International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307, Volume-1, Issue-3, July 2011

COMPARISON OF DIFFERENT PARAMETERS USED IN GMM BASED AUTOMATIC SPEAKER RECOGNITION

Archana Shende, Subhash Mishra , Shiv Kumar

***Abstract*—The performance of Speaker recognition systems has improved due to recent advances in speech processing techniques but there is still need of improvement. In this paper we present the comparison of different parameters used in automatic speech recognition system to increase the accuracy of the system. The main goal here is a detailed evaluation of the parameters used in Automatic speech recognition system such as window type, MFCC frame size, number of Gaussian mixtures and GMM & VQ/GMM technique .In this paper we propose a decision function by using vector quantization techniques to decrease the training model for GMM in order to reduce the processing time.**

Index Terms: **Gaussian Mixture Model (GMM), Mel Frequency Cepstral Coefficient (MFCC), Speaker Identification (SI), Speaker Verification (SV), Vector Quantization (VQ).**

I. INTRODUCTION

The recognition of a human being through his voice is one of the simplest forms of automatic recognition because it uses biometric characteristics which come from a natural action, the speech. Speech, may be the cheapest form to supply a growing need of providing authenticity and privacy in the worldwide communication networks [1]. Speaker recognition refers to two fields: Speaker Identification (SI) and Speaker Verification (SV) [2], [3]. In speaker identification, the goal is to determine which one of group of known voices (closed set) best matches the input voice sample. Speaker verification is the task of verifying if a speech signal belongs or not to a certain person, which means a binary decision. There are two tasks: text-dependent and text-independent speaker identification. In text-dependent identification, the spoken phrase is known to the system whereas in the text-independent case, the spoken phrase is unknown. Speaker identification system involves two main stages, the enrolment stage and the verification stage. These phases involve two main parts: Feature Extraction & Pattern Classification. Feature extraction and pattern classification model are the basic parts in speaker identification and verification systems. In our implementation, we will use MFCC technique to extract the speech feature in order to obtain the best result for pattern classification. For pattern classification part , There are proposed a lot of methods for speaker modeling and recognition. In text dependent speaker recognition the most popular methods are dynamic time warping (DTW), Hidden Markov Models (HMM) [2]. In text independent speaker recognition the most popular methods are: Vector Quantization (VQ) [4], fully connected (ergodic) HMM‘s, artificial neural networks (ANN), support vector machines (SVM), and Gaussian Mixture Models (GMM) [6], [7].Currently ,statistical based methods such as variants of Gaussian Mixture Models (GMM) are the most powerful methods for speaker recognition ,including speaker identification and speaker verification.

In this paper we would like to propose text independent speaker recognition with

Manuscript received June 3, 2011.

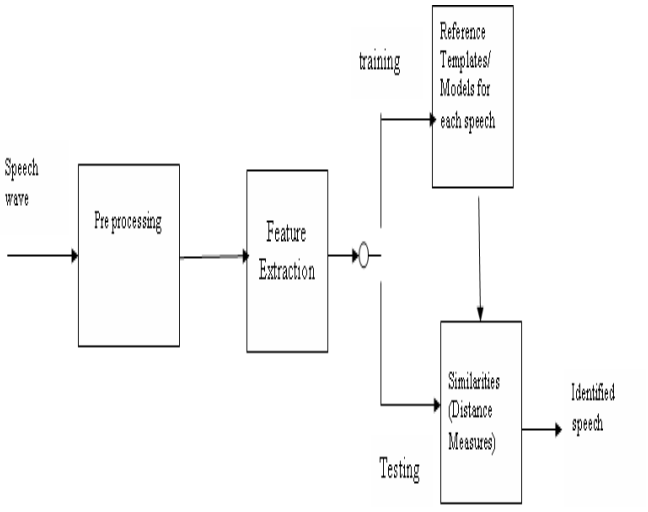
Archana Shende, Department of Electronics and Communication (M.Tech. Scholar), Technocrat Institute of Technology-Bhopal (M.P.), India (e-mail: archanashende01@gmail.com).

Prof. Subhash Mishra, Department of Electronics and Communication, Technocrat Institute of Technology-Bhopal (M.P.), India (e-mail: subhashmishra67@gmail.com).

Prof. Shiv Kumar, Asst. Professor, Department of Information Technology, Technocrat Institute of Technology-Bhopal (M.P.), India (e-mail: shivksahu@rediffmail.com).

II. AUTOMATIC SEARCH REGOGNITION

Speech recognition is the process of automatically recognizing who is speaking on the basis of individuality information in speech waves. The three important components of speaker recognition are preprocessing. Feature Extraction, Speaker modeling or classification system.



[Figure-1- Components of Automatic Speaker Recognition]

Pre-processing is the first phase of a speaker recognition task. Pre-processing consists of pre-emphasis, end-point detection, linear time alignment normalization

and a 512 point frame windowing with 256 point overlapping.

A. Windowing

Windowing is performed to avoid unnatural discontinuities in the speech segment and distortion in the underlying spectrum. We can use three types of window Triangular, Rectangular & hamming window.In speaker recognition, the most commonly used window shape is the hamming window.If we define the window $w(n)$ as $0 \le n \le N-1$, where N is the number of samples in each frame, then the result of windowing is the signal $y_i(n) = x_i(n)w(n)$, $0 \le n \le N-1$

Typically, the Hamming window is used, which has the form:

W(n) = 0.54 – 0.46 cos $\left(\frac{2\pi n}{N-1}\right)$, $0 \le n \le N-1$

(source-Book- Thomas s .Quitery)

The Mel Frequency Cepstrum Coefficient - The Mel Frequency Cepstrum Coefficient (MFCC) is used to resolve the speech signal into sum of two components[5].

Mel Filterbank- The equation below shows the relationship between frequency in hertz and mel scaled frequency. Frequency (mel scaled) = 2595 log (1+ f (Hz) / 700) .

In order to perform mel-scaling, a number of triangular filters or filterbank are used. Mel scale is less vulnerable to the changes of speaker's vocal cord in course of time.

III. THE GAUSSIAN MIXTURE MODEL (GMM)

The Gaussian mixture model (GMM) is a density estimator and is one of the most commonly used types of classifier[6],[7]. In this method, the distribution of the feature vector x is modeled clearly using a mixture of M Gaussians.

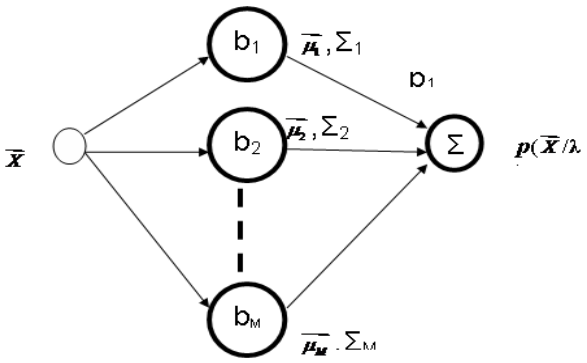
A Gaussian mixture model is modeled by many different Gaussian distributions. Each of the Gaussian distribution has its mean, variance and weights in the Gaussian mixture model. A Gaussian mixture density is a weighted sum of M component densities (Gaussians) as depicted in following figure and given by equation.

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

Where x is a L dimensional vector, pi are mixture weights. bi(x) = component densities. Where i = 1,….., M. Each component density is a L variate Gaussian function of the form:

$$b_i(\vec{x}) = \frac{1}{(2\Pi)^{L/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\}$$

Where μi is the mean, ∑i is covariance matrix. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} p_i = 1$. T is the total no. of feature vectors or total no of frames. In the following figure you can see the pdf p(x/λ) is modeled using M component densities {b1(X), b2(X),b3(X),--------bM(X)}with mixture weights p1,p2---------- pM.



[Fig-2. M probability densities forming a GMM]

The mean vectors, covariance matrices and mixture weights of all Gaussians together represent a speaker model and parameterize the complete Gaussian mixture density. These parameters are collectively represented by the notation

$\lambda =$

Hence now each speaker is represented by a Gaussian mixture model with parameters described by 'λi ', where i = 1,2,3……S (S no. of speakers) .

IV. GMM PARAMETER ESTIMATION

The other task is to estimate the parameters of GMM $\overline{\lambda}$, which best matches the distribution of the training feature vectors, given by speech of the speaker. For Given observation sequence (or a set of sequences), the estimation problem involves finding the "right" model parameter values that specify a model most likely to produce the given sequence(7). The most popular method is maximum likehood (ML) estimation . The basic idea of this method is to find model parameters which maximize the likehood of GMM. For a given set of T training vectors

$X = \{\vec{x}_1 \ldots\ldots\ldots\ldots \vec{x}_T\}$

GMM likehood can be written:

$$p(x/\lambda) = \prod_{t=1}^{T} p\left(\vec{x}_t/\lambda\right)$$

ML parameter estimates can be obtained iteratively using special case of expectation-maximization (EM) algorithm. There the basic idea is, beginning with initial model λ, to estimate a new model $\overline{\lambda}$, that. $p\left(x/\overline{\lambda}\right) \ge p(x/\lambda)$

The new model then becomes the initial model for the next iteration. This process is repeated until some convergence threshold is reached. On each iteration, following re-estimation formulas are used:

Mixture weights are recalculated

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

Means are recalculated :

$$\overline{\mu}_i = \frac{\sum_{t=1}^{T} p\left(i/\vec{x}_t, \lambda\right)\vec{x}_t}{\sum_{t=1}^{T} p\left(i/\vec{x}_t, \lambda\right)}$$

Variances are recalculated

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

Then a posteriori probability for acoustic class i is given by:

$$p\left(i/\vec{x}_t, \lambda\right) = \frac{p_i b_i\left(\vec{x}_t\right)}{\sum_{k=1}^{M} p_k b_k\left(\vec{x}_t\right)}$$

V. EFFECT OF DIFFERENT NUMBER OF GAUSSIAN MIXUTE COMPONENTS AND AMOUNT OF TRAINING DATA

No objective way to determine the correct number of mixture components (model order) and the model dimension apriori. For saving the identification time, the objective is to choose the minimum number of components necessarily for adequate speaker modeling. However, too few components will not be able to accurately model the distinguished characteristics of a speaker distribution. Too many components relative to limited training data induce too many free parameters to be estimated reliably, thus degrade performance. Besides, small amount of training data is crucial to facilitate client enrolment to the system, with the trade-off that the insufficient data unable to train the model reliably(7).

VI. VECTOR QUANTIZATION SPEAKER IDENTIFICATION

Vector Quantization (VQ) is a pattern classification technique applied to speech data to form a representative set of features. It maps vectors to smaller regions called cluster. These cluster's center, centroid, are collected and will make up a codebook. The speaker identification engine are depends on the codebook to identify a speaker. In VQ training phase, Vector Quantization is executed using MFCC as input. Later on, the speaker identification engine will run the nearest-neighbor search to find the codeword in the current codebook that is closest and assign that vector to the corresponding cell. Then, its find centroids and update for each speech signal and the codebooks are created. In testing phase, a function will computes the Euclidean distance between training data and testing data. The system will identify which calculation yields the lowest value and checks this value against a constraint threshold. If the value is lower than the threshold, the system outputs an answer. vector quantization technique is used to minimize the amount of data to be handled(10).

VII. SIMULATION RESULTS

In this paper we are doing simulation at four level (1)Window type (2)Mel Frequency Cepstral Coefficient size (3) No. of Gaussian mixtures(4)GMM & VQ/GMM comparison.The ASR software is designed in such a way that the user can adjust certain parameters that is associated to the feature extraction and matching process. This enables analysis to be carried to study the effect of changing these parameters in the accuracy of the system. The following sections show the effect of changing the parameters associated to Window type , MFCC, No. of Gaussian Mixtures and vector quantization on the efficiency of the system.

A. Window Type

The system has been implemented in Matlab7.4 on windows XP platform. The result of the study has been presented in Table 1. The speech database consists of 20 speakers, Here, identification rate is defined as the ratio of the number of speakers identified to the total number of speakers tested.

[Table-1: Identification rate (in %) for different windows (using Mel Scale)]

Code book size	Triangular	Rectangular	Hamming
1	57.14	57.14	57.14
2	85.7	66.67	85.7
4	90.47	76.19	100
8	95.24	80.95	100
16	100	85.7	100
32	100	90.47	100
64	100	95.24	100

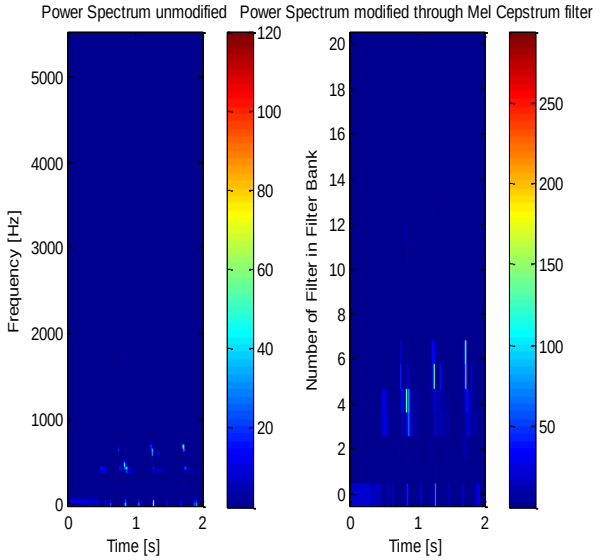
Above Table shows identification rate when triangular, or rectangular, or hamming window is used for framing in a Mel scale . The table clearly shows that as codebook size increases, the identification rate for each of the three cases increases, and when codebook size is 16, identification rate is 100% for both the triangular and hamming windows. When a codebook size is of 4 and hamming window is used 100% identification rate is obtained. Therefore in speaker recognition, the most commonly used window shape is the hamming window. The study reveals That combination of Mel frequency and Hamming window gives the best performance. It also suggests that in order to obtain satisfactory result, the number of centroids has to be increased as the number of speakers increases .

B. Mel Frequency Cepstral Coefficient Size (Frame Size)

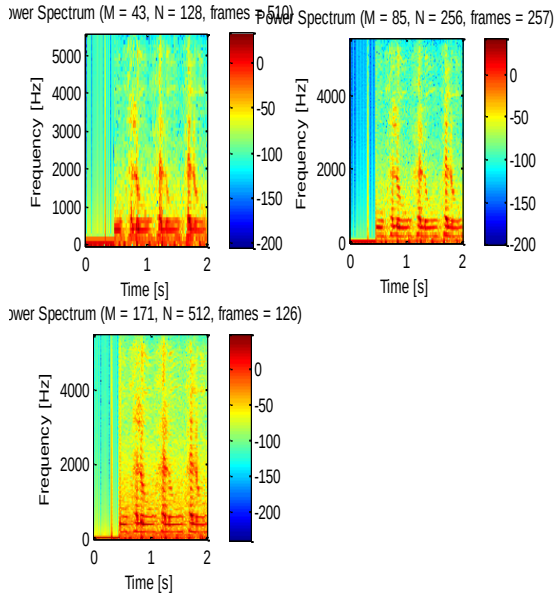
The frame size is changed from 200 to 500 samples per frame. Other parameters are left unchanged. Distance measurement for each case is shown in Table-2 .

Table-2. Matching score difference as a function of frame size

Frame size (Samples per frame)	Matching score Difference
200	1.3
300	1.4
400	1.6
500	1.1



[Fig.- 3 Power Spectrum with modified through Mel Cepstrum Filter]



[Fig.4 –Power Spectrum Plot with Different codebbok size and Frame size]

From Table-2, it can be seen that the matching score difference which is associated to the identification efficiency increases when the frame size which represents the number of samples per frame is increased. The efficiency reaches its highest point when the frame size is increased up until 400 samples per frame. Further increase in the frame size causes the efficiency to decrease. The frame size from 200 to 400 represents a time length of 20 to 40 ms. it is known that the speech signal shows quasi stationary behavior for the time interval of 20 to 40 ms. Therefore for the frame size higher than 400, the speech segment shows non stationary behavior and this causes the efficiency to decrease. Higher matching score difference shows higher efficiency because this enables the discriminative ability of the system to increase.

C. Comparison With a Fixed Number of Mixture

We have performed an experiments to find out the best number of mixtures which corresponds to the best system's performance [9] .This experiment is performed on 45 SPIDRE speakers .taking the slope 0.36 for all 45 speakers .here we present the results for 10,16,27,32,59 mixtures respectively.

[Table-3 Identification results with a Fixed Number Of Mixtures]

Number of Mixtures	***10***	***16***	***27***	***32***	***59***
Recognition (%)	***66.67***	***66.67***	***73.33***	***71.11***	***68.89***

Table-3 shows that the best result is obtained with a number of mixtures equal to 27.

D. GMM and VQ/GMM Comparison

The first method evaluated uses GMM as pattern classification techniques. The first set of experiments; we use the number of speakers from 10 to 50.

[Table-4 GMM based speaker identification performance]

Number of Speakers	**Accuracy percentage(%)**
10	98
20	95
30	91
40	89
50	83

Table-4 shows the effect of increasing the speakers on performance of the GMM speaker identification system. Accuracy starts off highly 98% as would be expected, and slowly declines to approximately 83%. As can be observed, GMM speaker verification accuracy rate has decrease when the training data increase; this is due to the complexity of the computation. Besides, it ignores knowledge of the underlying phonetic content of the speech therefore it does not take advantage of all available information.

E. Hybrid Vector Quantization/ Gaussian Mixture Model System Evaluation

The next method evaluated uses hybrid VQ decision/GMM as pattern classification techniques. This is the new hybrid pattern classification as we proposed for speaker identification system [10].

[Table- 5 Hybrid VQ decision /GMM based speaker identification performance]

Number of Speakers	Accuracy percentage(%)
10	99
20	98
30	97
40	95
50	93.38

Table-5 shows the effect of increasing the speakers on performance of the hybrid VQ/GMM speaker identification system. Accuracy starts off highly 99%, and slowly declines to approximately 93.38%. As can be observed, even hybrid VQ decision/GMM speaker identification accuracy rate has decrease when the training data increase, but it still obtain the better result if compare with baseline GMM. Besides, it seems more stable to handle the large data set.

[Table-6 COMPARISON OF TIME PROCESSING in GMM and VQ/GMM]

Algorithm	GMM	VQ/GMM
Time	62.49sec	50.78sec

The result of time processing for 10 speakers by using baseline GMM and hybrid VQ/GMM shows in table 6. We report that the baseline GMM need 62.49 seconds for the whole training and testing process whereas hybrid VQ/GMM just need 50.78 seconds. Thus, this implementation can categorized as more simplified version for classification techniques in speaker identification system. Obviously, a significant improvement compared to the baseline system is reported, a reduction in identification times up to 20% is reached. The results indicate that with the hybrid modeling, the performance of the speaker identification system is improved. Moreover, the speed of verification is significantly increased because number of features is reduced over 50% which consequently decreases the complexity of identification system.

VIII. CONCLUSION

This paper provided a performance evaluation of the model parameters used in a text independent speaker recognition system. From the above simulation result it is clear that ,the accuracy of the identification process can be influenced by certain factors such as the MFCC technique should be applied for feature extraction. It has been found that combination of Mel frequency and Hamming window gives the best performance , It also suggests that in order to obtain satisfactory result, the number of centroids has to be increased as the number of speakers increases. In a GMM based text-independent speaker identification system on increasing the amount of training data increases the identification rate. Experimental result shows that increasing the mixture components of the speaker model improves the performance, limited by amount of training data. VQ is used to minimize the data of the extracted feature. In This papar ,We are intended to improve the computation time ,the approximation quality and the accuracy of the speaker identification system by proposed method. Future work will be concentrating on investigation of the effectiveness of feature extraction techniques for more robust speaker recognition. Investigation on a better adaptation function also will be done to ensure that the hybrid classifier get the better accuracy.

REFERENCES

[1] CAMPBELL, Joseph P.,Jr." *Speaker Recognition: A Tutorial"*.Proceedings of IEEE, vol. 85,no. 9, pp. 1437-1462, September 1997.

[2] D. Reynolds, "An overview of automatic speaker recognition technology," *in Proc. Int. Conf. on Acoust. Speech and Signal Process (ICASSP 2002)*, Orlando, FL, 2002, pp. 4072-4075.

[3] F. Bimbot, J.-F. Bonastre, G. Gravier, I. Chagnolleau, S. Meignier, T.Merlin, J. Garcia, D. Delacrètaz, and D. Reynolds, "A tutorial on text independent speaker verification," *Eurasip Journal on Applied SignalProcess.*, vol. 4, pp. 430-451, 2004.

[4] Vlasta Radová and Zdenek Svenda, "Speaker Identification Based on Vector Quantization", Proceedings of the Second International Workshop on Text, Speech and Dialogue, Vol. 1692,1999, Pages: 341 -344.

[5] 'TEXT INDEPENDENT AUTOMATIC SPEAKER RECOGNITION' *Othman O. Khalifa, S. Khan, Md. Rafiqul Islam, M. Faizal and D. Dol* Electrical and Computer Engineering, International Islamic University Malaysia, *3rd International Conference on Electrical & Computer Engineering ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh* ISBN 984-32-1804-4 561.

[6] Reynolds, D. A. and Rose, R. C. "Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans.Speech Audio Process. 3, 1995, pp 72–83.

[7] "Text independent speaker verification using GMM" Charles B.de Lima,Abraham Alcaim and Jose A.Apoloniyaro Jr.

[8] "SPEAKER IDENTIFICATION USING MEL FREQUENCY CEPSTRAL COEFFICIENTS "565 *Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman, 3rd International Conference on Electrical & Computer Engineering ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh* ISBN 984-32-1804-4

[9] "GMM BASED SPEAKER IDENTIFICATION USING TRAINING-TIME-DEPENDENT NUMBER OF MIXTURES"- 1998 IEEE Chakib Tadjt , Pierre Dumouchelt~ and Pierre Ouellett tEcole de Technologie Supkrieure - Electrical Engineering, 1100 rue Notre Dame Quest Montr6ai (Qc) - H3C 1K3 - Canada

[10] "Vector Quantization Decision Function for Gaussian Mixture Model Based Speaker Identification".2008 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS2008) Swissôtel Le Concorde,Bangkok,Thailand .