Text - Dependent Speaker Recognition

Kalyan Bhattacharjee[1] Dr. Aditya Bihar Kandali[2]
[1]Student [2]Associate Professor
[1,2]Department of Electrical Engineering
[1,2]Jorhat Engineering College, Jorhat, Assam, India

Abstract— Speaker Recognition system has been an active area of research for several decades, as it finds application in many access control systems, recognizing a person from a telephonic conversation, etc. The goal is to automatically identify or recognize a particular person without the need of a human as an identifier. This paper presents a Text-Dependent Speaker Recognition system based on Gaussian Mixture Model (GMM) classifier which models speaker identity. The system is evaluated on Sylheti Speaker Database prepared from 100 Sylheti speaking people of various age groups.

Key words: GMM Classifier, MFCC, LPC, PLP, Wavelet Packet Transform.

I. INTRODUCTION

Speech, the most essential ingredient of communication among human beings, not only contains the information to be conveyed, but also the identity of the speaker. This property of speech is utilized in developing a system where a speaker is automatically identified from his/her speech, the task being known as the Speaker Recognition task. While the task of Speech Recognition is concerned with extracting all the linguistic information present in a speech sample, the task of Speaker Recognition is concerned with recognizing a speaker from his/her speech sample.

Depending upon the application, there are two specific sub-tasks of Speaker Recognition which are the Speaker Verification and the Speaker Identification [1]. In the Speaker verification task, the objective is only to verify from a voice sample, whether or not the particular speaker is known to the system. On the other hand, speaker's identity is obtained in the Speaker Identification task by determining which one of a group of known voices best matches the input voice sample. Speaker Recognition task can also be sub-grouped depending on the text to be spoken: text-dependent speaker recognition and text-independent speaker recognition. In the text-dependent case, the speaker has to speak the same predefined text both for training and testing the system [2], [3]. But there is no such bound on the text to be spoken in the text-independent case, in which the recognition task is solved based on factors such as the shape and size of the vocal tract, dynamics of the articulators, rate of vibration of the vocal folds, accent imposed by the speaker and speaking rate [4],[5]. The most popular and efficient classifier used for the purpose of speaker recognition is GMM [4], [6].

In this paper, a work on GMM based text-dependent Speaker Recognition is discussed for four types of features which are Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coefficients (LPC), Perceptual Linear Analysis (PLP) and Wavelet Packet Transform (WPT) based features. The speech corpus was prepared in Sylheti Language, which is considered a dialect of the Bengali language by the government of Bangladesh [7], and is primarily spoken in the Sylhet Division of Bangladesh and the Barak Valley region of southern Assam, India. Also, due to many similarities between Sylheti and Assamese languages, Sylheti is often considered as an Assamese dialect and also often considered as a separate language due to significant differences between them all and lack of mutual intelligibility.

The rest of the paper is organized as follows. Section II discusses about the way of speech database preparation followed by a brief introduction to GMM in section III. In section IV we discuss about the features extracted from the speech signal. Speaker Recognition system is presented in section V and we discuss the results obtained in section VI. Finally, some concluding remarks are discussed in section VII.

II. DATA COLLECTION

The speech samples were collected in Sylheti language from 100 speakers whose first language was Sylheti. First a set of 20 sentences along with two short paragraphs were prepared in typical Sylheti language that was phonetically balanced. Then, each of the 100 speakers were made to read out the prescribed texts one by one clearly in an almost noise-free environment and the recording was done on a cell phone in WAV format, sampling rate and bit-depth being 16kHz and 8 bits respectively. Out of the 100 speakers, 63 were male speakers and remaining 37 were female speakers and all were from various age groups.

III. GAUSSIAN MIXTURE MODEL

A particular speech from a speaker is never spoken with exactly the same vocal tract shape and glottal flow due to the context, co articulation, anatomical and fluid dynamical variation [1]. This variability in speech production is represented in probabilistic way through a multi-dimensional Gaussian probability density function (pdf) [6]. For a 'D' dimensional feature vector 'x' of a particular speaker 'S', the Gaussian pdf for ith state denoted as $b_i^S(x)$ is given as [6]:

$$b_i^S(x) = \frac{1}{(2\pi)^{D/2}|\Sigma_i^S|^{1/2}}\, exp\left\{\frac{-1}{2}\left(x-\mu_i^S\right)^T\left(\Sigma_i^S\right)^{-1}\left(x-\mu_i^S\right)\right\} \quad (1)$$

Where, $\mu_i^S$is the mean vector and $\sum_i^S$ is the covariance matrix. The covariance matrix represents the cross-correlations (off-diagonal terms) and the variance (diagonal terms) of the elements of the feature vector.

The probability of a feature vector 'x' being in any of 'M' states (or acoustic classes), for a particular speaker model, denoted by λ_S, is represented by the union, or mixture, of different Gaussian pdfs, thus giving us the Gaussian Mixture Model (GMM) of a speaker 'S' as:

$$p(x|\lambda_S) = \sum_{i=1}^{M} p_i^S b_i^S(x) \quad (2)$$

Where, $b_i^S(x)$ are the component mixture densities given in (3.1) and p_i^S are the mixture weights satisfying the constraint$\sum_{i=1}^{M} p_i^S = 1$, to ensure that the mixture density represents a true pdf. The speaker model thus represents a set of three parameters which are the GMM weights, means and the covariance matrix:

$$= p_i^S, \mu_i^S, \Sigma_i^S \quad (3)$$

These parameters are estimated by maximum-likelihood estimation i.e., maximizing with respect to λ_S, the conditional probability$p(x|\lambda_S)$. Maximum-likelihood model parameters are estimated using the iterative Expectation-Maximization (EM) algorithm [1], [4].

IV. Feature Extraction

In this section, we discuss the features extracted from the speech signal. But prior to feature extraction, some preprocessing has been done on the speech signal. At first, the silence periods were removed from the speech samples. Then the silence removed speech signals were passed through a pre-emphasis filter (high pass filter) having transfer function $\mathrm{H(z)} = 1 - \alpha \mathrm{z}^{-1}$. This was necessary to spectrally flatten the signal and to make it less sensitive to finite precision effects later in the signal processing

A. *Mel Frequency Cepstral Coefficients (MFCC)*

The method of MFCC computation is shown in the block diagram of Fig. 1 [1]. For each speaker, 12 MFCC, 12delta MFCC and 12delta-delta MFCC were extracted from each speech frame using 24 triangular shaped filter banks.

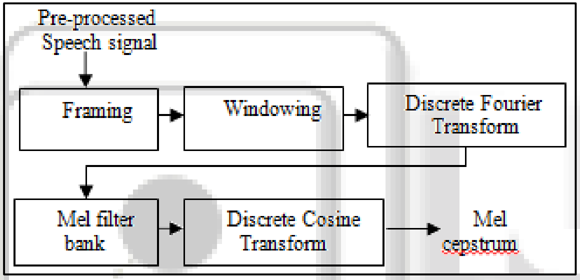


Fig. 1: Block diagram showing MFCC extraction method.

B. *Linear Prediction Coefficients (LPC)*

The idea behind linear predictive analysis [8] is that each speech sample at a given time instant is approximated as a linear combination of past speech samples. Linear prediction is used to obtain the filter coefficients (corresponding to the vocal tract) by minimizing the mean square error between the input and the estimated sample. Block diagram in fig. 2 shows the procedure for obtaining the filter coefficients using autocorrelation method. In this work, 12 LPC features were computed.

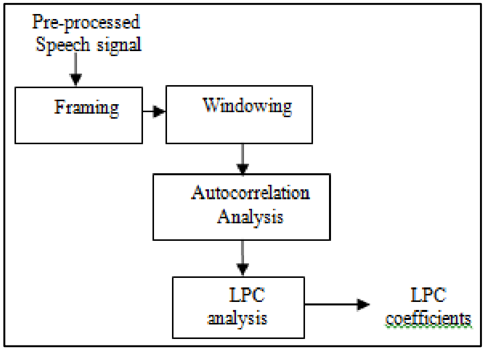


Fig. 2: Block diagram showing LPC extraction method.

C. *Perceptual Linear Prediction (PLP)*

The Linear Predictive analysis can be viewed as a means for obtaining the smoothed spectral envelope of the short term power spectrum of speech, but with an associated disadvantage that it approximates the power spectrum equally well at all frequencies which is not the case with human hearing. PLP, proposed by Hermasky [9] is one technique used for rectifying this problem. In this work, 12 PLP coefficients were extracted.

D. *Wavelet Packet Transform (WPT)*

In addition to the above three types of features, Wavelet Packet Transform based features were extracted. The Wavelet Packet transform may be thought of as a tree of subspaces with $\Omega_{0,0}$ representing the original signal space, i.e., the root node of the tree (fig. 3) [10]. Then appropriate features are extracted from each of the decomposed signals. In this work, 7-level decomposition was done.

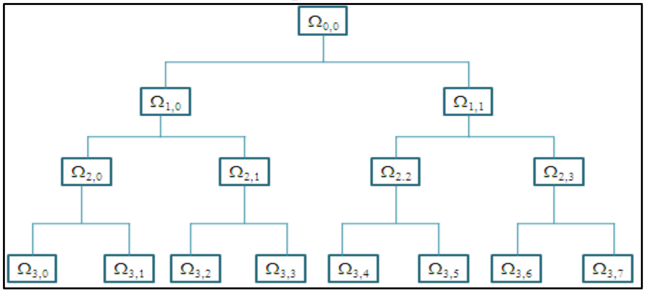


Fig. 3: An example of the wavelet packet decompositions of Ω0,0 into tree structured subspaces

V. Recognition System

The recognition system is based on the Gaussian Mixture Model classifier. Two phases are involved in the recognition system, namely, the training phase and the testing phase (Fig. 4 and Fig. 5).

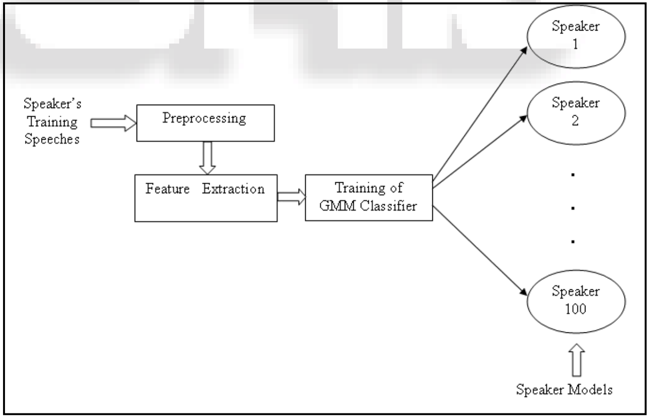


Fig. 4: Training blocks diagram representation.

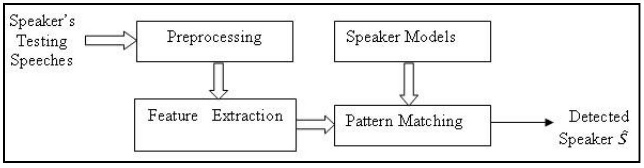


Fig. 5: Testing block diagram representation.

In the training phase, a speaker model for each speaker was built from the extracted features by estimating the GMM parameters which are the GMM mean (μ), covariance (Σ) and weight (p) parameters represented by the set $\lambda_S = \mathrm{p}, \mu, \Sigma$. Out of the 22 speech recordings collected per speaker, features for 11 of them were used for training the GMM and hence for building the speaker models. Expectation-maximization algorithm was used for estimating the parameters.

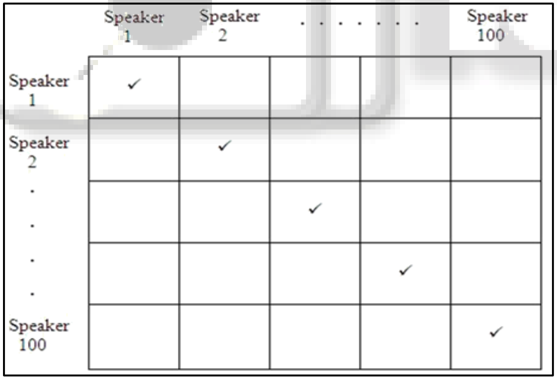
Testing phase is the actual recognition phase where decision is taken regarding to which class (speaker in this case) a particular test utterance (speech) belongs. Features from the other 11 speeches for a particular speaker which were not used for the training phase are used here as the testing features. The maximum a posterior probability (MAP) classification technique was used for taking the decision regarding which speaker model is a best fit for the particular test feature. In MAP classification, we compute the probability of each speaker model given the features i.e., $\mathrm{p}(\lambda_{\mathrm{j}}|\mathrm{x}_{\mathrm{n}})$, and then choose the speaker with the highest probability. Here x_{n} is a stream of feature vectors i.e., $\mathrm{x}_{\mathrm{n}} = \mathrm{x}_0, \mathrm{x}_1, \ldots.., \mathrm{x}_{\mathrm{M}-1}$, where M is the number of feature vectors for the test utterance. To express $\mathrm{p}(\lambda_{\mathrm{j}}|\mathrm{x}_{\mathrm{n}})$ in terms of GMM, by Bayes' rule we write,

$$\mathrm{p}(\lambda_{\mathrm{j}}|\mathrm{x}_{\mathrm{n}}) = \frac{\mathrm{p}(\mathrm{x}_{\mathrm{n}}|\lambda_{\mathrm{j}})\mathrm{p}(\lambda_{\mathrm{j}})}{\mathrm{p}(\mathrm{x}_{\mathrm{n}})} \quad (1)$$

As $\mathrm{p}(\mathrm{x}_{\mathrm{n}})$ is constant, we maximize the quantity $\mathrm{p}(\mathrm{x}_{\mathrm{n}}|\lambda_{\mathrm{j}})\mathrm{p}(\lambda_{\mathrm{j}})$ where $\mathrm{p}(\lambda_{\mathrm{j}})$ is the a priori probability of speaker λ_{j}. The prior probabilities $\mathrm{p}(\lambda_{\mathrm{j}})$ are assumed equal and the problem is to find the $\mathrm{p}(\lambda_{\mathrm{j}})$ that maximizes $\mathrm{p}(\mathrm{x}_{\mathrm{n}}|\lambda_{\mathrm{j}})\mathrm{p}(\lambda_{\mathrm{j}})$. By applying the logarithm, we can write the speaker recognition solution as:

$$\hat{\mathrm{S}} = \max_{1\leq \mathrm{j} \leq \mathrm{S}} \sum_{\mathrm{m}=0}^{\mathrm{M}-1} \log[\mathrm{p}(\mathrm{x}_{\mathrm{m}}|\lambda_{\mathrm{j}})] \quad (2)$$

A confusion matrix was prepared as shown in Fig. 6, to store the results obtained. Three such confusion matrices were prepared, one for storing the results obtained for male speaker, other for female speaker and the third one for all male and female speakers. Only the diagonal entries of the matrix represent success in percentage.



	Speaker 1	Speaker 2	. . . .		Speaker 100
Speaker 1	✓				
Speaker 2		✓			
. .			✓		
. .				✓	
Speaker 100					✓

Fig. 6: A confusion matrix whose diagonal entries represent success.

VI. Results and discussions

MATLAB v2013a software is used for all computations.

A comparative analysis of the performance of GMM based text dependent speaker recognition for the four types of features i.e., MFCC, LPC, PLP and Wavelet packet transform based features have been done and is illustrated in Fig. 7. The figure illustrates, in four subplots, the average percentage accuracy obtained for male, female and all speakers, one by one for 4, 8, 12 and 16 Gaussian Mixtures respectively. Average percentage accuracy is obtained by taking the mean of the diagonal entries of the respective confusion matrices. Observing the figure, first of all it can be said that with the increase in the number of Gaussian probability density functions (pdf), the system performance also increases for all the features.

MFCC along with delta and delta-delta energy provide better result every time than the rest three. MFCC bear resemblance to the human auditory system as it incorporates a mel-scale filter bank, whose center frequencies and bandwidths roughly match those of the actual auditory critical band filters, and hence MFCC has proved to be one of the most successful feature in Speaker Recognition task.

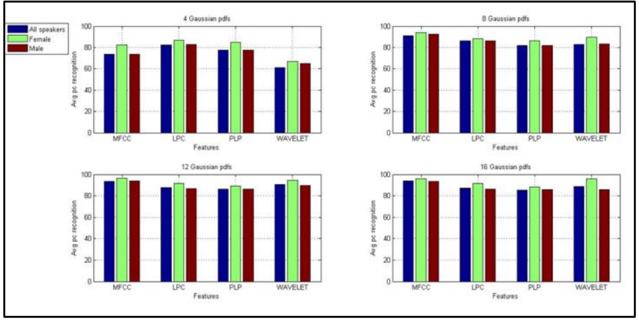


Fig. 7: Results of the speaker recognition system for four types of features.

VII. Conclusion

This paper has introduced and evaluated the Gaussian Mixture Model for Text-Dependent Speaker Recognition using four types of features on an almost noise-free database. Any algorithm employed for Speaker Recognition, be it GMM or other any other, relies on correct feature extraction as well as the conditions in which the speech recordings are taken. In this work, MFCC has provided a better result than the other three types of features.

A further research work can be carried out in this domain considering a noisy database collected from a chaotic environment like public places, market places, etc. Investigations can then be done on whether the same algorithms and the same features can achieve the same performance as obtained on a clean database.

References

[1] T. F. Quatieri, "Discrete Time Speech Signal Processing: Principles and Practice", Pentice Hall Signal Processing Series.

[2] W. Chen, Q. Hong and X. Li, "GMM-UBM for Text-Dependent Speaker Recognition", IEEE-2012 International onference on Audio, Language and Image Processing, 2012.

[3] D. D. Thi Thu, L. T. Van, Q. N. Hong and H. P. Ngoc, "Text-Dependent Speaker Recognition for Vietnamese", IEEE-2013 International Conference on Soft Computing and Pattern Recognition, pages 196-200, 2013.

[4] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models", IEEE Transactions on Speech and Audio Processing, vol. 3, January 1995.

[5] B. Xiang and T. Berger, "Efficient Text-independent Speaker Verification with Structural Gaussian Mixture Models and Neural Network", IEEE Transactions on Speech and Audio Processing, vol. 11, September 2003.

[6] D. A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models", Speech Communication, vol. 17, pages 91-108, 1995.

[7] Sylheti Language, https://en.wikipedia.org/wiki/Sylheti_language

[8] L. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition", Pentice Hall International.

[9] H. Hermansky, "Perceptual Linear Prediction (PLP) analysis of Speech", Journal of The Accoustic Society of America, pages 1738-1752, 1990.

[10] S. Mallat, "A Wavelet Tour of Signal Processing", Academic Press, 2e.