Performance evaluation of Statistical Approaches for Automatic Text-Independent Speaker Recognition using Robust Features

R. Rajeswara Rao[1] , A. Prasad[2] and Ch. Kedari Rao[3]

[1] Professor of CSE, MGIT
Hyderabad, India

[2]Professor of CSE,Vignan University, Guntur
Guntur, India

[3] Asst. Prof, Dept of CSE, DVRCET
Hyderabad, India

Abstract
This paper introduces the performance evaluation of statististical approaches for Automatic-text-independent Speaker Recognition system. Automatic-text-independent Speaker Recognition system is to quickly and accurately identify the person from his/her voice. The study on the effect of feature vector size for good speaker recognition demonstrates that the feature vector size in the range of 18-22 can capture speaker related information effectively for a speech signal sampled at 16 kHz. it is demonstrated that the timing varying speaker related information can be effectively captured using hidden Markov models (HMMs) than GMM. It is established that the HMM based speaker recognition system requires significantly less amount of data during both during training as well as in testing than GMM based Speaker Recognition System. The performance evaluation of speaker recognition study using robust features for HMM based method and GMM based method is exploited for different mixtures components, training and test durations We demonstrate the speaker recognition studies on TIMIT database.

Keywords: hidden Markov models (HMMs), Gaussian Mixture Model (GMM)), MFCC, Robust Features, Speaker

1. Introduction

Speaker recognition refers to recognizing persons from their voice. No two individuals sound identical because their vocal tract shapes, larynx sizes, and other parts of their voice production organs are different. In addition to these physical differences, each speaker has his or her characteristic manner of speaking, including the use of a particular accent, rhythm, intonation style, pronunciation pattern, choice of vocabulary and so on. State-of-the-art speaker recognition systems use a number of these features in parallel, attempting to cover these different aspects and employing them in a complementary way to achieve more accurate recognition.

An important application of speaker recognition technology is forensics. Much of information is exchanged between two parties in telephone conversations, including between criminals, and in recent years there has been increasing interest to integrate automatic speaker recognition to supplement auditory and semi-automatic analysis methods.
Automatic speaker recognition is an application of pattern recognition. Speaker recognition system, like any other pattern recognition system, can be represented as shown in Fig. 1. This task involves three phases, feature extraction phase, training phase and testing phase [1]. Training is the process of familiarizing the system with the voice characteristics of a speaker, whereas testing is the actual recognition task.

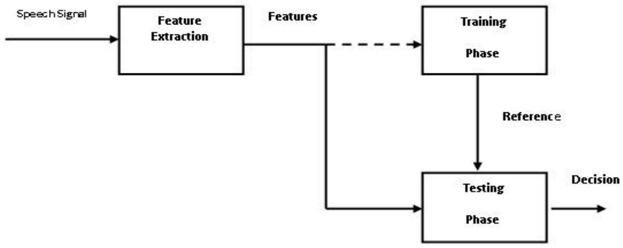


Fig. 1: A typical Block diagram representation of a speaker recognition task.

1.1 Feature extraction

For any pattern recognition task like Automatic Speaker Recognition (ASR), the relevant information has to be captured in terms of suitable feature vectors. In speaker recognition, the feature vectors are in general some parameter vectors extracted from frames of the speech signal. Most of the present day ASR systems are developed using parameters that are derived based on spectral analysis, and the speaker variability is captured in terms of the distribution of these feature vectors. But, it is a fact that the spectrum of a signal is prone to channel characteristics and noise. Channel characteristics and noise play a prominent role in the performance of spectral feature-based systems [2]. Another drawback with the existing techniques is the way in which speaker-specific information is being captured. Mostly, they are statistical techniques, capturing the variability in terms of distribution of the feature vectors and hence large amount of data is required for a better estimate.

Since all the real world services have to deal with speech coming over telephone channel, the ASR systems have to be robust to environmental variations. Also, the requirement of large amount of data has to overcome, as in the real world applications we may not have large amount of data to recognize a person. Hence, in order to make the ASR work in noisy conditions, and with less amount of data, features other than those derived based on spectral analysis also need to be explored.

1.2 Selection of Features

Speech signal includes many features of which not all are important for speaker discrimination. An ideal feature would

- have large between-speaker variability and small within-speaker variability
- be robust against noise and distortion
- occur frequently and naturally in speech
- be easy to measure from speech signal
- be difficult to impersonate/mimic
- not be affected by speaker's health or long-term variations in voice.

1.3 Motivation to use Mel frequency cepstral coefficients (MFCC)

Since our interest is in capturing global features which corresponds the low frequency or pitch components are to be emphasized. To fulfil this requirement it is felt that MFCC are most suitable as they emphasize low frequency and de-emphasize high frequencies

1.4 Mel frequency cepstral coefficients (MFCC)

In this phase the digital speech signal is partitioning into segments (frames) with fixed length 10-30 ms from which the features are extracted due to their spectral qualities. Spectrum is achieved with fast Fourier transformation [3]. Then an arrangement of frequency range to mel scale follows according to relation

$$f_{mel} = 2595 \log\left(1 + \frac{f_{Hz}}{700}\right) \quad (1)$$

By logarithm of amplitude of mel spectrum and applying reverse Fourier transformation we achieve frame cepstrum:

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

The FFT-base cepstral coefficients are computed by taking IFFT of the log magnitude spectrum of the Speech signal. The mel-warped cepstrum is obtained by inserting a intermediate step of transforming the frequency scale to place less emphasis on higher frequencies before taking the IFFT [4][5][6].

1.5 High-Level Features

Speakers differ not only in their voice timbre and accent/pronunciation, but also in their lexicon – the kind of words the speakers tend to use in their conversations. The work on such "high-level" conversational features was initiated in [7] where a speaker's characteristic vocabulary, the so-called idiolect, was used to characterize speakers. The idea in "high-level" modeling is to convert each utterance into a sequence of tokens where the co-occurrence patterns of tokens characterize speaker differences. The information being modeled is hence in categorical (discrete) rather than in numeric (continuous) form.

The tokens considered have included words [7], phones [8][9], prosodic gestures (rising/failing pitch/energy) [10][11][12], and even articulatory tokens (manner and place of articulation) [13]. The top-1 scoring Gaussian mixture component indices have also been used as tokens [14][15][16].

Sometimes several parallel tokenizers are utilized [9][14][17]. This is partly motivated by the success of parallel phone recognizers in state-of-the-art spoken language recognition [18][19]. This direction is driven by the hope that different tokenizers (e.g. phone recognizers trained on different languages or with different phone models) would capture complementary aspects of the utterance. As an example, in [14] a set of parallel GMM tokenizers [15[[16] were used. Each tokenizer was trained from a different group of speakers obtained by clustering.

One of the issues in speaker recognition is how to represent utterances that, in general, have a varying number of feature vectors. In early studies [20] speaker models were generated by time-averaging features so that each utterance could be represented as a single vector. The average vectors would then be compared using a distance measure [21] which is computationally very efficient but gives poor recognition accuracy. Since the 1980's, the predominant trend has been creating a model of the training utterances followed by "data-to-model" type of matching at run-time (e.g. likelihood of an utterance with respect to a GMM). This is computationally more demanding but gives good recognition accuracy.
Interestingly, the speaker recognition community has recently re-discovered a robust way to present utterances using a single vector, a so-called super vector [22].

2. Exploring Robust Features for Speaker Recognition

Here, the GMM is used as front-end to extract features vectors from speech signal. For the ASR task, the basic requirement is to obtain the feature vectors form the speech signal. Recently, some attempts are made to explore the alternative representation of feature vectors based on GMM feature extraction.
For Speaker Recognition task, robust features are derived from the speech signal based on estimating a Gaussian mixture model. The underlying speaker discrimination information is represented by Gaussians. The estimated GMM parameters means, co-variance and component weight can be related to the formant locations, bandwidths and magnitudes.
For the proposed new feature vectors, from the speech signal of a speaker S_i, a 12 dimensional MFCC feature vectors are obtained with a window size of 20ms and window shift of 5 ms. These MFCC feature vectors are distributed into 'R' Gaussians mixtures as shown in Fig. 2.

Fig. 2: R Gaussians for Speaker S_i .

The feature vector X=(X1, X2,……, X12) is passed through a Gaussian G1 by calculating a Gaussian probability P1 using Gaussian probability density function. This P1 is first coefficient in the new feature vector. In the same way feature vector X is passed through R Gaussians by creating R feature vector coefficients namely P1,P2,….,PR, as shown in Fig. 3. These R coefficients create a new R dimensional feature vector. The newly created R dimensional feature vector is shown in the Fig. 4.
Experiments are carried to find the dimension new feature vector for good speaker recognition performance. This is done by varying the number of Gaussians from 12 to 30, i.e number of coefficients in the new feature vectors. When the numbers of coefficients are 20, the good identification performance is achieved.

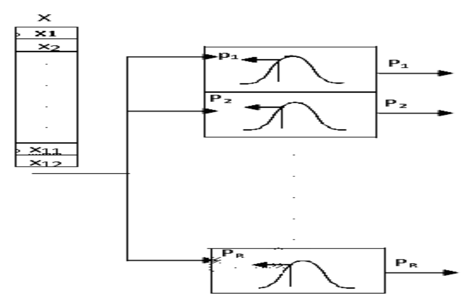


Fig. 3: Parameter estimation for new vector P. When R=15, the optimal recognition performance has been achieved.

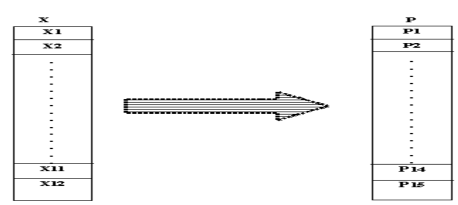


Fig. 4: Transforming from 12 dimensional MFCC feature vector to R dimensional feature vector.

3. Performance evaluation of Parametric approaches

Parametric approaches are model-based approaches. The parameters of the model are estimated using the training feature vectors. It is assumed that the model is adequate to represent the distribution. The most widely used parametric approaches are GMM and HMM based approaches.

3.1 Gaussian Mixture Models for LID

GMM is a classic parametric method best used to model speaker identities due to the fact that Gaussian components have the capability of representing speaker discrimination information effectively. Gaussian classifier has been successfully employed in several text-independent speaker recognition applications. As shown in Fig. 5 in a GMM model, the probability distribution of the observed data takes the form given by the following equation [23][24].

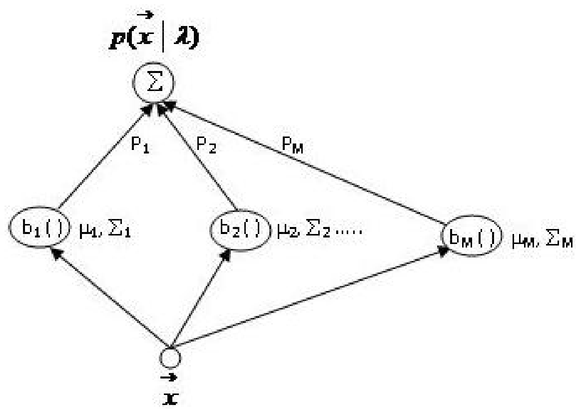


Fig. 5: Gaussian Mixture Model

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

Where M is the number of component densities, $\bar{x}$ is a D dimensional observed data (random vector), $b_i(\bar{x})$ are the component densities and p_i are the mixture weights for i = 1, .., M.

$$b_i(\bar{x}) = \frac{1}{(2\pi)^{D/2} |\sum_i|^{1/2}} \exp\left\{-\frac{1}{2}(\bar{x}-\bar{\mu}_i)^T \sum_i{}^{-1} (\bar{x}-\bar{\mu}_i)\right\}$$

Each component density $b_i(\bar{x})$ denotes a D-dimensional normal distribution with mean vector $\bar{\mu}_i$ and covariance matrix $\sum_i$. The mixture weights satisfy the condition $\sum_{i=1}^{M} p_i = 1$ and therefore represent positive scalar values. These parameters can be collectively represented as $\lambda = \{p_i, \bar{\mu}_i, \sum_i\}$ for i = 1 … M. Each speaker in a language system can be represented by a GMM and is referred by the language respective model λ.

The parameters of a GMM model can be estimated using maximum likelihood (ML) [25] estimation. The main objective of the ML estimation is to derive the optimum model Parameters that can maximize the likelihood of GMM. Unfortunately direct maximization using ML

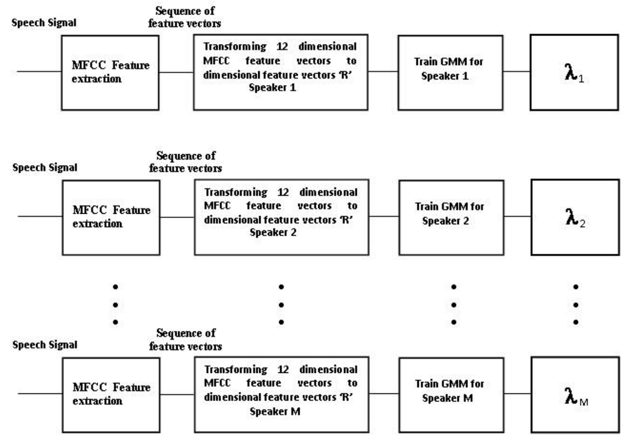


Fig. 6: Training GMM for Speaker Recognition Task

Parameter estimation is not possible and therefore a special case of ML estimation known as Expectation-Maximization (EM) [25] algorithm is used to extract the model parameters.

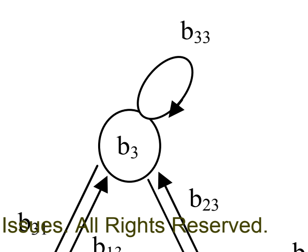
The GMM likelihood of a sequence of T training vectors $X = \{\bar{x}_1, .. \bar{x}_T\}$ can be given as [25]

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\bar{x}_t \mid \lambda)$$

The EM algorithm begins with an initial model λ and tends to estimate a new model $\bar{\lambda}$ such that $p(X \mid \bar{\lambda}) \geq p(X \mid \lambda)$ [19]. This is an iterative process where the new model is considered to be an initial model in the next iteration and the entire process is repeated until a certain convergence threshold is obtained.

3.2 Continuous Ergodic Hidden Markov model for speaker recognition

The HMM is a doubly embedded stochastic process where the underlying stochastic process is not directly observable. HMMs have the capability of effectively modeling statistical variations in spectral features. In a variety of ways, HMMs can be used as probabilistic speaker models for both text-dependent and text-independent speaker recognition [27][28][29]. HMM not only models the underlying speech patterns but also the temporal sequencing among the sounds. This temporal modeling is advantageous for text-dependent speaker recognition system. Left Right HMM can model temporal sequence of patterns only, where as to capture the patterns of different type ergodic HMM is used [21].



IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012
ISSN (Online): 1694-0814
www.IJCSI.org

FIG. 4:

Fig. 7: Three-state ergodic HMM.

As shown in the Fig. 8 in the training phase, one HMM for each speaker is obtained (i.e., parameters of model are estimated) using training feature vectors. The parameters of HMM are [19] State-transition probability distribution: It is represented by $A = \lfloor a_{ij} \rfloor$

Where

$$a_{ij} = P(q_{t+1} = j \mid q_t = i) \quad 1 \le i, j \le N \qquad (2)$$

defines the probability of transition from state i to j at time t.

For a three state left-right model the state transition matrix is given as

$$A = \{a_{ij}\} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix} \qquad (3)$$

The state transition matrix of three state ergodic model is given by

$$A = \{a_{ij}\} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \quad (4)$$

Observation symbol probability distribution: It is given by $B = \lfloor b_j(k) \rfloor$ in which

$$b_j(k) = P(O_t = V_k \mid q_t = j) \quad 1 \le k \le M \ (5)$$

defines the symbol distribution in state $j = 1,2,3.....N$. The initial state distribution is given by $\pi = P(q_1 = i)$ where

$$\pi_i = P(q_1 = i) \quad 1 \le i \le N \quad (6)$$

Here, N is the total number of states, and q_t is the state at time t, M is the number of distinct observation symbols per state, and O_t is the observation symbol at time t. In testing phase, $P\left(O/\lambda\right)$ for each model is calculated, where $O = (O_1 O_2 O_3 ....O_T)$ Here the goal is to find out the probability for a given model to which the test utterance belongs to. The speaker whose model gives the highest score is declared as the identified speaker. GMM corresponds to a single-state continuous ergodic HMM.

The model parameters can be collectively represented as $\lambda = (A_i, B_i, \pi_i)$ for $i = 1........M$. Each speaker in a speaker identification system can be represented by a HMM and is referred to by the speaker's respective models λ.

In the testing phase, p (O/λ) for each model is calculated [29]. Where O=(o1o2o3…OT) is the sequence of the test feature vectors. The goal is to find the probability, given the model that the test utterance belongs to that particular model. The speaker model that gives the highest score is declared as the identified speaker.

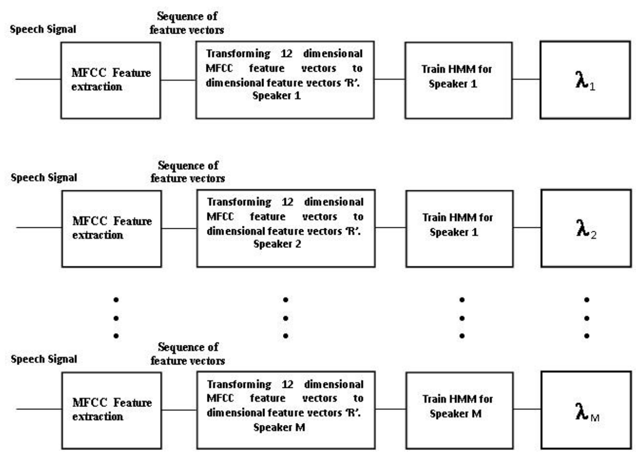


Fig. 8: Training HMM for Speaker Recognition Task

The parameter estimation is not possible and therefore a special case of ML estimation known as Expectation-Maximization (EM) [30] algorithm is used to extract the model parameters.

The GMM likelihood of a sequence of T training vectors $X = \{\bar{x}_1,..\bar{x}_T\}$ can be given as [30]

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\bar{x}_t \mid \lambda)$$

The EM algorithm begins with an initial model λ and tends to estimate a new model $\bar{\lambda}$ such that $p(X \mid \bar{\lambda}) \ge p(X \mid \lambda)$ [29]. This is an iterative process where the new model is considered to be an initial model in the next iteration and the entire process is repeated until a certain convergence threshold is obtained.

4. Experimental Evaluation

4.1 Database used for the study

Speaker Recognition is the task of identifying the speaker from the registered set of speakers. In this paper we consider identification task for TIMIT Speaker database [30].

The TIMIT corpus of read speech has been designed to provide speaker data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speaker recognition systems. TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. We consider 100 male speakers and 100 female out of 630 speakers for speaker recognition. Maximum of 30 sec. of speech data is used for training and minimum of 1 sec. of data for testing. In all the cases the speech signal was sampled at 16 kHz sampling frequency. Through out this study, closed set identification experiments are done to demonstrate the feasibility of capturing the speaker-discrimination information from the speech signal. Requirement of significantly less amount data for speaker-discrimination information and Gaussian mixture models is also demonstrated.

4.2 Experimental setup

The system has been implemented in Matlab 7 on Windows XP platform. We have trained the GMM model using Gaussian Components as 4, 8, 16, 32 and 64 and HMM with 2, 3 and 4 states with 4, 8, 16, 32 and 64 components at each state for training speech duration of 10, 20 and 30 sec. testing is performed using different test speech durations such as 1 sec., 3 sec., and 5 sec..

5. Performance Evaluation

The system has been implemented in Matlab7 on windows XP platform. The result of the study has been presented in Table 3. We have used coefficient order of 20 for all experiments. We have trained the GMM and HMM models using Gaussian components as 4, 8, 16, 32 and 64 components at each state for different training speech lengths as 10 sec., 20 sec., and 30 sec. by varying HMM states such as 2, 3 and 4. Testing is performed using different test speech lengths such as 1 sec, 3 sec, and 5 sec.. Here, recognition rate is defined as the ratio of the number of speaker identified to the total number of speakers tested. As shown in Fig. 9, the speaker recognition performance for varying co-efficient \order is depicated. It is established that the co-efficient order range of of 18-22 is found to be optimal for good speaker recognition. As shown in Fig. 10, Fig. 11, Fig. 12 for a 3-state HMM, the recognition rate for testing length for 5 sec. outperformed, where as for testing length of 3 sec. is also on par with 5 sec. testing length.

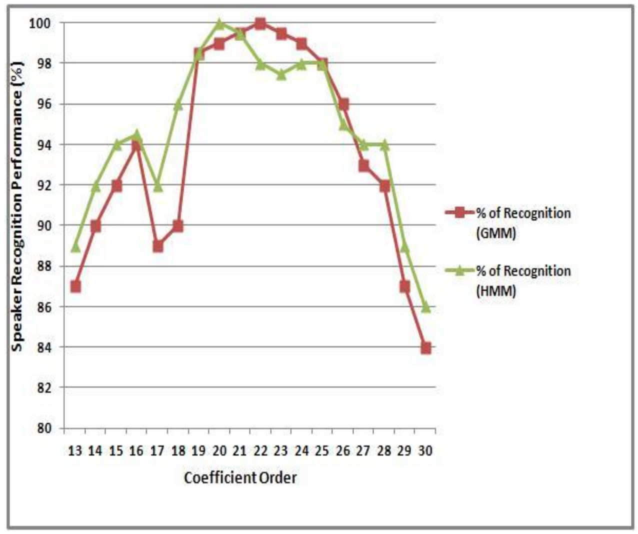


Fig. 9: Speaker Recognition Performance for varying Co-efficient order

Table 1: Speaker Recognition Performance for varying Co-efficient order

Coefficients Order	Speaker Recognition (%) of GMM	Speaker Recognition (%) of HMM
13	87	89
14	90	92
15	92	94
16	94	94.5
17	89	92
18	90	96
19	98.5	98.5
20	99	100
21	99.5	99.5
22	100	98
23	99.5	97.5
24	99	98
25	98	98
26	96	95
27	93	94
28	92	94
29	87	89
30	84	86

Table 2: Avg. Speaker Recognition Performance for varying test duration.

As shown in Table 3,, for a 3-state HMM the percentage (%) recognition for Gaussian Components such as 4, 8, 16, 32 and 64 seems to be uniformly increasing. The minimum number of Gaussian components to achieve good speaker recognition performance seems to be 16 and thereafter the recognition performance is minimal. As shown in Fig. 12.

No. of Mixture Components	Avg. Speaker Recognition (%)					
	Test Duration					
	1 Sec.		3 Sec.		5 Sec.	
	GMM	HMM	GMM	HMM	GMM	HMM
4	54.66	86.66	86	92.83	92.66	96.16
8	67.83	90.5	95	96.33	97.33	98.66
16	74.66	92.16	97.5	98.16	99	99
32	82.85	94	97.83	97.66	99	98.83
64	79.83	92	96	96.33	98.16	98.16

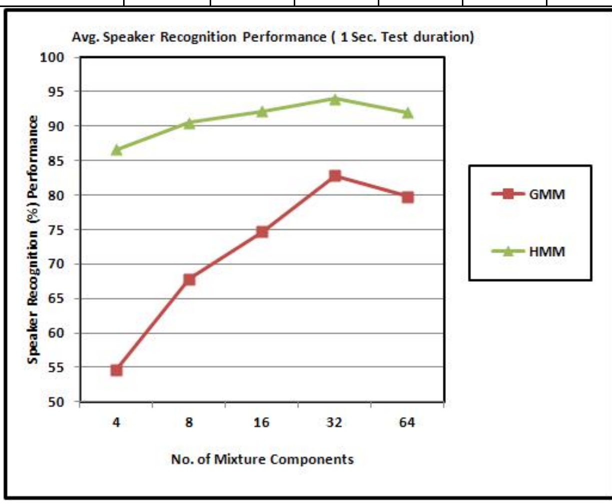


Fig. 10: Avg. Speaker Recognition Performance of GMM/HMM for varying test duration of 1Sec.

As shown in Fig. 10, the average speaker recognition performance for 10 sec., 20 sec. and 30 sec. training duration for varying mixture components as 4, 8, 16, 32 and 64 tested with 1 sec., 3 sec., and 5 sec., test durations for 2, 3 and 4 states indicate that for 20 sec., of training speech duration with 16 mixture components test duration of 3 sec. And for 3-state HMM gives good speaker recognition performance.

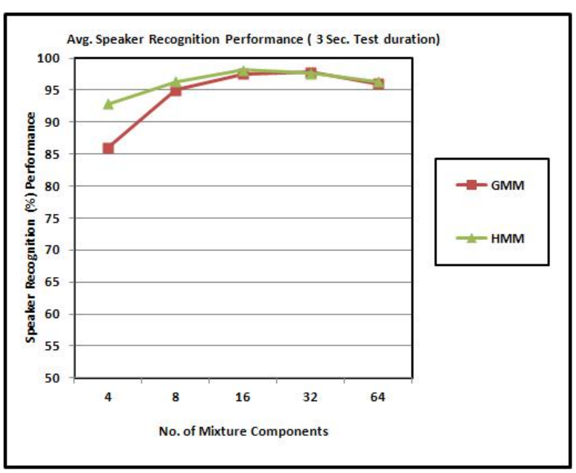


Fig. 11: Avg. Speaker Recognition Performance of GMM/HMM for varying test duration of 3 Sec.

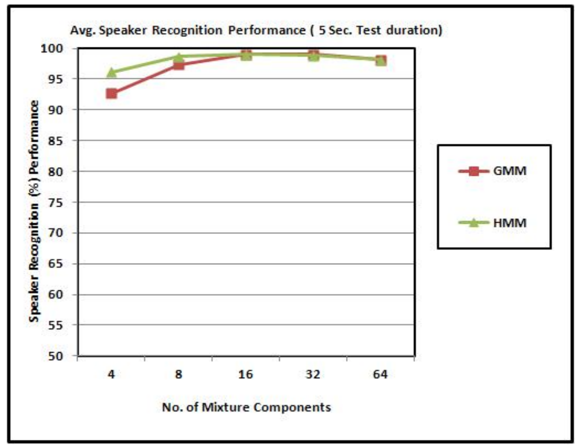


Fig. 11: Avg. Speaker Recognition Performance of GMM/HMM for varying test duration of 5 Sec.

Table 3: Avg. Speaker Recognition Performance of GMM & HMM for 20 Sec. Training duration.

No. of Mixture Components	Avg. Speaker Recognition Performance of GMM	Avg. Speaker Recognition Performance of HMM
4	78.33	96.83
8	87.66	98.33
16	90.83	99.5
32	93.33	99.16
64	94.16	97.83

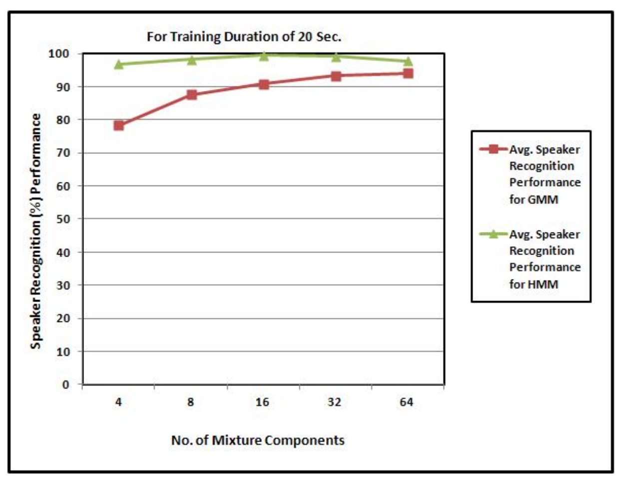


Fig. 11: Avg. Speaker Recognition Performance of GMM & HMM for 20 Sec. training duration

1 CONCLUSION

In this work we have demonstrated the importance of coefficient order for speaker recognition task. Speaker discrimination information is effectively captured for coefficient order 20 using a HMM .The recognition performance depends on the training speech length selected for training to capture the speaker-discrimination information. Larger the training length, the better is the performance, although smaller number reduces computational complexity.

Effectiveness of the HMM for speaker recognition task using the time varying speech signal is demonstrated. GMM based approaches do not capture prosodic information and acoustic variations, which vary with time. Hence, in order to capture time varying properties in the speech signal effectively, HMM based system is well suited for speaker recognition task. Continuous HMM embeds benefits associated with GMM approaches. We have not made any attempt to optimize the parameters of the model used for feature extraction, and also the decision making stage. Therefore the performance of language identification may be improved by optimizing the various design parameters

The objective in this paper was mainly to demonstrate the significance of the speaker-discrimination information present in the speech signal for speaker recognition. We have not made any attempt to optimize the parameters of the model used for feature extraction, and also the decision making stage. Therefore the performance of speaker recognition may be improved by optimizing the various design parameters.

.

References

[1]. Furui, "An overview of speaker recognition technology", in Automatic Speech and Speaker Recognition (C.-H. Lee, F. K. Soong, and K. K. Paliwal, eds.), ch. 2, pp. 31-56, Boston: Kluwer Academic, 1996.

[2] D. A. Reynolds, "The effects of handset variability on speaker recognition performance: Experiments on switch board corpus," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, pp. 113-116, 1996.

[3] Dempster, A., Laird, N., and Rubin, D., "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, vol. 39, pp. 1-38, 1977.

[4] Makhoul, J., 1975. Linear prediction: a tutorial review. Proc. IEEE 63, 561-580.

[5] Molau, S., Pitz, M., Schluter, R., and Ney, H., "Computing Mel-frequency cepstral coefficients on the power spectrum,“ Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. 73-76, May. 2001.

[6] Picone, J. W., "Signal modeling techniques in speech recognition," Proceedings of IEEE, vol. 81, no.9, pp. 1215-1247, Sep. 1993.

[7] Doddington.G., Speaker Recognition based on idiolectal differences between speakers. In Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001) (Aalborg, Denmark, September 2001), pp. 2521 – 2524.

[8] Andrews.W., Kohler.M., Campbell.J., Godfrey.J., and Hernandez –Cordero.J. Gender-dependent phonetic refraction for speaker recognition. In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002) (Orlando, Florida, USA, May 2002), vol. 1, pp. 149-152.

[9] Campbell.W., Campbell.J., Reynolds.D., Jones.D., and Leek.T., Phonetic speaker recognition with support vector machines. In Advances in Neural Information Processing Systems 16, S.Thrun, L.Saul, and B.Scholkopf, Eds. MIT Press, Cambridge, MA, 2004.

[10] 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 (ICASSP 2003) (Hong Kong, China, April 2003), pp. 788-791.

[11] Chen, Z.-H., Liao, Y.-F., and Juang, Y.-T. Eigen-Prosody analysis for robust speaker recognition under mismatch handset environment. In Proc. Int. Conf. on Spoken language processing (ICSLP 2004) (Jeju, South Korea, October 2004), pp. 1421-1424.

[12] Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., and Stolcke, A. Modeling Prosodic feature sequences for speaker recognition. Speech Communication 46, 3-4 (July 2005), 455-472.

[13] Leung, K., Mak, M., Siu, M., and Kung, S. Adaptive articulatory feature – based conditional pronunciation modeling for speaker verification. Speech Communication 48, 1 (January 2006), 71-84.

[14] MA, B., Zhu, D., Tong, R., and Li, H. Speaker cluster based GMM tokenization for Speaker recognition. In Proc. Interspeech 2006 (ICSLP) (Pittsburgh, Pennsylvania, USA, September 2006), pp. 505-508.

[15] Torres-Carrasquillo, P., Reynolds, D., and Jr., J. D. Language identification using Gaussian mixture model tokenization. In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002) (Orlando, Florida, USA, May 2002), vol. 1, pp. 757-760.

[16] Xiang, B. Text-independent speaker verification with dynamic trajectory model. IEEE Signal Processing Letters 10 (May 2003), 141 -143.

[17] Jin, Q., Schultz , T., and Waibel, A. Speaker identification using multilingual phone strings. In Proc. Int. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002) (Orlando, Florida, USA, May 2002), vol. 1, pp. 145-148.

[18] Zissman, M. Comparison of four approaches to automatic language identification of telephone speech. IEEE Trans. on Speech and Audio Processing 4, 1 (January 1996), 31-44.

[19] MA, B., Li, H., and Tong, R. Spoken language recognition with ensemble classifier. IEEE Trans. Audio, Speech and Language Processing 15, 7 (September 2007), 2053-2062.)

[20] Markel, J., Oshika, B., and A.H. Gray, J. Long-term feature averaging for speaker recognition. IEEE Trans. Acoustics, Speech, and Signal Processing 25, 4 (August 1977), 330-337.

[21] 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 (ISCSLP'06) (Singapore, December 2006), pp.559-567.

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

[23] Gish, H., Krasner, M., Russell, W., and Wolf, J., " Methods and Experiments for text-independent speaker recognition over telephone channels," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 11, pp. 865-868, Apr. 1986.

[24] Reynolds, D.A., and Rose, R.C., "Robust Text-Independent Speaker Identification using Gaussian Mixture Models," IEEE-Transactions on Speech and Audio Processing, vol. 3, no.1, pp. 72-83,1995.

[25] A. P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", J. Royal Statist. Soc. Ser. B. (methodologies), vol. 39, pp. 1-38, 1977.

[26] Gish, H., Krasner, M., Russell, W., and Wolf, J., " Methods and Experiments for text-independent speaker recognition over telephone channels," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 11, pp. 865-868, Apr. 1986.

[27] Reynolds, D.A., and Rose, R.C., "Robust Text-Independent Speaker Identification using Gaussian Mixture Models," IEEE-Transactions on Speech and Audio Processing, vol. 3, no.1, pp. 72-83,1995.

[28] A. P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", J. Royal Statist. Soc. Ser. B. (methodologies), vol. 39, pp. 1-38, 1977.

[29] K.N. Stevens, Acoustic Phonetics. Cambridge, England: The MIT Press, 1999.