A Novel, Robust, Hierarchical, Text-Independent Speaker Recognition Technique

Prateek Srivastava *prateek.k.srivastava@gmail.com*
National Institute of Technology
Rourkela, India, 769008

Reena Panda *reena.panda@gmail.com*
National Institute of Technology
Rourkela, India, 769008

SankarsanRauta sankarsan.1946*@gmail.com*
National Institute of Technology
Rourkela, India, 769008

Abstract

Automatic speaker recognition system is used to recognize an unknown speaker among several reference speakers by making use of speaker-specific information from their speech. In this paper, we introduce a novel, hierarchical, text-independent speaker recognition. Our baseline speaker recognition system accuracy, built using statistical modeling techniques, gives an accuracy of 81% on the standard MIT database. We then propose and implement a novel state-space pruning technique by performing gender recognition before speaker recognition so as to improve the accuracy/timeliness of our baseline speaker recognition system. Based on the experiments conducted on the MIT database, we demonstrate that our proposed system improves the accuracy over the baseline system by approximately 2%, while reducing the computational time by more than 30%.

Keywords: Speaker Recognition, Gender classification, Mel Frequency Cepstral Coefficients, Cepstral Mean Subtraction, Gaussian Mixture Model.

1. INTRODUCTION

Speaker recognition is the task of automatically recognizing/identifying an unknown speaker among several reference speakers using speaker-specific information included in speech waves [10]. Such a system can have several potential applications such as a biometric tool for security purposes. Speech being one of the most natural and common form of communication, any speech-based security system would be non-intrusive and havehigher user acceptance. Also such systems can be easily integrated into the ubiquitous telephone network, thereby providing access controlfor banking transactions by telephone, automatictelephonetransactions such as voice mail and credit card verification, and remote access to computers via modems on dial-up telephone lines. Such a system can also have potential applications in forensics.

Speaker recognition [5, 21, 22] combines both speaker verification and speaker identification. Speaker verification is the technique to verify a person's claimed identity by making use of the speech cures. On the other hand, in speaker identification, no identity claims are made and the system has to identify the speaker. Significant work has been done in the area of speaker recognition over the past years. The most notable and widely referred approaches are:- the Gaussian mixture model (GMM - UBM) [19], and the mixed GMM- UBM and SVM technique [23]. Speaker recognition systems can be further divided into text-dependent and text-independent systems. In text-dependent systems [24], the recognition phrases/words are constant or known a

priori. On the other hand, in text-independent systems, there are no constraints on the words which the speakers are allowed to use and thus, text independent recognition is considered to be a more challenging task.

In this paper, we propose a text-independent speaker recognition system based on Gaussian Mixture Models (GMMs) which is proven to be a powerful tool and is often employed in text-independent classification tasks. We also propose a technique for speeding up and improving the accuracy of the speaker identification task by pruning the search space by dropping out the unlikely speakers by making use of gender recognition before speaker identification.

The rest of the paper is organized as follows. In Section 2, we describe our speaker recognition technique. Section 3 discusses the improvements that we propose to our speaker recognition system. Section 4 provides a description of the experiments along with a detailed analysis of the results. We finally conclude the paper in Section 5 with notes regarding the future work.

2. GMM-Based Speaker Recognition

The recognition system is divided into two phases namely training phase and testing phase. In training phase, speech samples are collected pre-processed and then speaker-specific features are extracted from them. Thereafter, the different speaker classes are statistically modeled using GMMs. In the testing phase, features are extracted from the test samples and their likelihood of match is estimated against the trained models. The model against which the test sample yields the highest likelihood score is identified as the speaker class. This is shown in Figure 1.

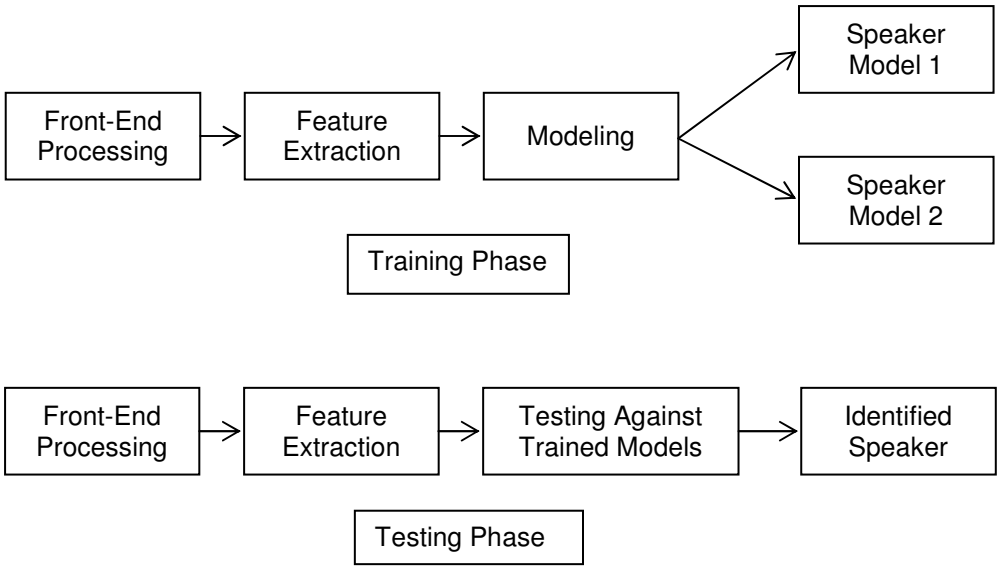

FIGURE 1: Pictorial Representation of the Speaker Recognition System.

In the following subsection, we describe the techniques that we use for front-end processing, feature extraction and feature matching respectively.

2.1 Front end Processing

2.1.1 Pre-emphasis

The sampled speech is pre-emphasized to enhance the high frequency components of the spectrum, especially the so-called formants, against the lower frequencies which contain most of the signal's power, but are known to be rather irrelevant for speech intelligibility. Pre-emphasis of

the high frequencies is done to obtain similar amplitudes for all the formants [8]. This is performed by applying a first order FIR filter to the speech signal:

$$s[k] = s[k] - a_1 . s[k-1] \text{ where } a_1 = 0.97$$

2.1.2 Framing

The resulting pre-emphasized speech signal is then divided into smaller parts out of which certain features essential for recognition are extracted. These short-time intervals of the speech signal are called frames. Since the frame duration is very small, each frame is assumed to be a stationary process and is assumed to have a constant spectrum. Overlapping of the frames is done so that the adjoining frames would overlap to achieve a smoother development of the short-time characteristics of the individual signal blocks [10]. Overlapping is done mainly to avoid loss of information.

2.1.3 Windowing

All these frames are then multiplied by a window function. This is required to smooth the edges of each frame to reduce the discontinuities or abrupt changes at the endpoints. Windowing also serves to reduce the spectral distortion that arises from the windowing itself [10]. Here, in our experiments, we have made use of a hamming window, which is characterized by:

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

where, N = width in samples and n is an integer with values $0 < n < N-1$

2.2 Feature Extraction and Modeling

The acoustic signal contains different kinds of information about the speaker. The signal processing involved changes depending on the type of characteristics we are interested in the speaker. The basic aim of feature extraction in our recognition system is to reduce the amount of data while retaining the speaker-dependent and gender-specific information.

2.2.1 Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs have by far, proved to be the most successful and robust feature for recognition purposes. The MFCC feature set is based on the human perception of sound i.e., on the known evidence that the information carried by low-frequency components of the speech signal are phonetically more important for humans than the high-frequency components [9]. This is expressed in the mel-frequency scale, which is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz .The MFCC feature extraction algorithm [3, 4, 11] is shown in Figure 2.

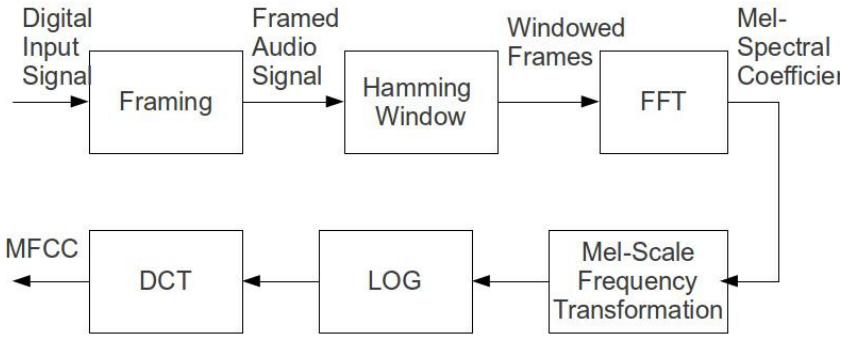

FIGURE 2:MFCC Feature Extraction Process.

The final MFCC feature vector is composed of 39 parameters (including the delta and delta- delta coefficients which are added to model the inter-frame dependencies in speech and are the time derivatives of the basic static parameters). However these delta and delta-delta coefficients can increase the feature vector by up to 24 dimensions. So, in this paper we have used the Delta Cepstral energy (DCE) and Delta-Delta Cepstral Energy (DDCE) that can compactly represent the delta and delta-delta cepstral information in one-dimensional feature [12]. For any one frame, they are calculated as follows:-

$$\text{DCE} = \sum_{l=1}^{L} (\Delta MFCC)^2$$

$$\text{DDCE} = \sum_{l=1}^{L} (\Delta^2 MFCC)^2$$

where, $\Delta MFCC_l$, $\Delta^2 MFCC_l$ are the l^{th}delta and delta-delta cepstral coefficients and L is the number of MFCCs.

2.2.2 Maximum Auto-Correlation Value (MACV)

The pitch frequency is an extremely important property of speech and defines the periodicity of a speech signal. However the accurate pitch extraction is not an easy task due to the non-stationarity and quasi-periodicity of speech signal, as well as the interaction between the glottal excitation and the vocal tract. Also speech frames are not always periodic and pitch cannot be determined for the unvoiced frames. So, here we have used the Maximum Auto-correlation algorithm [10] (MACV) which does not use pitch value directly as a feature and works well for both voiced and unvoiced frames. It captures the periodicity characteristics of speech signal in an indirect manner in the form of voicing information.

2.2.3 Cepstral Mean Subtraction (CMS)

Practically, the speech samples in the database are collected using different microphones, each having its own inbuilt channel noise. This channel noise gets convolved with the environmental noise. To remove the variability in different speech samples owing to the use different microphones, we make use of the Cepstral mean features (CMS). After the features are extracted from each speech sample, the mean of the whole feature set is calculated and is subtracted from each frame to get the Cepstral mean features. It is assumed throughout that the speech signal has a zero mean and the channel noise is finite. It has been established experimentally in prior research work that CMS yields more robust features than MFCC by itself.

2.3 Gaussian Mixture Modeling

Gaussian mixture models (GMMs) [19, 20] are parametric representation of a probability density function. When trained to represent the distribution of a feature vector, GMMs can be used as classifiers. GMMs have proved to be a powerful tool for distinguishing acoustic sources with different general properties. The use of GMMs for modeling activity is motivated by the interpretation that the (1) uni-variate Gaussian densities have a simple and concise representation, depending uniquely on two parameters, mean and variance, (2) they are capable to model arbitrary densities, (3) the Gaussian mixture distribution is universally studied and its behaviors are widely known, (4) a linear combination of Gaussian basis functions is capable of modeling a large class of sample distributions. In principle, the GMM can approximate any probability density function to an arbitrary accuracy.

A GMM is a weighted sum of M component densities as shown in figure, given by the equation:-

$$P(x_t / \lambda_s) = \sum_{i=1}^{M} p_i b_i \quad (x(t))$$

Here, x_t is a sequence of feature vectors from the activity data, x(t) is feature vector having D-dimensionality. $b_i(s)$ is the Gaussian probability distribution function (PDF) associated with the i^{th} mixture component and is given by:

$$b_i(x_t)=\frac{1}{2\pi^{D/2}|\sum_s^i|^{1/2}} e^{\left(-\frac{1}{2}\right)(x-\mu i)\hat{}T \sum_i^s \quad ^{-1}(x-\mu i)}$$

Here, $\mu_i$is the mean vector and $\sum^s_i$ is the covariance matrix of the i^{th} mixture component.

The mixture weights are such that:-

$$\sum_{i=1}^{M} p_i \quad = 1$$

Each trained speaker is thus, represented by a Gaussian mixture model, collectively represented by:-

$$\lambda_s=\{\mu_i , \sum_i , p_i \}$$

where, i=1,2 ,…M, μ_i , $\sum_i$, p_i represent the mean, covariance and weights of the i^{th}mixture respectively.

In this paper, the models are trained using the expectation-maximization (EM) algorithm [18]. The basic idea of the EM algorithm is as follows:- Beginning with an initial model λ, to estimate a new model λ', such that p (X | λ') ≥ p(X | λ).The new model then becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached.

But during the implementation of the EM algorithm, a singularity problem arises which limits the training to a limited number of Gaussians. To avoid these problems, a variance flooring method is generally used. However in our experiments, we find that the variance flooring method is also not able to solve the singularity problem altogether. In our experiments, we found out that if, an optimum splitting of mean is implemented during the EM implementation, then it deals with the singularity problem completely and we are able to train the data to any number of Gaussians. So, in our experiments, we have proposed and implemented this optimum splitting of mean technique so as to overcome the singularity problem.

3 Proposed Improvements to the Speaker Recognition System

In this paper, we propose a novel technique to improve both the accuracy and computational speed of the speaker identification task by pruning the state search space. We propose to do so by dropping out the more unlikely speakers from the search space by preceding the speaker identification stage with a gender recognition stage. The basic idea of the two approaches we have followed for Speaker Identification purposes is depicted below in Figure 3, where the top figure corresponds to the original speaker identification system and the bottom figure demonstrates our proposed changes.

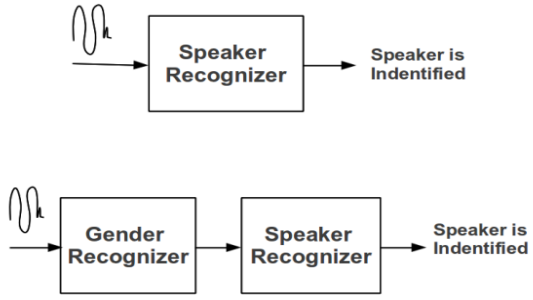

FIGURE 3: Speaker recognition system and Hierarchical Speaker Recognition System.

From our speaker recognition experiments, we observed that some of the incorrect recognition cases resulted from confusions with speakers of a different gender. In a separate set of experiments, where we performed gender recognition using speech features, we achieved significantly higher recognition accuracies. So, we tried to improve our speaker recognition system by implementing a pruning stage before the actual speaker recognition stage where we first estimate the speaker's gender and then perform speaker recognition on the identified smaller

speaker set. Such a system has two main advantages :- a) First of all, if the implemented gender recognition system is highly accurate, then it would allow to reduce the inter-gender confusions thereby resulting in a higher overall recognition accuracy, b) A gender recognizer before the speaker recognizer prunes the state space and thus, the computational speed of the overall system improves. The results of this improved recognition system is provided in the next section.

The hierarchical recognition system improves the performance by reducing the inter gender misclassification. The hierarchical approach exploits the difference in statistical properties of male and female during 1st phase of recognition. This approach also provides us the flexibility to use more targeted feature for different gender cluster. In this paper, we have used different feature set for gender recognition phase and speaker recognition phase in hierarchical recognition system.

4 Experiments and Results

4.1 Dataset

In this paper, we have conducted the experiments on the MIT database and a self-collected database:-

- MIT Database
  - Was collected by a prototype hand held device in order to simulate scenarios encountered by real-world speech recognition and verification systems.
  - Used different locations as well as different microphones.
  - Total 48 speakers with 22 females and 26 males.
  - Sampled at 16k Hz.

- Self-Collected Database
  - To deal with real time noisy condition
  - Total 20 Indian speakers including 9 females and 11 males.
  - Sampled at 16k Hz.
  - Different microphone and collected in different sessions.

4.2 Experiments and Results

During the training phase, the speech signals from each speaker class were pre-emphasized using a first order FIR filter (pre-emphasis coefficient = 0.97). Then they were divided into 20 ms frames with an overlap of 10ms. Each frame was then, windowed using a hamming window. Features are then extracted from each windowed frame. In our experiments, we have used feature vectors composed of 12 lowest Mel-frequency cepstral coefficients computed using 21 Mel-spaced filters (the 0th coefficients being excluded because they carry little speaker-specific information), the delta and delta-delta coefficients, the delta and delta-delta cepstral energy, 5 MACV features derived from the auto-correlation function and the cepstral mean subtraction features (determined for each utterance). After extraction of features, the speaker classes were statistically modeled using GMMs. EM algorithm was then used for estimating the parameters of the GMM class. At the end of the training phase, we were thus, left with Gaussian mixture models, representing each speaker class. Experiments were conducted with 32 and 64 mixtures.

In the testing phase, similarly features were extracted for the test utterances of the corresponding databases. Their likelihoods were estimated against the trained models. The model against which it yielded the highest likelihood score was identified as the speaker.

4.2.1 Gender Recognition in Standard Database

The first set of experiments was conducted to perform gender recognition on the complete set of male and female files in the MIT database. The accuracies obtained for different sets of features are shown in Table 1.

Sl No	Features used	No. of mixtures	Male	Female	Overall accuracy
1	MFCC+DCE+ DDCE	16	90.01	97.96	93.795
2	MFCC+DCE+ DDCE+ MACV(5)	32	95	98.4	96.619
3	MFCC+DCE+ DDCE+ MACV(3)	64	94.36	97.7	95.95

TABLE1: Gender Recognition Accuracies on MIT Database.

The MFCC+DCE+DDCE served as our baseline system which gave an accuracy of 93.795%. Including 3 MACV features improved the results by almost 2% with a marginal increase in the dimensionality. Including 5 MACV features again increased the accuracy of the system.

4.2.2 Gender Recognition on Self-Collected Database

We repeated the same set of gender recognition experiments as discussed above on the self-collected database. For the feature set composed of MFCC, DCE, DDCE and 5 MACV features, we obtained 100% accuracies in distinguishing between the male and female speaker classes.

4.2.3 Speaker Recognition on Standard Database

I) In the 3rd set of experiments, 48 speaker models (48 male/female speakers) were trained with MFCC+ΔMFCC+ ΔΔMFCC (39 feature vector set) using 64 Gaussian mixture models and tested using the test utterances of the speakers (other than the training utterances). The results are shown in Table 2. We obtained an overall accuracy of 81.058% and out of which the female and male accuracies are 78.6209% and 83.1204%.

Spk.	Accuracy	Spk.	Accuracy	Spk.	Accuracy
f00	85.19	f16	85.19	m10	81.48
f01	81.48	f17	72.22	m11	92.59
f02	61.11	f18	70.37	m12	81.48
f03	59.26	f19	64.18	m13	77.78
f04	68.54	f20	90.74	m14	66.67
f05	79.63	f21	87.04	m15	88.89
f06	70.37	m00	92.59	m16	68.54
f07	98.15	m01	87.04	m17	77.78
f08	81.48	m02	83.33	m18	88.89
f09	87.04	m03	68.52	m19	83.33
f10	90.04	m04	83.33	m20	77.78
f11	75.93	m05	90.74	m21	77.78
f12	90.74	m06	98.15	m22	87.04
f13	85.19	m07	94.44	m23	83.33
f14	83.33	m08	88.89	m24	72.26
f15	61.11	m09	100	m25	68.52

TABLE2: Speaker Recognition Accuracies for MFCC+ΔMFCC+ ΔΔMFCC on MIT Database.

II) In the next set of experiments, the 48 speaker recognition system were built using CMS with 64 Gaussian mixtures. We improved overall accuracy to 83.869 % as compared to the baseline system accuracy of 81.058%. It is observed that though for few speakers, the accuracy went down as compared to standard MFCC but in general, it increased for all the speakers. The results are shown in Table 3.

Spk.	Accuracy	Spk.	Accuracy	Spk.	Accuracy
f00	88.89	f16	68.52	m10	85.19
f01	70.37	f17	66.67	m11	100
f02	85.19	f18	66.67	m12	74.07
f03	68.52	f19	81.48	m13	72.22
f04	53.70	f20	100	m14	70.37
f05	94.44	f21	92.59	m15	98.15
f06	74.07	m00	98.15	m16	100
f07	100	m01	94.44	m17	87.04
f08	83.33	m02	85.19	m18	83.33
f09	62.96	m03	83.33	m19	74.07
f10	92.59	m04	100	m20	92.59
f11	74.07	m05	85.19	m21	87.21
f12	85.19	m06	81.48	m22	81.48
f13	94.44	m07	100	m23	81.48
f14	98.15	m08	90.74	m24	62.96
f15	68.15	m09	100	m25	87.04

TABLE3: Speaker Recognition Accuracies for CMS Feature set on MIT Database.

We observe that system accuracy for 64 Gaussians is best for this dataset.

4.2.4 Speaker Recognition on Self-Collected Database

In the next set of experiments, we performed speaker recognition on the self-collected database. Preliminary 8-speaker models were built with single Gaussian mixture modeling. With the 39-vector set MFCC, 100% accuracies were obtained for the training data. In another set of experiments, 8 speaker models were made using MFCC and CMS features and the accuracies obtained was found to be 95.413%. The results are shown in Table 4.

Sl. No	Speaker	Accuracy
1	Spk 1	100
2	Spk 2	100
3	Spk 3	100
4	Spk 4	96.67
5	Spk 5	76.67
6	Spk 6	96.67

7	Spk 7	93.3
8	Spk 8	100

TABLE4: Speaker Recognition Accuracies for CMS Feature set on Self-Collected Database.

4.2.5 Hierarchical Speaker Recognition System

As discussed before in Section III, we then performed experiments to improve our speaker recognition accuracies by combining the system with a gender recognizer. The final system that was built would first classify the speaker's gender and then, it will recognize the speaker's identity in that speaker class. We performed the baseline experiment on the modified system and the results are shown in Table 5.

Spk.	Accuracy	Spk.	Accuracy	Spk.	Accuracy
f00	87.04	f17	68.52	m11	88.89
f01	59.26	f18	57.41	m12	74.07
f02	88.81	f19	94.07	m13	57.41
f03	66.67	f20	100	m14	75.47
f04	35.42	f21	96.30	m15	66.67
f05	90.74	m00	96.30	m16	96.30
f06	70.37	m01	94.44	m17	79.63
f07	98.15	m02	88.89	m18	83.33
f08	66.67	m03	66.67	m19	96.30
f09	72.22	m04	98.15	m20	88.89
f10	92.59	m05	90.74	m21	88.89
f11	81.48	m06	88.89	m22	81.48
f12	90.74	m07	100	m23	88.89
f13	96.30	m08	66.67	m24	70.37
f14	83.33	m09	96.30	m25	88.89
f15	79.63	m10	81.48	-	-
f16	77.78				

TABLE5: Table showing accuracies for speaker recognition using gender recognition

The overall system accuracy improved from 81.058% (with the baseline system) to 82.56% (with our novel proposed system). Also, the computational time of the system reduced from 64.812s to 44.078 seconds. The comparison of the accuracies of the systems 4.2.2(I) and 4.2.3 can be graphically seen in Figure 4. It is expected that performing the same experiments using the CMS feature would also improve the performance in a similar fashion, which we may perform in future.

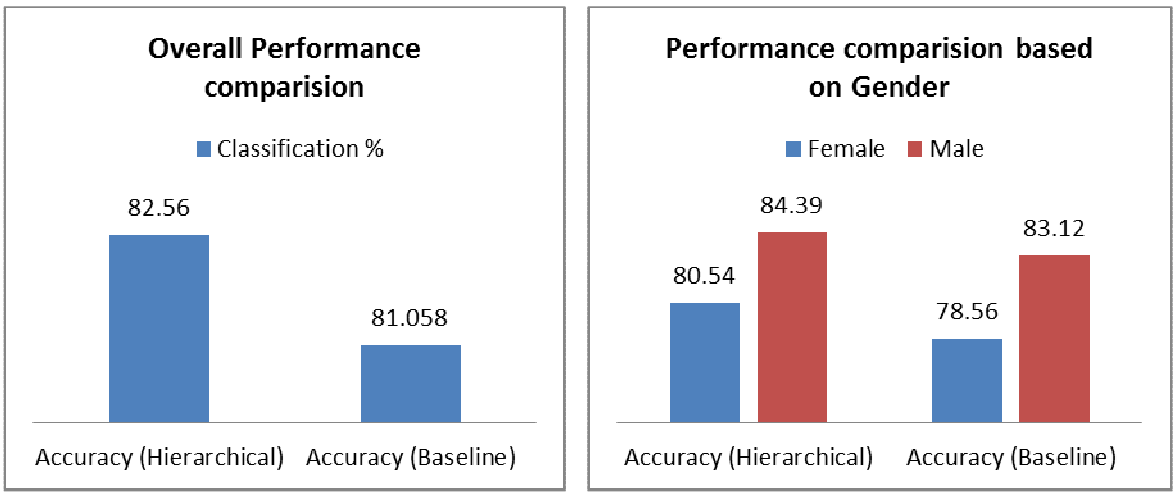

FIGURE 4: Figure showing the performance improvement using the hierarchical recognizer

5 Conclusion and Future Work

In this paper, we first propose a text-independent speaker recognizer using Gaussian Mixture Models. For a combination of 39 MFCC features, we obtained an accuracy of 81.058% on the MIT Database, which served as our baseline system. In another set of experiments, we demonstrated that using the CMS feature improves the accuracy of the system to 83.869%. Since the number of speakers in MIT Dataset is 48, the performance of system is around 83% but the recognition of self-collected dataset, which contains only 8 speakers is relatively high around 96%.

We then proposed a novel technique to improve the performance of our baseline speaker recognizer by implementing gender recognition before the speaker recognition. Through experimental results, we finally show that the enhanced system has improvedthe system accuracy by more than 1.85% while reducing computational time by over 30%. Thus, the proposed hierarchical approach provides a better performance compared to our baseline.

As a part of the future work, we would suggest to implement some additional features which would be having more speaker-relevant information. In future, the learning from this system can be adopted to build a real time system as this approach effectively reduces the recognition time. Also, the overall accuracy of the system can be possibly improved by developing GMMs which take care of the degree of overlap between different speaker classes and thereby, giving more weightage to the non-overlapped segments.

6 Acknowledgement

The authors gratefully acknowledge the contributions of Prof S. Umesh (*IIT Madras*), Dr Shakti Rath and DrSanand for their guidance during the initial phase of this project.

7 References

[1] X. Huang, A.Acero and H.-W.Hon, *Spoken language processin*g, Upper Saddle River, New Jersey, Prentice Hall PTR, 2001.

[2] S. Furui, Digital Speech Processing, Synthesis and Recognition, New York, Marcel Dekker, 2001.

[3] J. R. Deller, J. H. L. Hansen, J. G. Proakis, Discrete-Time Processing of Speech Signals, Piscataway (N.J.), IEEE Press, 2000.

[4] X. Huang, A. Acero and H.-W.Hon, Spoken language processing, Upper Saddle River, New Jersey, Prentice Hall PTR, 2001.

[5] D. A. Reynolds, “An Overview of Automatic Speaker Recognition Technology”, ICASSP 2002, pp 4072-4075.

[6] EvgenyKarpov, ‘Real-Time Speaker Identification’, University of Joensuu Department of Computer Science Master’s Thesis

[7] Mohamed FaouziBenZeghibaa, ‘Joint Speech And Speaker recognition' IDIAP RR 05- 28, February 2005

[8] J.R Deller, J.H.L. Hansen, J .G. Proakis, Discrete –Time processing of speech signals, Piscataway (N.J.),/IEEE Press,2000

[9] Brett Richard Wildermoth,'Text Independent Speaker Recognition using source based features', January 2001, Griffith university , Australia.

[10] H. Gish and M. Schmidt, “Text Independent Speaker Identification”, IEEE Signal Processing Magazine, Vol. 11, No. 4, 1994, pp. 18-32.

[11] MohaddesehNosratighods ,EliathambyAmbikairajah ,and Julien Epps “SPEAKER VERIFICATION USING A NOVEL SET OF DYNAMIC FEATURES”

[12] J .M.Naik ,”Speaker Verifiaction-A tutorial”, IEEE Communications Magazine, January 1990,pp.42-48.

[13] D. A. Reynolds, “An Overview of Automatic Speaker Recognition Technology”, ICASSP 2002, pp 4072-4075.

[14] J.P. Campbell, “Speaker Recognition: A Tutorial”, Proc. of the IEEE, vol. 85, no. 9, Sept 1997, pp. 1437-1462

[15] H. Gish and M. Schmidt, “Text Independent Speaker Identification”, IEEE Signal Processing Magazine, Vol. 11, No. 4, 1994, pp. 18-32.

[16] D. Reynolds, R. Rose, “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, IEEE transactions on speech and audio processing, Vol. 3, No1, 1995, pp. 72-83

[17] Jeff A. Bilmes , “A Gentle Tutorial of the EM Algorithm and its application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models”, TR-97-021, April 1998

[18] leonard, R. ,G. ,' A Database for speaker independent digit recognition' , Proc. ICASSP 84 , Volume 3, p. 42.11, 1984

[19] 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.

[20] S. Roberts, D. Husmeier, I. Rezek, andW.Penny, “Bayesian approaches to gaussian mixture modeling,” IEEE Trans. Pattern Anal. Machine Intell., vol. 20, pp. 1133–1142, Nov. 1998.

[21] Atal, B.S“Automatic recognition of speakers from their voices,” Proc. IEEE, vol. 64, pp. 460–475, 1976.

[22] SadaokiFurui“Speaker-dependent-feature extraction, recognition and processing techniques,” Speech Commun., vol. 10, pp. 505–520, 1991.

[23] Campbell W, Sturim D, Reynolds D, Solomonoff A. SVM-based speaker verification using a GMM supervector kernel and NAP variability compensation. In: Proceedings of the international conference on acoustics, speech and signal processing; 2006. p. 1–97.

[24] Herbert, M., 2008. Text-dependent speaker recognition. In: Benesty, J., Sondhi, M., Huang, Y. (Eds.), Springer Handbook of Speech Processing. Springer-Verlag, Heidelberg, pp. 743–762.