J Inf Process Syst, Vol.14, No.4, pp.807~823, August 2018
https://doi.org/10.3745/JIPS.01.0030

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Speaker Verification with the Constraint of Limited Data

Thyamagondlu Renukamurthy Jayanthi Kumari* and Haradagere Siddaramaiah Jayanna*

Abstract

Speaker verification system performance depends on the utterance of each speaker. To verify the speaker, important information has to be captured from the utterance. Nowadays under the constraints of limited data, speaker verification has become a challenging task. The testing and training data are in terms of few seconds in limited data. The feature vectors extracted from single frame size and rate (SFSR) analysis is not sufficient for training and testing speakers in speaker verification. This leads to poor speaker modeling during training and may not provide good decision during testing. The problem is to be resolved by increasing feature vectors of training and testing data to the same duration. For that we are using multiple frame size (MFS), multiple frame rate (MFR), and multiple frame size and rate (MFSR) analysis techniques for speaker verification under limited data condition. These analysis techniques relatively extract more feature vector during training and testing and develop improved modeling and testing for limited data. To demonstrate this we have used mel-frequency cepstral coefficients (MFCC) and linear prediction cepstral coefficients (LPCC) as feature. Gaussian mixture model (GMM) and GMM-universal background model (GMM-UBM) are used for modeling the speaker. The database used is NIST-2003. The experimental results indicate that, improved performance of MFS, MFR, and MFSR analysis radically better compared with SFSR analysis. The experimental results show that LPCC based MFSR analysis perform better compared to other analysis techniques and feature extraction techniques.

Keywords

Gaussian Mixture Model (GMM), GMM-UBM, Multiple Frame Rate (MFR), Multiple Frame Size (MFS), MFSR, SFSR

1. Introduction

Speech is one of the communication media between people [1] and it can be used as an example of a biometric authentication to recognize a person [2]. When a speaker is recognized by use of vocal characteristic is called speaker recognition [3]. Speaker recognition contains speaker verification and speaker identification [4]. Accepting or rejecting the identity claim of a speaker is called speaker verification [5]. Speaker verification is one of the present days energizing technologies with very high potential [6].

Speaker verification under limited data conditions defines that verifying speakers with small amount of train/test data. In the current scenario, speaker verification works very well for sufficient data. Sufficient data refers to speech data of few minutes (greater than 1 minute) and on the other hand,

Manuscript received January 27, 2016; first revision January 3, 2017; accepted February 9, 2017.
Corresponding Author: Thyamagondlu Renukamurthy Jayanthi Kumari (trjayanthikumari@gmail.com)
* Dept. of Information Science and Engineering, Siddaganga Institute of Technology, Karnataka, India ({trjayanthikumari, jayannahs}@gmail.com)

limited data means speech data of few seconds (less than 15 seconds). Different techniques can be used to analysis speech data. The analysis techniques contains frame size (FS) and frame rate (FR) and to extract speaker specific information different size of FS and FR are used [7].

State-of-the-art speaker verification systems contains different analysis techniques [7–9] as shown in Fig. 1.

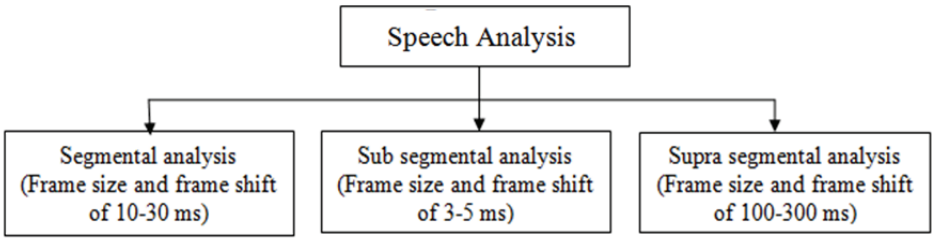

Fig. 1. Speech analysis techniques.

In segmental analysis, FS and FR in the range of 10–30 ms are considered to extract vocal tract information known as single frame size and single frame rate (SFSR) [9,10].

In sub-segmental analysis, speech can be analyzed using FS and FR in the range of 3–5 ms is preferred because excitation source information varies more quickly than that of vocal tract information [7,9,11]. In suprasegmental analysis, the behavioral aspect of speaker can be captured by varying FS and FR in the range of 100–300 ms [11–13]. When compared to vocal tract information, the behavioral characteristics vary slowly, therefore we require larger duration for FS and FR to capture specific information of speaker's [14].

The normal speaker verification systems use features that are extracted using SFSR. In case of SFSR analysis the speech signals are windowed into FS of 20–30 ms and FR of 10–20 ms [15]. This is because speech signal are non-stationary signals and shows quasi-stationary behavior at shorter durations. The problem in case of SFSR is that sometimes test speaker's speaking rate or pitch frequency does not match with the speaker's data during training [15,16]. Another problem in case of SFSR is due to single frame size capturing sudden changes in spectral information which may not be possible [16]. The SFSR analysis may not be able to provide sufficient feature vectors for training and testing the speakers in limited data [15].

In the existing speaker verification under limited data condition both FS and FR is fixed throughout the experiment and extracted feature vectors is also less in numbers. To resolve this problem we need more feature vectors. This can be done by varying FS and FR for different sizes. In case of variable frame size and rate (VFSR) [17–20], the spectral information changes with time due to the change of FR. The only disadvantage to this is that it needs additional time for computation. The problem can be overcome by using multiple frame size and rate (MFSR) analysis technique [15].

The leftover of the paper is organized as follows: MFSR analysis technique for speaker verification studies explained in Section 2. The studies related to speaker verification system presented in Section 3. The detailed experimental results with discussion are given in Section 4. The present work summary and possible path for future work are highlighted in Section 5.

2. Speaker Verification Using MFSR Analysis

In case of SFSR, speech is analyzed with FS of 20 ms and with FR of 10 ms is considered. But in case of multiple frame size (MFS) speech is analyzed by varying FS and maintaining FR constant, in case of multiple frame rate (MFR), maintaining FS constant and FR is varied and in case of MFSR both FS and FR are varied.

2.1 MFS Analysis

In case of MFS, speech data is analyzed by varying different FS and maintaining constant FR and also called as multi-resolution analysis technique [15]. The feature vectors extracted from speech data for different FS and magnitude spectra are noticeably dissimilar for different frequency resolutions [15]. The reason is spectral domain information obtained by convolution of spectral domain window and true spectrum of speech [10]. In addition to this, there will be little variation of FS of each speech samples. These two factors demonstrate the speaker information in different feature vectors. From this it is clear that by varying FS, the speaker specific information obtained from feature vector are different and spectral information also varies.

The actual feature vectors (N_f) vary from speaker to speaker for the similar quantity of speech data (DS) and is given by [15]:

$$N_f = \left(\frac{1-FS}{FR}\right) + 1 - N_{VAD} \tag{1}$$

Number of frames due to energy-based voice activity detection (VAD) technique is represented by N_{VAD}. In this analysis technique the features are extracted by considering FS = {5, 10, 15, 20} ms and FR constant as 10 ms. Eq. (2) represents extracted features for different frames to verify speaker.

$$N_f = \sum_{i=1}^{4} DS\left(\frac{1-FS_i}{FR}\right) + 1 - N_{VAD} \tag{2}$$

2.2 MFR Analysis

In MFR analysis, maintaining constant FS and varying different FR used to analysis the speech data. We know that pitch and speaking rate varies from one speaker to the other.

Pitch rate and speaking rate are different for different speakers and their spectral information is also different [21]. The single FR may not be possible to managing these variations. To overcome this problem, we need to analyze speech data for different FR and feature vectors extracted from vocal tract information will be different. Hence MFR analysis is also called multi-shifting analysis technique [15]. By varying FR spectral resolution remains same and there will be new set of feature vectors for each rate.

In MFR technique the features are extracted by considering FR= {2.5, 5, 7.5, 10} ms and FS constant as 20 ms. The extracted features for different FR is given by

$$N_f = \sum_{j=1}^{4} DS\left(\frac{1-FS}{FR_j}\right) + 1 - N_{VAD} \tag{3}$$

2.3 MFSR Analysis

In MFSR analysis, the multiple FS and FR analysis are used to analysis speech data. This analysis is the combination of MFR and MFS techniques [22].

The magnitude spectra will be different in MFS and MFR, resulting feature vectors are also different from each other [15]. The extracted feature vectors are more in number as compared to SFSR, MFR and MFS. In case of MFSR the feature vectors contains rich speaker-specific information. Further, the extracted features are obtained by combination of MFS and MFR analysis techniques. In the present work, we considered FS of {5, 10, 15, 20} ms and FR of {2.5, 5, 7.5, 10} ms. The extracted features for different FS and FR are given by

$$N_f = \sum_{i=1}^{4} \sum_{j=1}^{4} DS\left(\frac{1-FS_i}{FR_j}\right) + 1 - N_{VAD} \quad (4)$$

In order to study of these analysis techniques, we used mel-frequency cepstral coefficients (MFCC) and linear prediction cepstral coefficients (LPCC) feature extraction methods. Here we considered 13 and 39 dimensional feature vectors.

3. Speaker Verification Studies

3.1 Speech Database for Current Work

In present study, NIST-2003 database is used to evaluate the performance of speaker verification system [23]. The database contains 356 and 2556 train and test speakers respectively. The train data contain 149 male and 207 female speakers. The database also contains universal background model (UBM) speech with 502 speech data. In that 251 male and 251 female are available. The database speech data varies from seconds to few minutes. This work belongs to limited data, we have created train and test speakers are of duration three, four, five, six, nine and twelve seconds for the present study. The new database is used to conduct experiments for SFSR, MFR, MFS, and MFSR analysis techniques.

3.2 Feature Extraction Using MFCC and LPCC

The state-of-the-art speaker verification systems widely use either the MFCC or LPCC as features [24]. The feature extraction is used to extract speaker-specific information in the form of feature vector [25]. In speaker verification system to verify the speaker these features are extensively used [3]. The MFCC and LPCC techniques are used to extract features in present work. In MFCC, the spectral distortion can be minimized by using hamming window and applying Fourier Transform to windowed frame signal to get magnitude frequency response. In order to get cepstral coefficients, the discrete cosine transform (DCT) is applied to the output of the mel filters.

LPC can be calculated either by the autocorrelation or covariance methods directly from the windowed portion of speech [26]. The LPCC can be easily obtained by Durbin's recursive procedure without computing the discrete Fourier transform (DFT) and the inverse DFT, which are computationally complex and time consuming processes [27]. In both cases features are extracted using

SFSR, MFR, MFS, and MFSR methods.

The feature set LPCC or MFCC contains only static properties of a given frame of speech. The dynamic characteristics also contain some speaker specific information, these features also useful for recognition of speakers [28]. There are two types of dynamics in speech processing [28]:

- Δ*features* is the average first-order temporal derivative. This is determined by its velocity of the features.
- ΔΔ*features* is average second order temporal derivative. This is determined by its acceleration of the features.

Literature survey reveals that the significance of MFCC and LPCC feature for speaker verification in the constraint of limited data is not studied. Therefore in this work, effectiveness of both feature extraction techniques is studied.

3.3 Speaker Modeling and Testing

Different modeling techniques are available for speaker modeling including vector quantization (VQ), hidden Markov model (HMM), Gaussian mixture model (GMM) and GMM-UBM, etc. For the present work GMM-UBM modeling method is used. The GMM-UBM normally used when both training and testing data are less in size [29].

The final stage in speaker verification system is testing stage. In testing stage, test feature vectors are compared with reference models [5]. The log likelihood ratio test method [30] is adopted in this work.

4. Experimental Results and Discussions

The features are extracted using MFCC and LPCC techniques. The features are in the dimension of 13 and 39. The speakers are modeled using GMM and GMM-UBM techniques. The ideal speaker verification system should accept all the true speakers and reject all the false speakers [4]. Speaker verification performance can be measured in terms of equal error rate (EER). It is an operating point where the false rejection rate (FRR) equals the false acceptance rate (FAR) [31]. NIST-2003 database is used to test trained model.

In the 1st experiment, the features are in the dimension of 13 extracted using MFCC and LPCC techniques and modeling is done using GMM is shown in Fig. 2(a) and (b), respectively. The analysis technique used is SFSR. The experiment is conducted for 3, 4, 5, 6, 9 and 12 sec data for different Gaussian mixtures. Further, the minimum EER is 45.17%, 44.21%, 42.36%, 41.59%, 38.25%, and 36.68% for 3, 4, 5, 6, 9, and 12 seconds, respectively. It was observed that, the minimum EER 36.68% is obtained for 12 seconds data for Gaussian mixtures of 16 compared to remaining data sizes. The average EER can be calculated by considering minimum EER obtained for all the data size of different Gaussian mixtures. The average EER of MFCC-SFSR is 41.39%.

In LPCC-SFSR, the minimum EER is obtained for 3, 4, 5, 6, 9, and 12 seconds data is 43.08%, 41.41%, 39.97%, 38.7%, 31.34%, and 28.18% respectively for different Gaussian mixtures. Further, among these different data size, the least minimum EER is for 12 seconds data. The average EER in case of LPCC-SFSR is 37.21%. The average EER of LPCC-SFSR is less compared to MFCC-SFSR analysis. When

compared the EER of both analysis, LPCC-SFSR is 4.18% lesser than MFCC-SFSR. Further comparing both MFCC-SFSR and LPCC-SFSR for all data size, minimum EER obtained for 12 seconds data. In limited data both train/test data size is limited to less than or equal to 15 seconds. Therefore remaining experimental results are analyzed only for 12 seconds data. In SFSR analysis both FS and FR are fixed and available train/test data are also limited. Because of this, extracted features are also less in numbers. This will not create good speaker modeling and also speaker testing may not occur accurately.

To overcome this problem, we need to increase the features vectors. This can be done by using MFR, MFS, and MFSR analysis techniques.

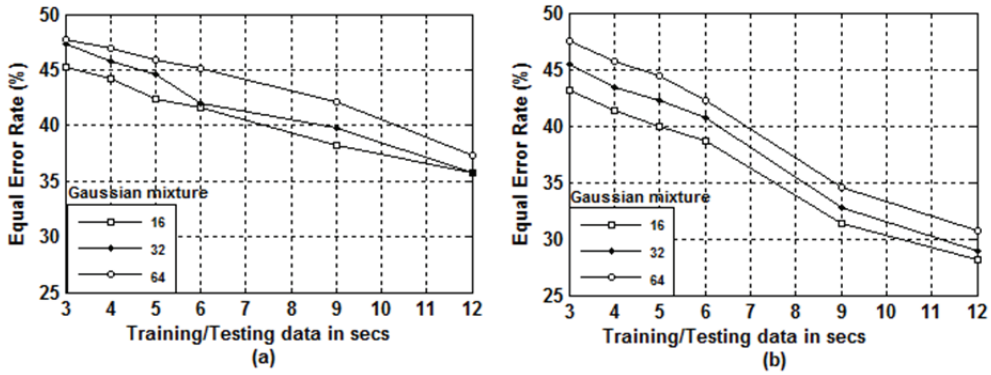

Fig. 2. Performance of SFSR using (a) MFCC and (b) LPCC features and GMM modeling.

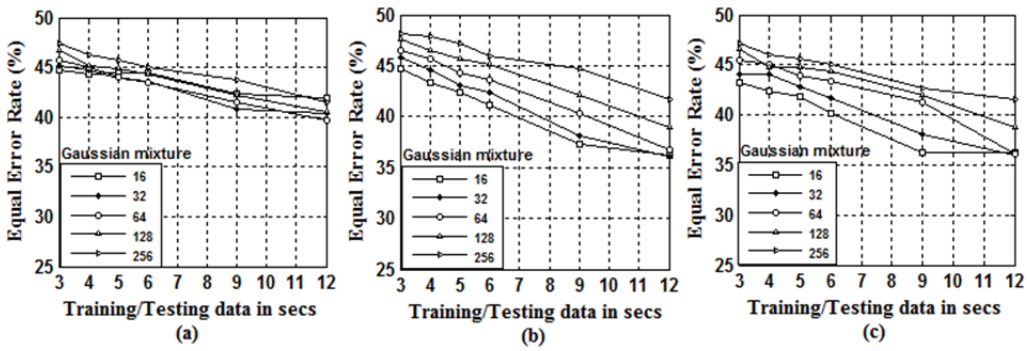

Fig. 3. Performance comparison of speaker verification system using (a) MFCC-MFR (b) MFCC-MFS, and (c) MFCC-MFSR and for GMM modeling.

In the second experiment, MFR along with MFS and MFSR analysis techniques are analyzed with the help of 13 dimensional MFCC features and results are plotted in Fig. 3(a)–(c), respectively.

The modeling is done using GMM. In case of MFCC-MFR, the minimum EER of 39.7% is obtained for 12 seconds data for Gaussian mixtures of 64 compared to other data sizes. The average EER in case of MFCC-MFR is 42.82%. Further, it can be observed that compared to the average EER of MFCC-MFR and MFCC-SFSR, the MFCC-MFR is 1.44% higher than MFCC-SFSR. The experimental results indicate that there is no progress in the performance.

In MFCC-MFS, the minimum EER is 36.04% for Gaussian mixture of 32 for 12 seconds data as compared with remaining data sizes. The average EER is 40.79%. MFCC-MFS performance is better as

compared to MFCC-MFR for all data sizes. Compare to the average EER of MFCC-MFS with MFCC-MFR, MFCC-MFS is having 2.04% lower in EER than MFCC-MFR. This is because the magnitude spectra and feature vectors extracted from speech data for different FS are different due to different frequency resolution [10].

The MFCC-MFSR gives minimum EER of 35.9% is obtained for 12 seconds for Gaussian mixture of 32 compared with remaining data sizes. The average EER is 40.38%. Compared to the average EER of MFCC-MFSR with MFCC-MFR and MFCC-MFS is 2.45% and 0.41% less in EER, respectively. This is because MFSR is the combination of both MFS and MFR. Further, MFSR is having more feature vectors compared to MFS and MSR.

In third experiment, 13 dimension LPCC features are extracted using MFR, MFS, and MFSR analysis techniques and GMM modeling is used to get the speaker models and the results are shown in Fig. 4(a)–(c), respectively. The results show that minimum EER of LPCC-MFR is 27.95% which is obtained for 12 seconds data for Gaussian mixture of 16 compared with all data sizes. The average EER is 37.14%. Further, when we compared EER of both analysis, it was observed that LPCC-MFR is 0.7% lesser than LPCC-SFSR.

In LPCC-MFS analysis, minimum value of EER is 27.64% obtained for 12 seconds train/test data for Gaussian mixture of 16 compared with remaining data sizes. The average EER is 36.33%. Further, it can be observed that the average EER of LPCC-MFS is 0.81% less as compared to LPCC-MFR. Also for all the data sizes LPCC-MFS has lesser EER than LPCC-MFR.

In LPCC-MFSR analysis, there is considerable improvement in the EER as compared to LPCC-MFS and LPCC-MFR. The least EER in case of LPCC-MFSR is 27.5% for Gaussian mixture of 16 for 12 seconds data size. The average EER of LPCC-MFSR is 35.95%. The LPCC-MFSR is 1.19% and 0.38% lesser in average EER as compared with LPCC-MFR and LPCC-MFS, respectively. Other than average EER in LPCC-MFSR, the individual EER for all data sizes are also substantially lesser than that of LPCC-MFR and LPCC-MFS.

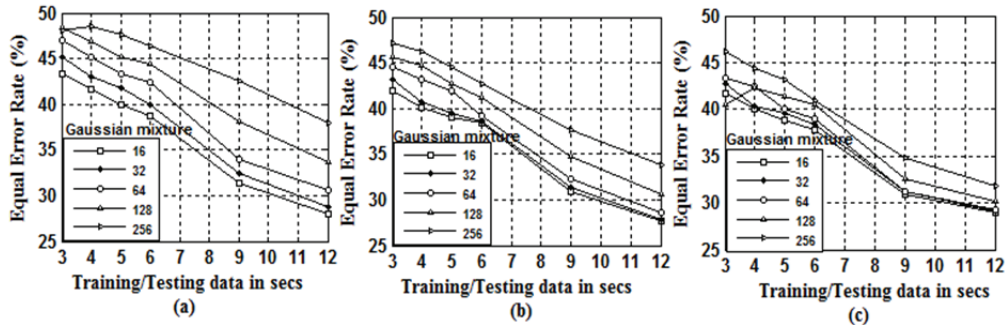

Fig. 4. Performance comparison of speaker verification system using (a) LPCC-MFR, (b) LPCC-MFS, and (c) LPCC-MFSR for GMM modeling.

From these experimental study, we noticed that when train/test both data are small, MFSR analysis technique improves the verification performance as compared to SFSR in both feature extraction methods. Further, LPCC-MFSR provides an average EER of 4.43% less as compared with EER of MFCC-MFSR.

To study the significance of GMM-UBM modeling, we conducted experiments using GMM-UBM

modeling. In GMM-UBM, UBM is usually constructed from large number of speakers' data and UBM is trained using EM algorithm. The speaker dependent model can be created by performing MAP adaptation technique [25]. UBM should contain equal number of male and female speakers. The total duration of male and female speakers is 1,506 seconds each. We also used NIST-2003 database for training the UBM. In this experiment also features are extracted using LPCC and MFCC by considering different speech analysis technique including MFR, MFS and MFSR.

Fig. 5(a) and (b) show the experimental results of SFSR analysis in case of MFCC and LPCC, respectively. The GMM-UBM modeling is used. The minimum EER in case of MFCC-SFSR is 27.28% obtained for 12 seconds data for Gaussian mixtures of 128 compared to remaining data sizes. The average EER of MFCC-SFSR is 35.12%.

In LPCC-SFSR, the least EER is obtained for 12 seconds data for Gaussian mixture of 32 is 26.91% compared with different data sizes. The average EER in case of LPCC-SFSR is 34.19%. The average EER of LPCC-SFSR is minimum as compared with MFCC-SFSR analysis. When we compare EER of LPCC-SFSR with MFCC-SFSR, LPCC-SFSR is 0.93% lower than MFCC-SFSR.

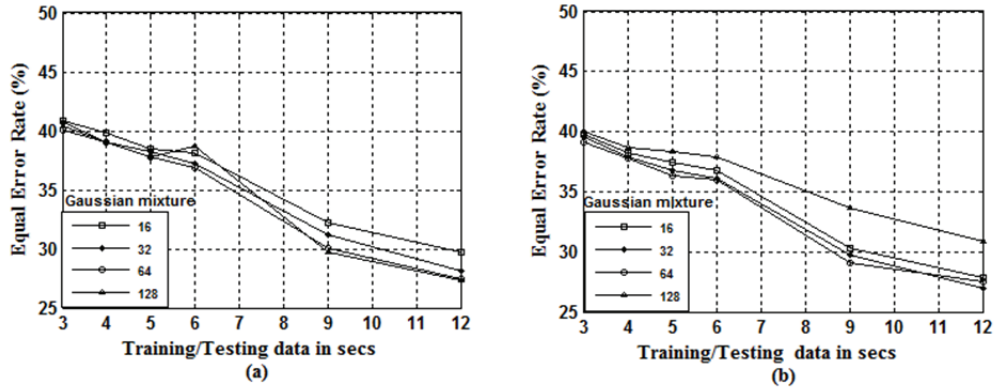

Fig. 5. Performance of SFSR using (a) MFCC and (b) LPCC features and GMM-UBM modeling.

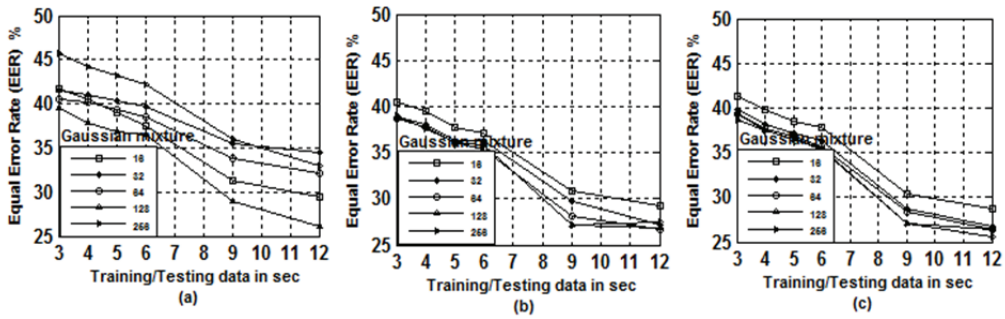

Fig. 6. Performance comparison of speaker verification system using (a) MFCC-MFR, (b) MFCC-MFS, and (c) MFCC-MFSR for GMM-UBM modeling.

The results of analysis techniques MFR, MFS and MFSR are shown in Fig. 6(a)–(c), respectively. The features used are MFCC and modeling technique used is GMM-UBM. In case of MFCC-MFR, the least EER is 26.1% obtained for 12 seconds data for Gaussian mixtures of 128 compared to other data sizes.

The average EER in case of MFCC-MFR is 34.29%. Further, it can be observed that compare the EER of MFCC-MFR and MFCC-SFSR, MFCC-MFR is 0.83% lower than MFCC-SFSR.

In case of MFCC-MFS, the minimum EER is 25.38% for Gaussian mixture of 128 for 12 seconds data as compared to other data sizes. The average EER is 33.58%. It was observed that, MFCC-MFS performance is better as compared to MFCC-MFR for all data sizes. When we the compare the average EER of MFCC-MFS with MFCC-MFR, MFCC-MFS is having 0.71% lesser than MFCC-MFR.

The MFCC-MFSR gives minimum EER of 25.57% is obtained for 12 seconds for Gaussian mixture of 256. The average EER is 33.42%. Average EER of MFCC-MFSR when compared with MFCC-MFR and MFCC-MFS is 0.87% and 0.16% lesser in EER, respectively.

The LPCC features are extracted using MFR, MFS, and MFSR analysis techniques and results are shown in Fig. 7(a)–(c), respectively. The modeling technique used is GMM-UBM. It shows that least EER of LPCC-MFR is 26.12% which is obtained for 12 seconds data for Gaussian mixture of 64 compared with all data sizes. The average EER is 34.02%. Further, the average EER of LPCC-MFR when compared with LPCC-SFSR, the LPCC-MFR average EER is 0.7% which is lesser than LPCC-SFSR.

In case of LPCC-MFS analysis, the least EER of 26.64% obtained for 12 seconds training and testing data for Gaussian mixture of 64 and average EER is 33.48%. Further, the average EER of LPCC-MFS is 0.10% which is less compared with LPCC-MFR. Also for all the data sizes LPCC-MFS has lesser EER than LPCC-MFR.

In case of LPCC-MFSR methods there is considerable improvement in the EER as compared to LPCC-MFS and LPCC-MFR. The least EER in case of LPCC-MFSR is 26.0% for Gaussian mixture of 128 for 12 seconds data size. The average EER of LPCC-MFSR is 33.39%. If we compare the average EER of LPCC-MFSR with LPCC-MFR and LPCC-MFS, LPCC-MFSR is 0.63% and 0.29% lesser in EER as compared with LPCC-MFR and LPCC-MFS, respectively. Other than average reduction in LPCC-MFSR, the individual EER for all the data sizes are also substantially lesser than that of LPCC-MFR and LPCC-MFS.

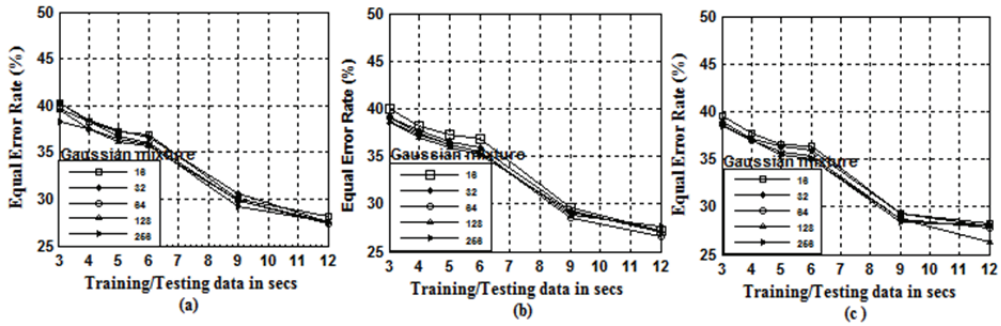

Fig. 7. Performance comparison of speaker verification system using (a) LPCC-MFR, (b) LPCC-MFS, and (c) LPCC-MFSR for GMM-UBM modeling.

Table 1 represents the minimum and average EER of MFCC and LPCC analysis techniques. It was observed in Table 1 that the LPCC analysis gives better performance in terms of minimum and average EER which is less as compared to MFCC analysis by using GMM as a modeling technique. The minimum EER of LPCC-SFSR, LPCC-MFR, LPCC-MFS and LPCC-MFSR is 8.5%, 11.75%, 8.4%, and 8.4% less as compared with MFCC-SFSR, MFCC-MFR, MFCC-MFS, and MFCC-MFSR, respectively

and the average EER of LPCC-SFSR, LPCC-MFR, LPCC-MFS and LPCC-MFSR is 4.18%, 5.62%, 4.46%, and 4.43% less as compared with MFCC-SFSR, MFCC-MFR, MFCC-MFS, and MFCC-MFSR, respectively, using GMM modeling.

Table 1. Comparison of average EER (%) of GMM and GMM-UBM modeling for 13 dimensions feature using SFSR and MFSR analysis techniques

Speech analysis	GMM		GMM-UBM	
	Min. EER (%)	Avg. EER (%)	Min. EER (%)	Avg. EER (%)
MFCC-SFSR	36.68	41.39	27.28	35.12
LPCC-SFSR	28.18	37.21	26.10	34.19
MFCC-MFR	39.70	42.83	26.12	34.29
LPCC-MFR	27.95	37.14	26.10	34.02
MFCC-MFS	36.04	40.79	25.38	33.58
LPCC-MFS	27.64	36.33	26.64	33.48
MFCC-MFSR	35.90	40.38	25.57	33.42
LPCC-MFSR	27.50	35.95	26.01	33.39

Another interesting point observed in Figs. 6 and 7 for GMM-UBM modeling is that, LPCC based MFR, MFS, and MFSR are having lesser EER as compared with MFCC based MFR, MFS, and MFSR for 3, 4, 5, and 6 seconds data. Further, if the train/test speech data are increased to 9 and 12 seconds, MFCC based MFR, MFS, and MFSR are having minimum EER compared to LPCC based MFR, MFS, and MFSR. From this experiment, we observed that when both training and testing data (3–6 seconds) are limited, LPCC performance is better as compared with MFCC. This is because LPCC is able to apprehend more information from speech data this will make a distinction between different speakers [32]. If we increase training and testing data (above 6 seconds), MFCC based feature extraction analysis improves the performance compared to LPCC based feature extraction analysis. The minimum EER of LPCC-SFSR and LPCC-MFR is 1.18%, 0.02% less as compared to MFCC-SFSR and MFCC-MFR, respectively. Further, in case of MFCC-MFR and MFCC-MFSR is having 1.26% and 0.44% less in EER as compared to LPCC-MFS and LPCC-MFSR, respectively. The average EER of LPCC-SFSR, LPCC-MFR, LPCC-MFS, and LPCC-MFSR is 0.93%, 0.27%, 0.10%, and 0.03% less as compared with MFCC-SFSR, MFCC-MFR, MFCC-MFS, and MFCC-MFSR, respectively.

To justify the above statement, 39 dimensional MFCC and LPCC features are extracted for different analysis techniques and modeled using GMM and GMM-UBM.

These feature vectors contain both static and transitional characteristics of the speaker-specific information [28]. The Δ and ΔΔ coefficients are calculated by capturing the transitional characteristics.

Fig. 8(a) and (b) show the experimental results of SFSR analysis using MFCC and LPCC features, respectively. The modeling technique used is GMM. The minimum EER in case of MFCC-SFSR is 30.35% is obtained for 12 seconds data for Gaussian mixtures of 16 compared to other data sizes. The average EER of MFCC-SFSR is 39.52%.

In case of LPCC-SFSR, the least EER is obtained for 12 seconds data for Gaussian mixture of 32 is 29.72% compared with other data sizes. The average EER in case of LPCC-SFSR is 37.95%. The average EER of LPCC-SFSR is minimum as compared with MFCC-SFSR analysis in case of GMM modeling. If we compare the EER of both analysis, LPCC-SFSR is 1.57% lower than MFCC-SFSR.

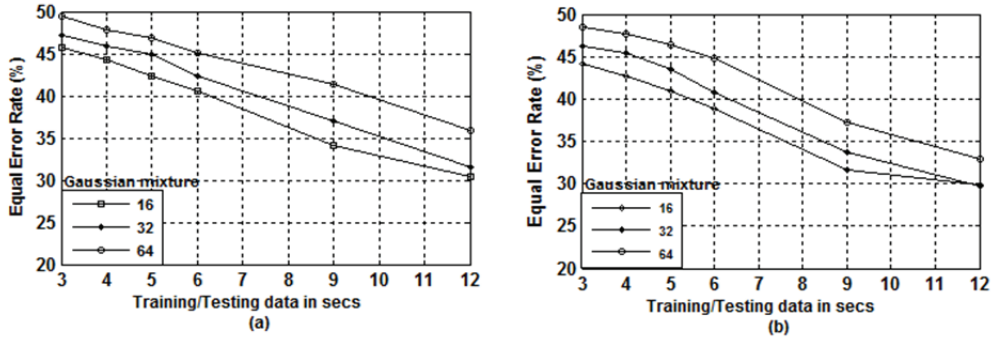

Fig. 8. Performance of SFSR using (a) ΔΔMFCC and (b) ΔΔLPCC features and GMM modeling.

The results of the analysis techniques MFR, MFS, and MFSR are shown in Fig. 9(a)–(c), respectively. The features used are MFCC and modeling is done using GMM. In MFCC-MFR, the least EER is 30.30% obtained for 12 seconds data for Gaussian mixtures of 16 compared to remaining data sizes. The average EER in case of MFCC-MFR is 39.39%. Further, the EER of MFCC-MFR when compared with MFCC-SFSR, MFCC-MFR is 0.13% lower than MFCC-SFSR.

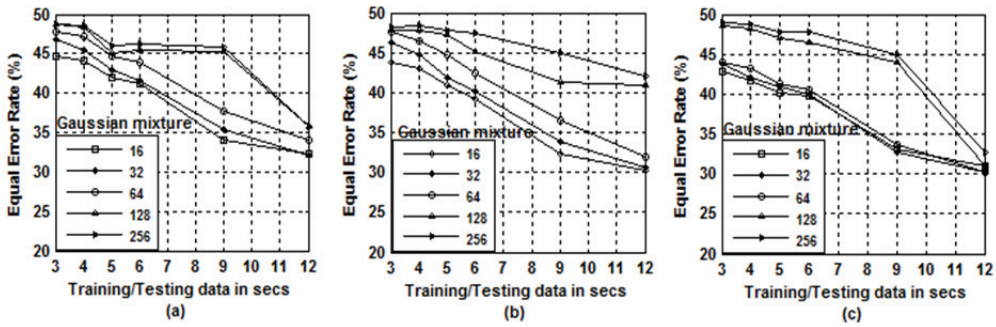

Fig. 9. Performance comparison of speaker verification system for ΔΔ using (a) MFCC-MFR, (b) MFCC-MFS, and (c) MFCC-MFSR for GMM modeling.

In MFCC-MFS, the minimum EER is 30.26% for Gaussian mixture of 16 for 12 seconds data as compared to other data sizes. The average EER is 38.25%. It was observed that MFCC-MFS performance is better as compared to MFCC-MFR for all data sizes. When we compared the average EER of MFCC-MFS with MFCC-MFR, MFCC-MFS is having 1.14% lesser EER than MFCC-MFR.

The MFCC-MFSR gives minimum EER of 30.26% which obtained for 12 seconds for Gaussian mixture of 32. The average EER is 37.89%. When the average EER of MFCC-MFSR is compared with MFCC-MFR and MFCC-MFS is 1.5% and 0.36% less in EER, respectively.

The LPCC features are extracted using MFR, MFS, and MFSR analysis techniques and results are shown in Fig. 10(a)–(c), respectively and modeling used is GMM. The result shows that least EER of LPCC-MFR is 30.08% which is obtained for 12 seconds data for Gaussian mixture of 32 compared with all data sizes. The average EER is 37.79%. Further, compare the average EER of LPCC-MFR with LPCC-SFSR, the LPCC-MFR average EER is 0.16% lesser than LPCC-SFSR.

In case of LPCC-MFS analysis, the least EER of 29.31% obtained for 12 seconds training and testing

data for Gaussian mixture of 16. The average EER is 37.49%. Further, the average EER of LPCC-MFS when compared with LPCC-MFR, LPCC-MFS average EER is 0.30% less than LPCC-MFR. Also for all the data sizes LPCC-MFS has lesser EER than LPCC-MFR.

In case of LPCC-MFSR analysis there is considerable improvement in the EER as compared to LPCC-MFS and LPCC-MFR. The least EER in case of LPCC-MFSR is 29.44% for Gaussian mixture of 64 for 12 seconds data size. The average EER of LPCC-MFSR is 36.81%. If we compared the average EER of LPCC-MFSR with LPCC-MFR and LPCC-MFS, LPCC-MFSR is 0.98% and 0.68% lesser in EER as compared with LPCC-MFR and LPCC-MFS, respectively. Other than average reduction in LPCC-MFSR, the individual EER for all the data sizes are also considerably lesser than that the LPCC-MFR and LPCC-MFS.

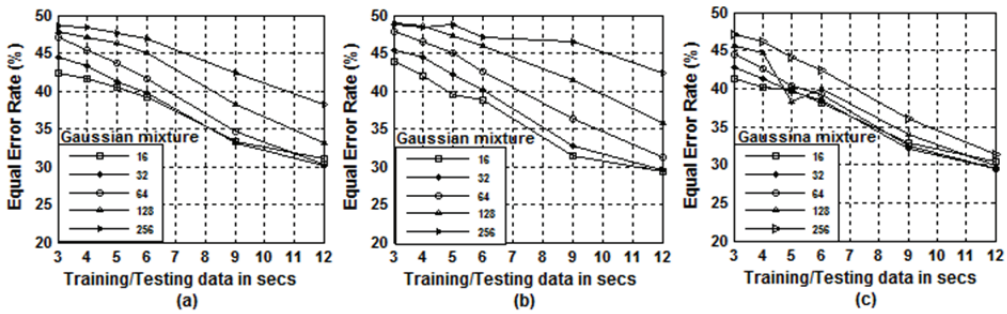

Fig. 10. Performance comparison of speaker verification system for ΔΔ using (a) LPCC-MFR, (b) LPCC-MFS, and (c) LPCC-MFSR for GMM modeling.

Fig. 11(a) and (b) shows the experimental results of SFSR analysis in case of MFCC and LPCC, respectively and modeling technique used is GMM-UBM. The minimum EER in case of MFCC-SFSR is 24.48% is obtained for 12 seconds data for Gaussian mixtures of 128 compared to other data sizes. The average EER of MFCC-SFSR is 33.58%.

In case of LPCC-SFSR, the least EER is obtained for 12 seconds data for Gaussian mixture of 64 is 23.71% compared with remaining data sizes. The average EER in case of LPCC-SFSR is 32.72%. The average EER of LPCC-SFSR is minimum as compared with MFCC-SFSR analysis. If we compare the EER of both analysis, LPCC-SFSR is 0.86% lower than MFCC-SFSR.

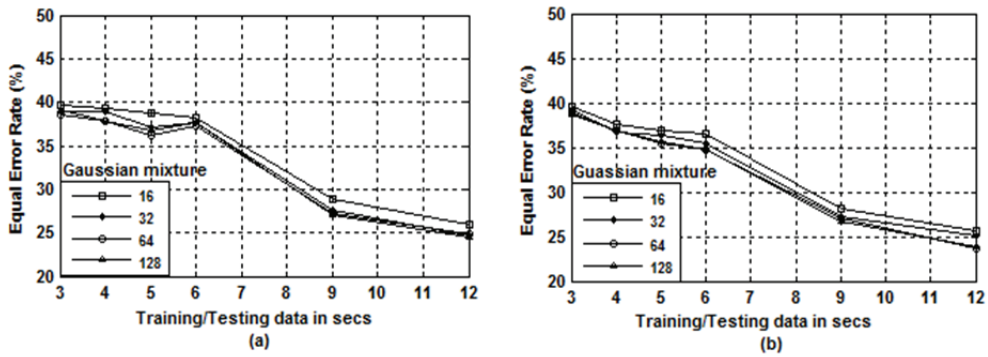

Fig. 11. Performance of SFSR using (a) ΔΔMFCC and (b) ΔΔLPCC features and GMM-UBM modeling.

The results of the analysis techniques MFR, MFS, and MFSR are shown in Fig. 12(a)–(c), respectively. The features used are MFCC and modeling technique used is GMM-UBM. In case of MFCC-MFR, the least EER is 22.4% is obtained for 12 seconds data for Gaussian mixtures of 128 compared to other data sizes. The average EER in case of MFCC-MFR is 32.66%. Further, if we compare the average EER of MFCC-MFR with MFCC-SFSR, MFCC-MFR is 0.06% lower than MFCC-SFSR.

In MFCC-MFS, the minimum EER is 23.25% for Gaussian mixture of 128 for 12 seconds data as compared to remaining data sizes. The average EER is 32.13%. Performance of MFCC-MFS is better as compared to MFCC-MFR for all data sizes. When comparing the average EER of MFCC-MFS with MFCC-MFR, MFCC-MFS is having 0.53% lower in EER than MFCC-MFR.

The MFCC-MFSR gives minimum EER of 22% obtained for 12 seconds for Gaussian mixture of 128. The average EER is 31.83%. While comparing the average EER of MFCC-MFSR with MFCC-MFR and MFCC-MFS is 0.83% and 0.3% less in EER, respectively.

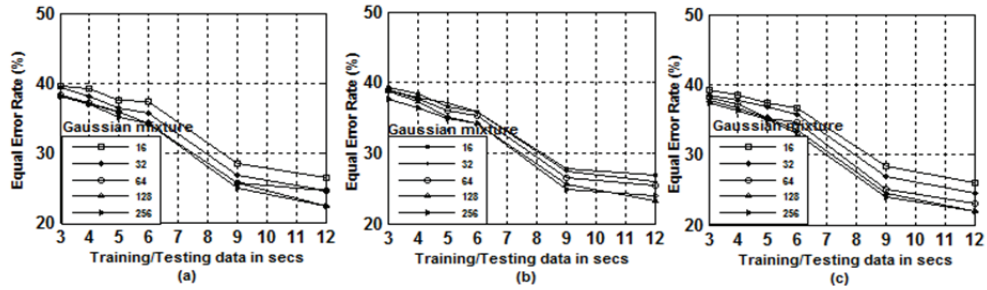

Fig. 12. Performance comparison of speaker verification system for ΔΔ using (a) MFCC-MFR, (b) MFCC-MFS, and (c) MFCC-MFSR for GMM-UBM modeling.

Fig. 13(a)–(c) shows performance analysis of MFR, MFS, and MFSR respectively using LPCC features and modeling is done using GMM-UBM. It shows that least EER of LPCC-MFR is 23.69% which is obtained for 12 seconds data for Gaussian mixture of 64 compared with all data sizes. The average EER is 31.91%. Further, while comparing the average EER of LPCC-MFR with LPCC-SFSR, the LPCC-MFR average EER is 0.81% lesser than LPCC-SFSR.

In LPCC-MFS analysis, the least EER of 23.7% obtained for 12 seconds training and testing data for Gaussian mixture of 64 compared with all remaining data sizes. The average EER is 31.7%. Further, the average EER of LPCC-MFS when compared with LPCC-MFR, LPCC-MFS average EER is 0.21% less than LPCC-MFR. Also for all the data sizes LPCC-MFS has lesser EER than LPCC-MFR.

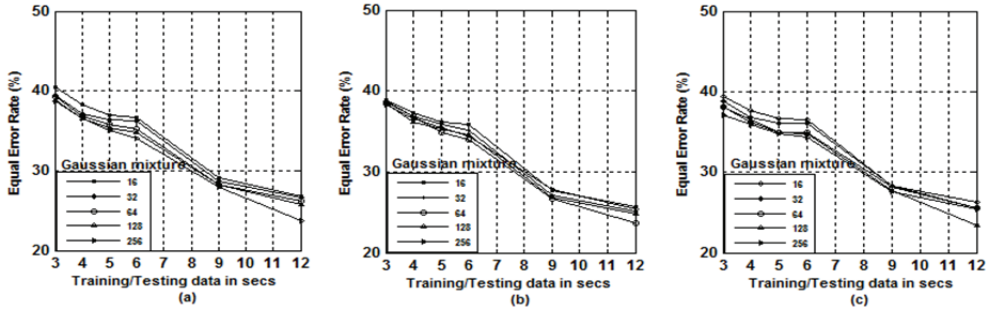

Fig. 13. Performance comparison of speaker verification system for ΔΔ using (a) LPCC-MFR, (b) LPCC-MFS, and (c) LPCC-MFSR for GMM-UBM modeling.

In LPCC-MFSR analysis there is considerable improvement in the EER as compared to LPCC-MFS and LPCC-MFR. The least EER in case of LPCC-MFSR is 23.33% for Gaussian mixture of 128 when compared with 12 seconds data size. The average EER of LPCC-MFSR is 31.36%. If we compare the average EER of LPCC-MFSR with LPCC-MFR and LPCC-MFS, LPCC-MFSR is 0.55% and 0.34% lesser in EER as compared with LPCC-MFR and LPCC-MFS, respectively. Other than average EER in LPCC-MFSR, the individual EER for all the data sizes are also considerably lesser than that of the LPCC-MFR and LPCC-MFS.

It was observed that, in this experiment also, ΔΔLPCC based MFR, MFS, and MFSR will be having lower EER as compared with MFCC based MFR, MFS and MFSR for 3, 4, 5, and 6 seconds data. If we increase the train/test speech data to 9 and 12 seconds, MFCC based MFR, MFS, and MFSR will have minimum EER compared to LPCC based MFR, MFS, and MFSR. Further, we observed that when both train/test data are limited (3–6 sec), LPCC performance is better as compared with MFCC.

It was observed in the Table 2, even in case of 39 dimension LPCC based analysis techniques will have lesser average EER when compared with MFCC based analysis techniques in case of GMM modeling, but in case of GMM-UBM, MFCC based analysis technique have lesser EER compared to LPCC based techniques. Further, the minimum EER of LPCC-SFSR, LPCC-MFR, LPCC-MFS, and LPCC-MFSR is 0.63%, 0.22%, 0.95%, 0.82% less compared with MFCC-SFSR, MFCC-MFR, MFCC-MFS, and MFCC-MFSR, respectively and the average EER of LPCC-SFSR, LPCC-MFR, LPCC-MFS and LPCC-MFSR is 1.57%, 1.6%, 0.76%, and 1.08% less as compared with MFCC-SFSR, MFCC- MFR, MFCC-MFS, and MFCC-MFSR, respectively as modeling done using GMM. Further, the minimum EER of LPCC-SFSR is 0.77% less as compared with MFCC-MFSR, but in other cases MFCC-MFR, MFCC-MFS and MFCC-MFSR is having minimum EER of 1.29%, 0.45%, and 1.33% less as compared with EER of LPCC-MFR, LPCC-MFS, and LPCC-MFSR, respectively. The average EER of LPCC-SFSR, LPCC-MFR, LPCC-MFS and LPCC-MFSR is 0.86%, 0.75%, 0.43%, and 0.47% less as compared with MFCC-SFSR, MFCC-MFR, MFCC-MFS, and MFCC-MFSR, respectively.

Table 2. Comparison of average EER (%) of GMM and GMM-UBM modeling for 39 dimensions feature using SFSR and MFSR analysis techniques

Speech analysis	GMM		GMM-UBM	
	Min. EER (%)	Avg. EER (%)	Min. EER (%)	Avg. EER (%)
MFCC-SFSR	30.35	39.52	24.48	33.58
LPCC-SFSR	29.72	37.95	23.71	32.72
MFCC-MFR	30.30	39.39	22.40	32.66
LPCC-MFR	30.08	37.79	23.69	31.91
MFCC-MFS	30.26	38.25	23.25	32.13
LPCC-MFS	29.31	37.49	23.70	31.70
MFCC-MFSR	30.26	37.89	22.00	31.83
LPCC-MFSR	29.44	36.81	23.33	31.36

5. Conclusions

In this paper, we verified the significance of MFR, MFS, and MFSR techniques for speaker verification under the limited data condition. Initially, we analyzed the importance of feature vectors

role in SFSR, MFR, MFS, and MFSR methods. Then, we experimentally analyzed to verify the performance for different conditions. The experimental results show that in both feature extraction methods SFSR was unable to capture the more speaker-specific information. The more speaker-specific features are extracted using MFR, MFS, and MFSR methods. The experimental results indicate that, speaker verification performance EER can be moderately improved by adapting appropriate analysis technique. Further, from the experiment results we observed that MFSR gives better verification performance compared with SFSR and other analysis technique. In case of GMM modeling for all data sizes LPCC-MFSR improves the performance of EER as compared to MFCC-MFSR.

We observed that, LPCC based MFR, MFS, and MFSR analysis yields lower EER compared to MFCC based analysis techniques by using GMM-UBM modeling technique. Further, we observed that, if training and testing data are increased, MFCC based analysis techniques improve the performance over LPCC analysis techniques. To verify the significance of various analysis techniques, different feature extraction and modeling technique need to be explored.

References

[1] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4-20, 2004.

[2] S. Dey, S. Barman, R. K. Bhukya, R. K. Das, B. C. Haris, S. R. M. Prasanna, and R. Sinha, "Speech biometric based attendance system," in *Proceedings of 2014 Twentieth National Conference on Communications (NCC),* Kanpur, India, 2014, pp. 1-6.

[3] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12-40, 2010.

[4] G. Pradhan and S. M. Prasanna, "Speaker verification under degraded condition: a perceptual study," *International Journal of Speech Technology*, vol. 14, no. 4, pp. 405-417, 2011.

[5] A. E. Rosenberg, "Automatic speaker verification: a review" *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475-487, 1976.

[6] A. Neustein and H. A. Patil, *Forensic Speaker Recognition*. Heidelberg: Springer, 2012.

[7] H. S. Jayanna and S. M. Prasanna, "Analysis, feature extraction, modeling and testing techniques for speaker recognition," *IETE Technical Review*, vol. 26, no. 3, pp. 181-190, 2009.

[8] H. S. Jayanna, "Limited data speaker recognition," Ph.D. dissertation, Indian Institute of Technology Guwahati, India, 2009.

[9] D. Pati and S. M. Prasanna, "Subsegmental, segmental and suprasegmental processing of linear prediction residual for speaker information," *International Journal of Speech Technology*, vol. 14, no. 1, pp. 49-64, 2011.

[10] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[11] S. M. Prasanna, C. G. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, no. 10, pp. 1243-1261, 2006.

[12] B. Yegnanarayana, S. M. Prasanna, J. M. Zachariah, and C. S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 575-582, 2005.

[13] F. Farahani, P. G. Georgiou, and S. S. Narayanan, "Speaker identification using supra-segmental pitch pattern dynamics," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, 2004, pp. 89-92.

[14] A. V. Jadhav and R. V. Pawar, "Review of various approaches towards speech recognition," in *Proceedings of 2012 International Conference on Biomedical Engineering (ICoBE),* Penang, Malaysia, 2012, pp. 99-103.

[15] H. S. Jayanna and S. M. Prasanna, "Multiple frame size and rate analysis for speaker recognition under limited data condition," *IET Signal Processing*, vol. 3, no. 3, pp. 189-204, 2009.

[16] G. L. Sarada, T. Nagarajan, and H. A. Murthy, "Multiple frame size and multiple frame rate feature extraction for speech recognition," in *Proceedings of 2004 International Conference on Signal Processing and Communications*, Bangalore, India, 2004, pp. 592-595.

[17] K. Samudravijaya, "Variable frame size analysis for speech recognition," in *Proceedings of the International Conference on Natural Language Processing,* Hyderabad, India, 2004.

[18] Q. Zhu and A. Alwan, "On the use of variable frame rate analysis in speech recognition," in *Proceedings of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing,* Istanbul, Turkey, 2000, pp. 1783-1786.

[19] P. Le Cerf and D. Van Compernolle, "A new variable frame analysis method for speech recognition," *IEEE Signal Processing Letters*, vol. 1, no. 12, pp. 185-187, 1994.

[20] R. Pawar and H. Kulkarni, "Analysis of FFSR, VFSR, MFSR techniques for feature extraction in speaker recognition: a review," *International Journal of Computer Science*, vol. 7, no. 4, pp. 26-31, 2010.

[21] T. Nagarajan, "Implicit systems for spoken language identification," Ph.D. dissertation, Indian Institute of Technology Madras, India, 2004.

[22] G. S. Ghadiyaram, N. H. Nagarajan, T. N. Thangavelu, and H. A. Murthy, "Automatic transcription of continuous speech using unsupervised and incremental training," in *Proceedings of the 8th International Conference on Spoken Language Processing*, Jeju Island, Korea, 2004.

[23] National Institute of Standards and Technology, "The NIST Year 2003 speaker recognition evaluation plan," 2013 [Online]. Available: https://www.nist.gov/sites/default/files/documents/2017/09/26/2003-spkrec-evalplan-v2.2.pdf

[24] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1085-1095, 2012.

[25] A. Salman, E. Muhammad, and K. Khurshid, "Speaker verification using boosted cepstral features with Gaussian distributions," in *Proceedings of IEEE International Multitopic Conference*, Lahore, Pakistan, 2007, pp. 1-5.

[26] D. Pat and S. M. Prasanna, "Processing of linear prediction residual in spectral and cepstral domains for speaker information," *International Journal of Speech Technology*, vol. 18, no. 3, pp. 333-350, 2015.

[27] W. C. Hsu, W. H. Lai, and W. P. Hong, "Usefulness of residual-based features in speaker verification and their combination way with linear prediction coefficients," in *Proceedings of the 9th IEEE International Symposium on Multimedia Workshops,* Beijing, China, 2007, pp. 246-251.

[28] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 342-350, 1981.

[29] V. Prakash and J. H. L. Hansen, "In-set/out-of-set speaker recognition under sparse enrollment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2044-2052, 2007.

[30] T. Hasan and J. H. Hansen, "A study on universal background model training in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1890-1899, 2011.

[31] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788-798, 2011.

[32] E. Wong and S. Sridharan, "Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification," in *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing,* Hong Kong, China, 2001, pp. 95-98.

Thyamagondlu Renukamurthy Jayanthi Kumari https://orcid.org/0000-0003-0020-4655

She received the B.E. degree from Bangalore University in 1997 and M.Tech. degree from Visvesvaraya Technological University in 2006. She is currently pursuing the Ph.D. degree at the Visvesvaraya Technological University, Karnataka, India. She has been working as a Researcher in the Department of Electronics and Communication Engineering, Siddaganga Institute of Technology, Karnataka, India. Her research interests include speech, speaker verification under limited data.

Haradagere Siddaramaiah Jayanna https://orcid.org/0000-0002-4342-9339

He received the B.E. and M.E. degrees from Bangalore University in 1992 and 1995, respectively, and Ph.D. from prestigious Indian Institute of Technology, Guwahati, India, in 2009. He has published a number of papers in various national and international journals and conferences apart from guiding a number of UG, PG and research scholars. Currently, he is working as a Professor in the Department of Information Science and Engineering, Siddaganga Institute of Technology, Karnataka, India. His research interests are in the areas of speech, limited data speaker recognition, image processing, computer networks and computer architecture.