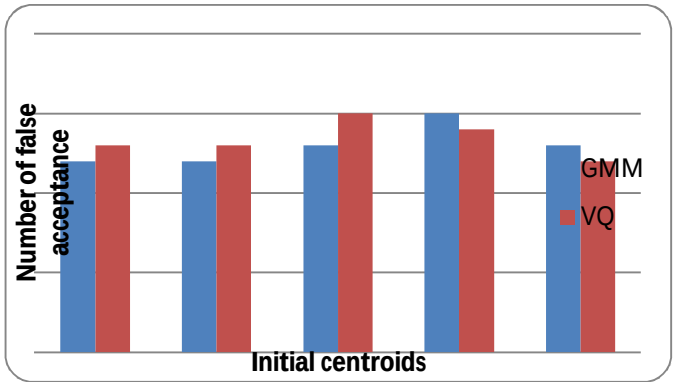
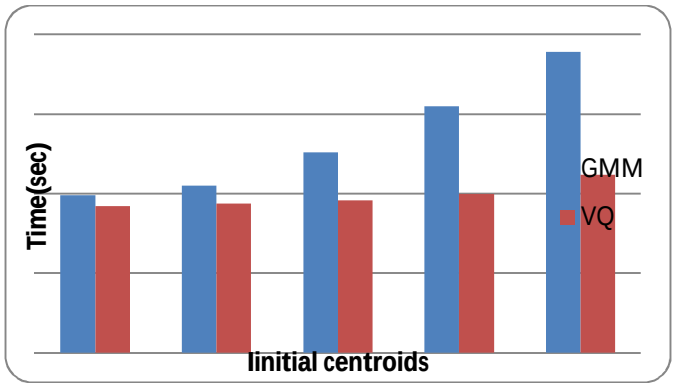


Figure. 6Comparison of false acceptance rate of GMM and VQ system using MFCC as the feature extraction technique

Computation time for training and testing the system is an issue while implementing the speaker identification into real time applications. Fig .7shows the variation in computational time for the two identification systems. Even though GMM consumes more computation time, it yields higher identification accuracy with lesser false rejection rate than VQ.



Figure/ 7 Comparison of computational time for different feature extraction techniques

6. CONCLUSION

An automatic speaker identification system has been proposed by extracting the MFCC features and by using GMM as the speaker modeling technique. A total number of 100 speakers were trained and tested. The proposed system yields an identification accuracy of 98.87% which is 1.01% higher than the system with VQ as the speaker modeling technique.This is because it has a very less false acceptance and false rejection rate. Further the system performance can be analyzed by increasing the size of database.The performance of the system can be increased by using GMM-UBM model and support vector machines.

REFERENCES

[1] J.P. Campbell, Speaker recognition: A tutorial,*proceedings of IEEE*, Vol. 85(9), pp. 1437-1462,1997.

[2] Herbert Gish and Michael schimdt, ”Text Independent Speaker Identification” IEEE signal processing magazine, October 1994

[3] Bing Xiang, Member, IEEE, and Toby Berger, “Efficient Text-Independent Speaker Verification with Structural Gaussian Mixture Models and Neural Network”, IEEE Transactions On Speech And Audio Processing, Vol. 11, No. 5, September 2003

[4] D.A.Reynolds, T.F.Quatieri, and R.B.Dunn, “SpeakerVerification Using Adapted Gaussian Mixture Models”,Digital Signal Processing, vol.lO, pp.19-41, Jan.2000

[5] Jayanna, H. S. and Mahadeva Prasanna, S. R. (2009) “Analysis, Feature Extraction, Modeling and Testing

Techniques for Speaker Recognition", *IETE technical review,* Vol .26

[6] L. Rabiner and B. H. Jaung ,"Fundamentals of Speech recognition", Prentice Hall Englewood Cliffs, New Jersey, 1993.

[7] S.B. Davis and P. Mermelstein ,Comparison of Parametric representations for Monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustics,Speech, Signal Processing, vol. ASSP-28(4),pp.357-366, August 1980

[8] Jeill-weill Hllng, "Optimization of Filter·bank to improve the extraction of Mfcc features in Speech Recognition"in proceedings of international Symposium on Intelligent Multimedia. Video and Speech Processing,2004

[9] L. Tue, H. Anant,E. Deniz, and C. Jose, "Vector Quantization using information theoretic concepts", Natural Computing: an international journal, vol .4, Issue. 1, pp. 39 - 51, January 2005

[10] Linde Y., Buzo A., Gray, R , "An Algorithm for Vector Quantizer Design," IEEE Transactions on Communications. Vol. 28(1), 84-95. 1980

[11] Reynolds, D.A.' Speaker identification and verification using Gaussian mixture speaker models', Speech Communication, vol. 17, pp. 91-108,1995

[12] Douglas.A.Reynolds and Richard.C.Bose ,"Robust Text Independent using Gaussian Mixture Models", IEEE Transactions on Speech and Audio Processing, vol 3,January 1995