Speaker Recognition System using Gaussian Mixture Model

Athira Aroon
Department of Electronics Engineering
A.I.S.S.M.S Institute of Information Technology
Pune, India

S.B. Dhonde
Assistant Professor, Department Of Electronics Engineering
A.I.S.S.M.S Institute of Information Technology
Pune,India

ABSTRACT
In this paper,features for text-independent speaker recognition has been evaluated. Speaker identification from a set of templates and analyzing speaker recognition rate by extracting several key features like Mel Frequency Cepstral Coefficients [MFCC] from the speech signals of those persons by using the process of feature extraction using MATLAB2013 .These features are effectively captured using feature matching technique like Gaussian Mixture Model [GMM] , with varying mixture components of mixture model and the analyzing its effect on recognition rate . Improve the speaker recognition rate by varying the input parameters of the classifier. The experiments are evaluated on TIMIT Database effectively for a speech signal sampled at 16kHz.

Keywords
Gaussian Mixture Model [GMM] , Mel Frequency Cepstral Coefficients [MFCC], Speaker Recognition rate.

1. INTRODUCTION
Speaker Identification and Speaker Verification are the subparts of speaker recognition . Determining the uniqueness of the person from produced speech among population of persons is known as speaker identification. The process to reject or accept the identity claim of the speaker is called speaker verification. Recognizing individuality from their respective audio is called speaker recognition. The use of a particular accent ,pronunciation pattern, intonation style, vocabulary are some of the individual characteristic of each speaker for its speech[1] .

Forensics is one of the most important application of speaker recognition . Informations are exchanged among two individual in telephone conversations and in recent years there has been mounting concern to incorporate automatic speaker recognition to supplement aural and semi-automatic analysis methods[1][2].

The Speaker Recognition System comprises of stages like the training stage and testing stage.

Speaker Recognition system can be represented by Figure 1 The training stage proceeds over on with feature extraction technique of individual speech and then characterizing the speech by speaker modeling in training stage by the method of familiarizing the system with the voice individuality of a speaker, whereas testing is the definite recognition task[2] [3].

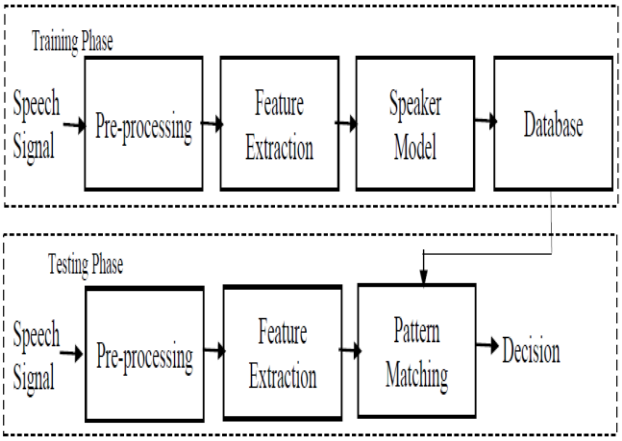


Fig 1. The Speaker Identification Process

2. FEATURE EXTRACTION
These features can be obtained from the spectrogram of the speech signal and we are using Mel-Frequency Cepstral Coefficients (MFCC) features in speaker identification, the advantages of perceptual frequency scale based critical bands with cepstrum analysis are combined[4].

On basis of literature survey MFCC is most accurate, popular and perhaps the best unknown. The Mel frequency scale is logarithmic spacing above 1000Hz and linear frequency spacing below 1000Hz.In order to capture the phonetically important characteristics of speech frequency filters are spaced logarithmically at high frequencies and linearly at low frequencies in accordance to properties of human ear. Thus the human ear perception is clearly mimicked by MFCC. This shortly describes is the process of feature extraction. Normally the speech signal is non-stationary but can be assumed as stationary for Although the speech signal is non-stationary, but can be assumed as stationary for a small tenure of time, so analysis is done by framing the speech signal; the frame width is about 20−30 milliseconds, and the frames are shifted by about 10 milliseconds[4][5].

MFCC Process includes the steps hierarchically. Framing is initially applied to the speech signal of the speaker partitioning the signal into N frames (segments) . Inorder to reduce the signal discontinuities at the start and end of each segment, the next step that is windowing is undertaken. Later the windowed frames are processed by Fast Fourier Transform (FFT) converting frames of N samples in time domain to frequency domain. Obtained spectrum is later wrapped and converting the frequency spectrum to Mel spectrum .And finally the log Mel spectrum is converted back to time resulting in Mel Frequency Cepstrum Coefficients

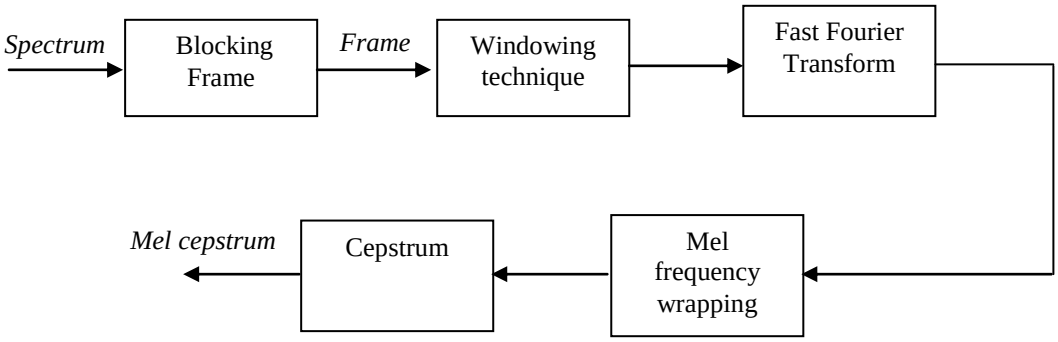


Fig 2. The Process of MFCC

(MFCC). Mel scale follows the relation for an arrangement of frequency range to mel scale.

$$f_{mel} = 2595 \log(1 + \frac{f_{hz}}{700})$$

Frame cepstrum is achieved by logarithm of amplitude of mel spectrum and applying reverse Fourier conversion:

mel- cepstrum (frame) $= FFT^{-1}$ *[mel (log| FFT(frame)|)]*

By taking the IFFT of the log magnitude spectrum of speech signal the FFT-base cepstral coefficients are computed[4]. The mel-warped cepstrum is obtained by inserting a intermediary step of transforming the frequency scale to place less prominence on higher frequencies before taking the IFFT[5].

3. SPEAKER MODELING

In scientific field and engineering the need of speaker recognition is a much broader topic so called pattern matching. Classifying the objects into number of classes and categories is the actual target of pattern matching .By using the techniques ,from sequences of acoustic vectors the patterns that are basically objects of interest are classified. An unknown speaker is the one with minimum matching score. For classification and for speaker recognition includes various speaker modeling techniques like Gaussian Mixture Modeling (GMM)[5].

Gaussians represent over the underlying speaker discrimination information. The formant bandwidths, magnitudes and location of speech signal are estimated by GMM parameters that is co-variances, means and component weight. Gaussian probability P1 is calculated using Gaussian probability density function by passing the feature vectors X= (X1,X2,X3….X12). The new feature vector X is passed through the Gaussians Resulting in the first coefficient P1[2][5].

4. GAUSSIAN MIXTURE MODEL

GMM statistical speaker model is created after extracting features .Conditions when single normal distribution fail at such moments finite mixture models and their typical parameter estimation methods can be approximated by a wide variety of probability density functions(pdf). A basic distribution to be followed that is by using predefined distribution type used to form a mixture

[3][6].

The multivariate normal distribution that is the Gaussian distribution is undoubtedly one of the most useful and well-known distribution in playing predominant role , in statistics and in other areas of applications[3][7].

In text independent speaker recognition applications we have inculcated Gaussian classifier..

$$P(x|\lambda) = \sum_{i=1}^{M} p_i \, b_i \, (x)$$

M defines the number of component densities, the mixture weights for i = 1, .., M. x is a D dimensional observed data are the component densities[7] .

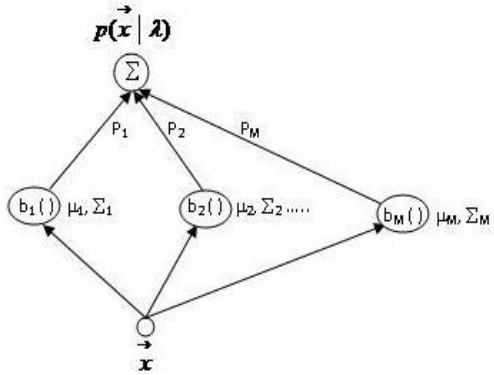


Fig 3. Gaussian Mixture Model

Each component density denotes a D-dimensional normal distribution with mean vector μ and covariance matrix Σ These parameters can be collectively represented as

$$\lambda = \{\omega, \mu, \Sigma\}$$

Each speaker in a language system can be represented by a GMM and is referred by the language respective model λ. We have to calculate a good estimation of the GMM parameters to obtain an optimum model representing each speaker .

$$N(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

The most popular and well-established method is maximum likelihood (ML) estimation. Using the training data Maximum Likelihood estimation is used in order to find the model parameters by maximizing the likelihood of the GMM. We use to do that, using a very efficient method by Expectation Maximization algorithm approach. For a sequence of T training vectors {X1…Xt} The GMM likelihood is written as

$$P(x|\lambda) = \prod_{i=1}^{T} p(x_t|\lambda)$$

ML parameter estimates can be obtained iteratively using a special case of the expectation-maximization (EM) algorithm .

The basic idea of the EM algorithm is, beginning with an initial model, x, to estimate a new model. The new model then becomes the initial model for the next iteration and the process continues until some convergence threshold is reached [7][8].

5. EXPERIMENTS

5.1 Experimental Setup

TIMIT database consists of recordings of 630 (438 male and 192 female) speakers. Each speaker has recorded 10 sentences which makes a total of 6300 sentences. TIMIT database is recorded from 8 major dialect regions of United States. Each sentence is of 3seconds. Concatenation of 8 sentences to form training duration of 24 seconds and test duration of 3 seconds. Evaluation for varying mixture components is conducted for 100 speakers and then evaluate the Speaker Recognition rate for increasing number of speakers.

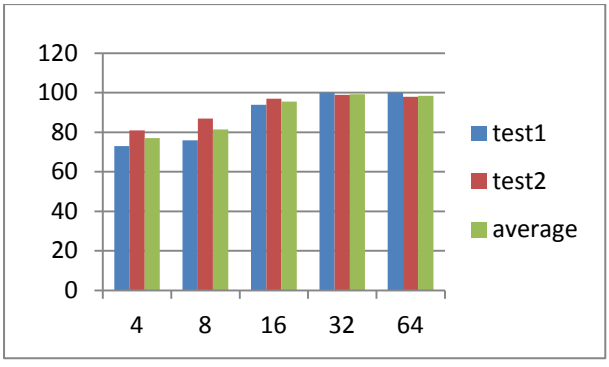
5.2 Experimental Results

We have evaluated speaker recognition rates for varying mixture components and its outperformance. The result observance by varying the input parameters for increasing the number of speakers. For Gaussian Mixture Model the improvement in recognition rate is observed by inserting K means algorithm with variance of 0.001 ,decrease in variance leads on to increase in recognition rate.

For 100 speakers with varying mixture components in TIMIT database the observed Average Speaker Recognition Rates are

Training Duration (24 sec) Test Duration (3 sec)	Varying Mixture Components	Speaker Recognition Rate (%)		
		Test1	Test2	Average
	4	73	81	77
	8	76	87	81.5
	16	94	97	95.5
	32	100	99	99.2
	64	100	98	98.5

It has been observed that for 32 mixture components we get the improved recognition rate.



Speaker Recognition rate for 100 speakers for varying mixture components

It can be further elaborated with define range proximity of closeness and in their respective values like mean , variance and weight. For lower mixture components like 4,18,16 the difference for successive values is more as comparative for 32, 64. For 64 mixture components time required is more and there is overlapping of values with increased number of speakers.

6. CONCLUSION

Tested for varying mixture components as 4, 8, 16, 32 and 64 tested for 24seconds training duration and 3 seconds testing duration and observed that with 32 mixture components test gives good speaker recognition performance.

By inserting K means algorithm for 32 mixture components the improvement in recognition rate has been observed. The improvement in recognition rate is observed by inserting K means algorithm with variance of 0.001 ,decrease in variance leads on to increase in recognition rate. The recognition performance degrades with increase in speakers.

In further analysis the average recognition rate needs to be improved and analyze the random variations for the test speech without any synchronization between the both set of test groups for varying mixture components like for 4,8 and 16 ,set test 2 has better recognition rate wherein for 32 and 64 test 1 has improved recognition rate as compared to test 2.

7. REFERENCES

[1] Campbell J.P. and Jr. "Speaker recognition: A Tutorial*" Proceeding of the IEEE. Vol 85, 1437- 1462 1997.*

[2] S.Furui. "Fifty years of progress in speech and speaker recognition," *Proceedings ASA Meeting, 2004.*

[3] Tomi Kinnunen., and Haizhou Li., An overview of Text-Independent Speaker Recognition: from Features to Supervectors. *Speech Communication, July 1, 2009.*

[4] Kinnunen, T., Hautamaki, V., and Franti, P. On the use of long-term average spectrum in automatic speaker recognition. In 5th Int. Symposium on Chinese Spoken Language Processing (December 2006), pp.559-567.

[5] Yuan Yujin, Zhao Peihua, Zhou Qun,, "Research of speaker recognition based on combination of LPCC and MFCC", IEEE International Conference , Oct. 2010, pp.765-767.

[6] D. A. Reynolds, A Gaussian mixture modeling approach to text independent speaker identification, Ph.D. thesis, Georgia Institute of Technology, Atlanta, Ga, USA, September 1992.

[7] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Communication, vol. 17, no. 1-2, pp. 91–108, 1995."

[8] Reynolds, D. "Speaker Verification Using Adapted Gaussian Mixture Models." Digital Signal Processing 10.13 (2000): 19-41. Print.

[9] Lu, X. and J. Dang (2008). An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. Speech Communication, 50(4), 312–322.

[10] Sinith, M.S., Salim, A., Gowri Sankar, K., Sandeep Narayanan, K.V. Soman, V., "A novel method for Text-Independent speaker identification using MFCC and GMM", , 2010 International Conference, Nov. 2010, pp.292-296

[11] Adami.A, Mihaescu.R, Reynolds.D, and Godfrey.J., Modelling Prosodic dynamics for speaker recognition. *In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing* , April 2003), pp. 788-791.