International Journal of Electrical and Computer Engineering (IJECE)
Vol. 7, No. 6, December 2017, pp. 3344 – 3357
ISSN: 2088-8708, DOI: 10.11591/ijece.v7i6.pp3344-3357

Limited Data Speaker Verification: Fusion of Features

T. R. Jayanthi Kumari[1] and H. S. Jayanna[2]
[1]Department of Electronics and Communication Engineering, Siddaganga Institute of Technology, Karnataka, India
[2]Department of Information Science and Engineering, Siddaganga Institute of Technology, Karnataka, India

Article Info

Article history:
Received: Mar 29, 2017
Revised: Jul 18, 2017
Accepted: Aug 3, 2017

Keyword:
MFCC
LPCC
LPR
LPRP
GMM
GMM-UBM

ABSTRACT

The present work demonstrates experimental evaluation of speaker verification for different speech feature extraction techniques with the constraints of limited data (less than 15 seconds). The state-of-the-art speaker verification techniques provide good performance for sufficient data (greater than 1 minutes). It is a challenging task to develop techniques which perform well for speaker verification under limited data condition. In this work different features like Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC), Delta (△), Delta-Delta (△△), Linear Prediction Residual (LPR) and Linear Prediction Residual Phase (LPRP) are considered. The performance of individual features is studied and for better verification performance, combination of these features is attempted. A comparative study is made between Gaussian mixture model (GMM) and GMM-universal background model (GMM-UBM) through experimental evaluation. The experiments are conducted using NIST-2003 database. The experimental results show that, the combination of features provides better performance compared to the individual features. Further GMM-UBM modeling gives reduced equal error rate (EER) as compared to GMM.

Corresponding Author:
T. R. Jayanthi Kumari
Department of Electronics and Communication Engineering
Siddaganga Institute of Technology
India, Karnataka, Bengaluru-560077
Email: trjayanthikumari@gmail.com

1. INTRODUCTION

Speech signals play a main role in communication media to understand the conversation between the people [1]. The speaker recognition is a technique to recognize a speaker using his/her original speech voice and can be used for either speaker verification or speaker identification [2]. Over the last decade, speaker verification has been used for many commercial applications and these applications prefer limited data conditions. Further, limited data indicates speech data of few seconds (less than 15 sec). Based on the nature of training and test speech data, text-dependent and text-independent [3] are two classification of speaker verification. In text-dependent mode, speaker training and testing data remains same and in case of text-independent, training and testing speech data are different. Text-independent speaker verification under limited data conditions has always been a challenging task.

The speaker verification system contains four stages, namely analysis of speech data, extraction of features, modeling and testing [4]. The analysis stage analyzes the speaker information using vocal tract [5], excitation source [6] and suprasegmental features like duration, accent and modulation [7]. The amount of data available in limited data condition is very small which gives poor verification performance. To improve the verification performance in limited data condition, we need different levels of information to be extracted from speech data and they have to be combined to good verification performance. The vocal tract and excitation source information are combined in the present study for improving the performance of speaker verification system under limited data condition.

Second stage of speaker verification is feature extraction. Speech production system usually generates

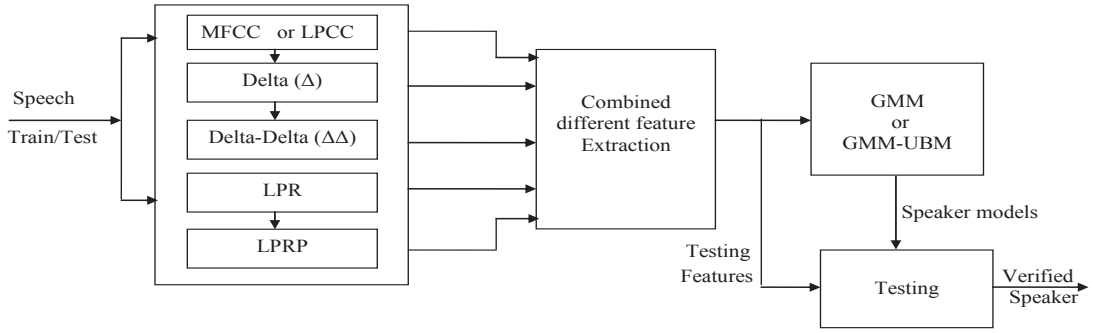

Figure 1. Block diagram of combination of different features for speaker verification system.

large amount of data which include sensor, channel, language, style etc [8]. The purpose of feature extraction is to extract feature vectors of reduced dimension. The extracted feature information are emphasized and other redundant factors are suppressed in these feature vectors [3][9]. The vocal tract information can be extracted using Mel-frequency cepstral coefficients (MFCC) [10] and Linear prediction cepstral coefficients (LPCC) [11] extraction methods. The speech signal contains both static and dynamic characteristics. The MFCC and LPCC feature set contain only static characteristics. The dynamic characteristics represented by Delta (△) and Delta-Delta (△△) contains some more speaker information, which are useful in speaker verification [4]. Excitation source features are extracted using Linear prediction residual (LPR) and Linear prediction residual phase (LPRP) [12].

In this work, LPR, LPRP, MFCC, △MFCC, △△MFCC, LPCC, △LPCC and △△LPCC features are used to evaluate the performance of the system under limited data condition. Further, each of these features offer different information and a combination of these may improve the performance of speaker verification. Hence, the performance of speaker verification considering combination of features is evaluated in the present work. Fig. 1. shows the block diagram representation of combination of different features for speaker verification system.

The paper has been organized into the following sections: Section 2 describes the speaker verification studies using different feature extraction techniques. Different modelling techniques and testing are presented in Section 3. Experimental results are reported in Section 4. Section 5 contains conclusion and future scope of work.

2. SPEAKER VERIFICATION STUDIES USING DIFFERENT FEATURES

The speaker-specific information can be extracted from feature extraction techniques at a reduced data rate[13]. These feature vectors contain vocal tract, excitation source and behavioral traits of speaker-specific information[4]. A good feature is one which contains all components of speaker-specific information. To create a good feature set, different feature extraction techniques need to be understood.

2.1. Vocal tract features for speaker verification

The vocal tract features are extracted using MFCC and LPCC feature extraction techniques. The features extracted from these techniques are different and therefore their performance varies. The reason for the same is as follows.

In case of MFCC, the spectral distortion is minimized using hamming window. The magnitude frequency response is obtained by applying Fourier Transformation to the windowed frame signal. The 22 triangular band pass filters are used to pass the resulting spectrum. Discrete cosine transform is applied to the output of the mel filters in order to obtain the cepstral coefficients. The obtained MFCC features are used to train and test speech data.

LPCC reflects the differences of the biological structure of human vocal tract. Computing method by LPCC is a recursion from LPC parameter to LPC cepstrum according to all-pole model. LPC is simply

the coefficients of this all-pole filter and is equivalent to the smoothened envelope of the log spectrum of the speech. LPC can be calculated either by the autocorrelation or covariance methods directly from the windowed portion of speech. The Durbin's recursive method is used to calculate LPCC without using the Discrete Fourier Transform (DFT) and the inverse DFT. These two methods are more complex and time consuming [14].

The MFCC and LPCC extraction techniques are widely used and have proven to be effective in speaker verification. However, they are not providing satisfactory performance under limited data condition. Therefore, there is a need to improve the performance of speaker verification system by obtaining extra information about the speech data. The feature set of MFCC and LPCC contains only static properties of speech signal. In addition, the dynamic characteristics of the speech signal can also be obtained to improve the performance of speaker verification. This will be helpful for verification of speakers[15]. Two types of dynamics are available in speech processing [16] :

- The velocity of the features which is known as $\triangle$ features obtained by average first-order temporal derivative.
- The acceleration of the features which is known as $\triangle\triangle$ features obtained by average second order temporal derivative.

2.2. Excitation source features for speaker verification

The spectral features extracted from vocal tract are in the range of 10-30 ms. These spectral features ignore some of the speaker specific excitation information like linear prediction (LP) residual and LP residual phase that can be used for speaker verification [6]. In order to calculate LP residual, first the vocal tract information is predicted from speech data using LP analysis and inverse filter formation is used to suppress them from the speech data [17][6]. To calculate LPRP, first we need to divide LP residual by its Hilbert envelop [17]. The LPRP contains speaker-specific information and LPR contains information obtained from excitation source mainly glottal closure instants (GCIS) [18]. The features of LPR and LPRP contain speaker-specific excitation source information, which are dissimilar in their characteristics. These two features can be combined to gain more advantage.

3. SPEAKER MODELING AND TESTING

Different modelling techniques are available for speaker modelling including Vector quantization (VQ), Hidden markov model (HMM), Gaussian mixture model (GMM) and GMM-Universal background model (UBM) etc. Among these GMM and GMM-UBM are used as a classifier for the present work. When the available training data is inadequate, the GMM-UBM is widely used for speaker verification [19]. UBM represents the speaker independent distribution of features. To construct UBM, we require large amount of speech data. UBM is the core part of GMM-UBM speaker verification system. A balance of male and female speakers must be ensured in UBM. The simplest approach to train a UBM is to pool all the data and use it via expectation-maximization (EM) algorithm [20]. The coupled target and background speaker model components are integrated effectively while performing speaker recognition, when Maximum a posteriori (MAP) adaptation is used [13].

The advantage of UBM model is that a large number of speakers are used to design speaker independent model and trained for the required task. Even with minimal speaker data, UBM-based modeling technique provides good performance. The drawback of UBM model is that a large gender-balanced speaker set is required for training [20]. The speakers are also modelled using GMM to verify its effectiveness under limited data speaker verification.

In case of testing, the reference models are compared by test feature vectors, if the test feature vectors are matches with the reference models scores is generated. The scores represent how well the test feature vectors match with reference models [4]. In practical applications, there will be chance of rejecting true speakers and chance of accepting false speakers. In the present work the log likelihood ratio test method [21] is adopted.

4. RESULTS AND DISCUSSIONS

4.1. Experimental setup

In current analysis, the NIST-2003 database is used for verifying the speakers [22]. This contains 356 train and 2559 test speakers. The train speaker contains 149 male and 207 female speakers. The UBM contains

251 female and male speakers. The duration of test, train and UBM speech varies from seconds to few minutes. The present work is for limited data, therefore we have taken each speakers is of durations 3s-3s (train-test), 4s-4s, 5s-5s, 6s-6s, 9s-9s and 12s-12s data to create the database for the study.

4.2. Speaker verification results

We conducted the text-independent speaker verification experiments. The verification performance of the system can be calculated by using equal error rate (EER). It is the ratio of false rejection rate (FRR) and false acceptance rate (FAR). The extracted features are MFCC, LPCC, LPR, LPRP and transitional characteristics like $\triangle$ and $\triangle\triangle$ are in the dimension of 13. In case of MFCC and LPCC and its derivatives, speech data is analyzed with the frame size (FS) of 20 ms and with frame rate (FR) of 10 ms. In case of LPCC, we considered 10^{th} order LP analysis because speech is sampled at 8 KHz. The LP order varies from 8 to 12 [23] and 10^{th} order shown to be appropriate to compute LPCC [23]. FS of 12 ms and FR 6 ms has been fixed for LPR and LPRP. The speaker specific information obtained for each of these features are different. Therefore the combination of these features may give better performance. The modeling techniques used are GMM and GMM-UBM. The speakers are modelled for Gaussian mixture of 16, 32, 64, 128 and 256.

4.3. Individual feature performance using GMM and GMM-UBM

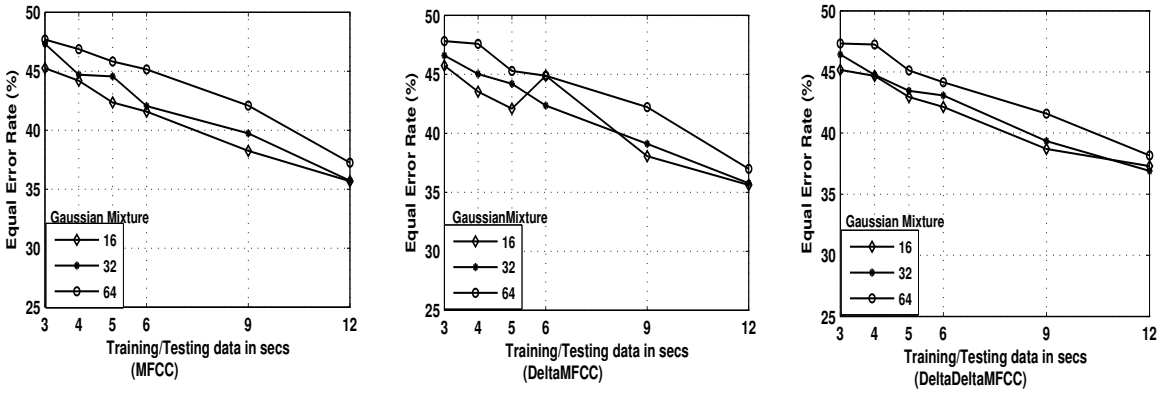

Figure 2. Performance of speaker verification system based on MFCC individual features using GMM modeling

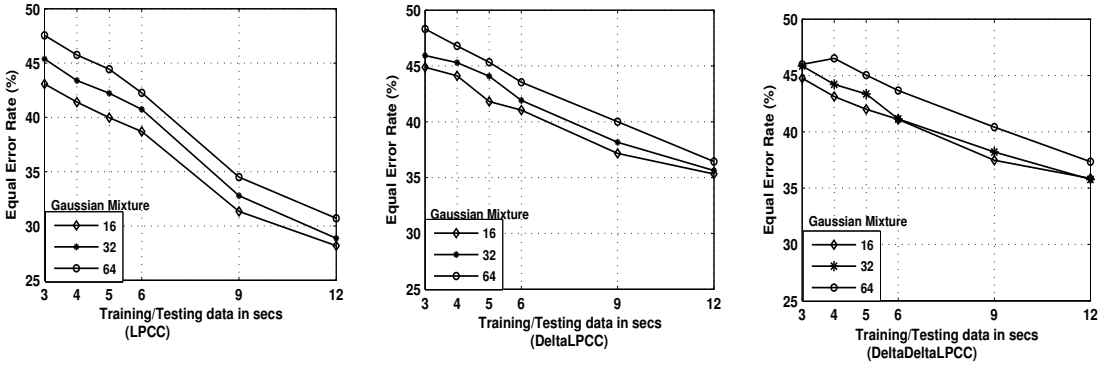

Figure 3. Performance of speaker verification system based on LPCC individual features using GMM modeling

The experimental results are shown in Figure. 2, 3 and 4 for individual features of (MFCC, $\triangle$MFCC, $\triangle\triangle$MFCC), (LPCC, $\triangle$LPCC, $\triangle\triangle$LPCC) and (LPR, LPRP) respectively. The experiment is conducted for 3s-

Table 1. Comparison of minimum EER(%) for individual features using different amount of training and testing data for GMM

Individual Features	Training/Testing data					
	3s-3s	4s-4s	5s-5s	6s-6s	9s-9s	12s-12s
MFCC	45.16	44.21	42.36	41.89	38.25	35.68
△MFCC	45.75	43.54	42.09	44.89	38.07	35.63
△△MFCC	45.27	44.67	42.95	42.14	38.70	37.30
LPCC	43.08	41.41	39.97	38.7	31.34	28.18
△LPCC	44.89	44.12	41.82	41.05	37.17	35.32
△△LPCC	44.76	43.13	42.00	41.10	37.48	35.86
LPR	47.85	48.34	47.34	47.06	46.59	46.09
LPRP	47.16	47.43	46.08	46.58	46.66	46.62

3s, 4s-4s, 5s-5s, 6s-6s, 9s-9s and 12s-12s for different Gaussian mixtures. Further, the modeling is done using GMM for Gaussian mixtures of 16, 32 and 64. Since the data is very small, the Gaussian mixtures are limited to 64. The minimum EER of each speech data are tabulated in Table 1, irrespective of Gaussian mixtures.

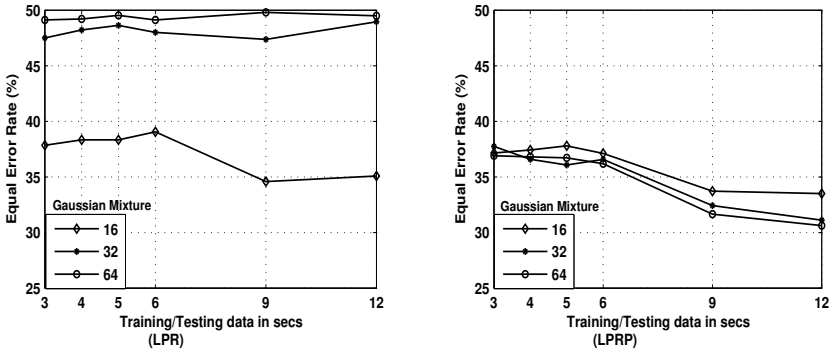

Figure 4. Performance of speaker verification system based on LPR and LPRP individual features using GMM modeling

The performance of individual features are analysed by considering 3s-3s data size as shown in Figure. 2. From the experimental results it was observed that, the individual feature MFCC provides a reduced EER which is less by 0.59% and 0.11% of △MFCC and △△MFCC respectively.

The results of LPCC features for the same data size is shown in Figure. 3, the individual feature LPCC provides a reduced EER which is less by 1.81% and 1.68% of △LPCC and △△LPCC respectively.

The two points can be noticed from these results. First point, static characteristics provides better performance as compared with dynamic characteristics. The second point is, the individual features of LPCC and its derivatives gives better verification performance than MFCC and its derivatives.

The results of LPCC features for the same data size is shown in Figure. 4. From the experimental results it was observed that, the reduced EER of LPR which is greater than 2.69% and 4.77% of MFCC and LPCC respectively. Further, the reduced EER of LPRP which is more by 2% and 4.08% of MFCC and LPCC respectively. It clearly shows that performance of vocal tract features gives better EER as compared to excitation source features.

The same study is also conducted for other data sizes of 4s-4s, 5s-5s, 6s-6s, 9s-9s and 12s-12s to verify the performance using individual features. In all the cases, the results shows that EER decreases as we increased the train and test data.

The GMM modeling works very well in case of sufficient data [20]. To overcome this problem, we used GMM-UBM modeling. UBM should be trained in such a way that it should have equal number of male and female speakers. In our experiment the total duration of male and female speakers is 1506 sec each.

To study the significance of GMM-UBM modeling, the same set of experiments are conducted. The experimental results are shown in Figure. 5, Figure. 6, Figure.7 for individual features using (MFCC, △MFCC, △△MFCC), (LPCC, △LPCC, △△LPCC) and (LPR, LPRP) respectively. The Gaussian mixtures considered are 16, 32, 64, 128 and 256 as additional UBM speech data is used for training. Table 2. represents the minimum EER of individual features for different speech data and different amount of Gaussian mixtures.

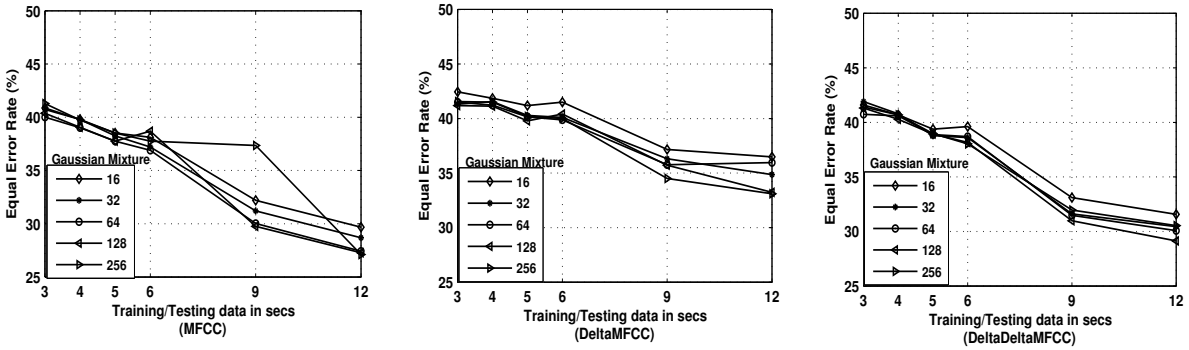

Figure 5. Performance of speaker verification system based on MFCC individual features using GMM-UBM modeling

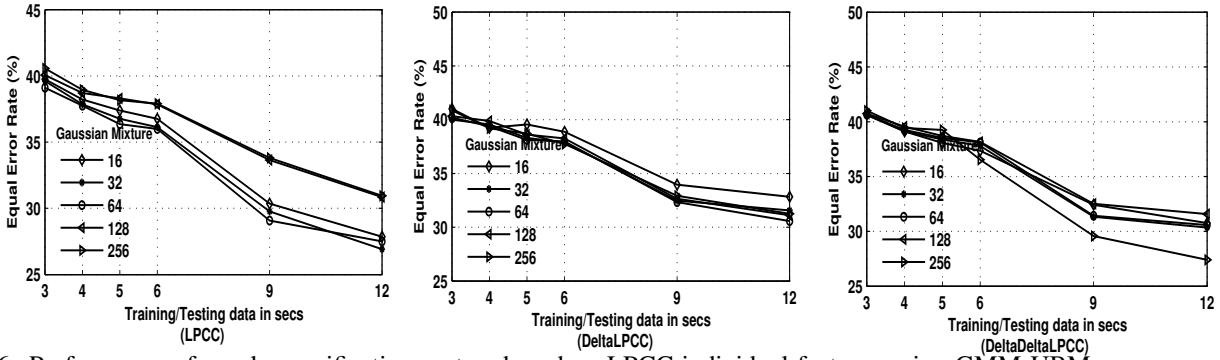

Figure 6. Performance of speaker verification system based on LPCC individual features using GMM-UBM modeling

Consider 3s-3s data for individual features MFCC and its derivatives, MFCC provides a reduced EER which is less by 1.18% and 0.73% of less △MFCC and △△MFCC respectively. In case of LPCC feature for same data size, the individual feature LPCC provides a reduced EER which is less by 0.9% and 1.49% of △LPCC and △△LPCC respectively.

In this modeling also, static characteristics provides better performance as compared with dynamic characteristics. Further, the individual features of LPCC and its derivatives gives better verification performance than MFCC and its derivatives.

Consider LPR and LPRP features for 3s-3s data size. The minimum EER of LPR which is more by 1.35% and 2.25% of MFCC and LPCC respectively. Further, LPRP is also having 1.15% and 2.05% higher in EER as compared with MFCC and LPCC respectively. The same study is also conducted for other data sizes of 4s-4s, 5s-5s, 6s-6s, 9s-9s and 12s-12s to verify the performance using individual features. Here also, the results shows that EER decreases as we increased the train and test data.

From these two modeling techniques it is clear that, performance of vocal tract features gives better EER as compared to excitation source features. Further, the individul features extracted from varies extraction techniques are different and hence they may combine to further improve the speaker verification performance under limited data condition.

In Table 1 and 2, it was observed that irrespective of speech data size and individual features, the minimum EER of GMM-UBM performance is better than GMM.

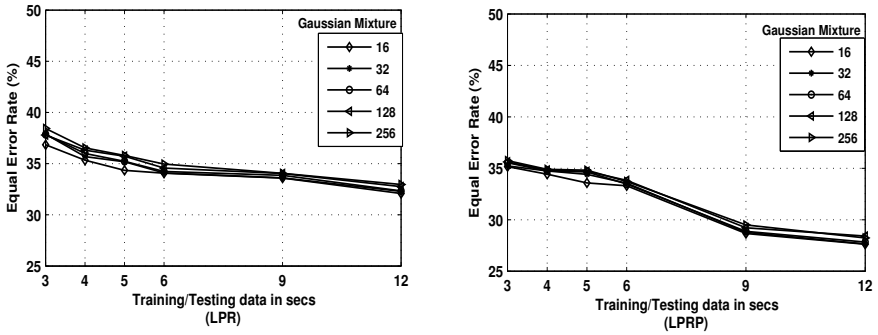

Figure 7. Performance of speaker verification system based on LPR and LPRP individual features using GMM-UBM modeling

Table 2. Comparison of minimum EER(%) for individual features using different amount of training and testing data for GMM-UBM

Individual Features	Training/Testing data					
	3s-3s	4s-4s	5s-5s	6s-6s	9s-9s	12s-12s
MFCC	40.01	39.02	37.75	36.90	29.35	27.12
△MFCC	41.19	41.14	39.79	39.88	36.31	33.10
△△MFCC	40.74	40.28	38.79	38.03	30.98	29.14
LPCC	39.11	37.75	36.35	35.99	29.08	26.91
△LPCC	40.01	39.20	38.12	37.75	32.33	30.57
△△LPCC	40.60	39.11	38.07	36.54	29.58	27.41
LPR	41.36	40.32	39.25	37.04	36.06	33.28
LPRP	41.16	40.43	39.16	37.23	36.16	33.02

4.4. Combination of features performance using GMM and GMM-UBM

The speaker verification system using limited data contains speech data of few seconds. Due to this the available feature vectors are less in numbers. The performance of speaker verification system can be increased by combining feature vectors of different features. The combination of features is accomplished by a simple concatenation of the feature sets obtained by different feature extraction techniques.

The performance of speaker verification system for combination of features (MFCC, △, △△, LPR and LPRP) for different data sizes and modeling is done by GMM. The experimental results are shown in Figure. 8 and 9 for combinations of MFCC and LPCC respectively. The minimum EER of varies Gaussian mixtures of each speech data are tabulated in Table 3. Further, consider Figure. 8(a) to analyse the performance for multiple combination of features with MFCC using 3s-3s data.

From the experimental results it was observed that, the combination of features (MFCC+△+△△) is providing minimum EER of 44.35% for Gaussian mixture of 32 and the individual features MFCC, △ and △△ are providing minimum EER of 45.27%, 45.75% and 45.16% respectively for the Gaussian mixture of 16. The (MFCC+△+△△) provides a reduced EER which is less by 0.92%, 1.4% and 0.81% MFCC, △ and △△ respectively. The performance of MFCC and its derivatives (MFCC+△+△△) is better than individual performance of MFCC, △MFCC, △△MFCC. This is due to combination of both static and dynamic characteristics of speech data in training and testing.

The (MFCC+LPR) is providing minimum EER of 37.75% for Gaussian mixture of 16. The individual features LPR is providing minimum EER of 47.85% for Gaussian mixture of 16 and which is more by 10.1% of (MFCC+LPR). The (MFCC+LPR) provides a reduced EER which is less by 6.6% of (MFCC+△+△△). The combination of (MFCC+LPR) performance is better than (MFCC+△+△△).

The (MFCC+LPRP) is having minimum EER of 37.62% for the Gaussian mixture of 64. The individual features LPRP is providing minimum EER of 47.16% for Gaussian mixture of 32 and which is more by

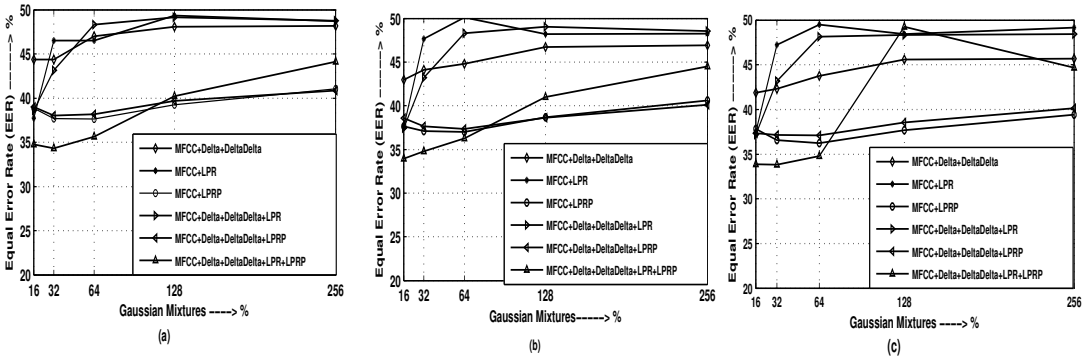

Figure 8. Performance of speaker verification system for MFCC and different combined system using (a) 3s-3s, (b) 4s-4s and (c) 5s-5s and modeling using GMM.

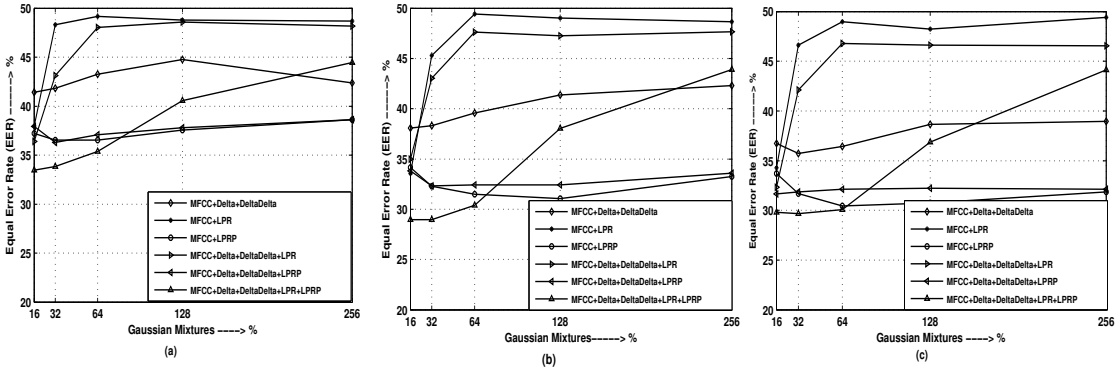

Figure 9. Performance of speaker verification system for MFCC and different combined system using (a) 6s-6s, (b) 9s-9s and (c) 12s-12s data and modeling using GMM.

9.54% of (MFCC+LPRP). The combination of (MFCC+LPRP) provides a reduced EER which is less by 6.73% of (MFCC+△+△△). This is due to combination of both vocal tract and excitation source information. The LPR contains Glottal Closures Instants (GCIs) related to excitation source information. Whereas, LPRP contains speaker-specific sequence information [24]. The LPR and LPRP features contains different characteristics of speaker-specific excitation information.

The (MFCC+△+△△+LPR) and (MFCC+△+△△+LPRP) is having minimum EER of 37.63% and 37.03% for the Gaussian mixture of 16 respectively and provids reduced EER which is less by 0.12% and 0.59% of (MFCC+LPR) and (MFCC+LPRP) respectively.

The combination of (MFCC+△+ △△+LPR+LPRP) provide minimum EER of 34.32% for Gaussian mixture of 32. Further, this combination provides reduced EER which is less by 10.03%, 3.43%, 3.3% 3.31% and 2.71% of (MFCC+ △+△△), (MFCC+LPR), (MFCC+LPRP), (MFCC+△+△△+LPR) and (MFCC+△+△△+LPRP) respectively.

The combined (MFCC+△+△△+LPR+LPRP) system performs better as compared to other combined systems performance for all training and testing data. This is because, in case of (MFCC+ △+△△+LPR+LPRP) the speaker-specific information includes static, transitional characteristics and excitation source. The same trend is observed for remaining data sizes are given in Figure. 8 and Figure. 9.

From above mentioned results, we have observed that, if we increase training and testing data the performance of combined system shows significant improvement in EER. To study the significance of LPCC and combined system the same set of experiments are conducted as in case of MFCC and combined system.

The experimental results are shown in Fig. 10 and Fig. 11 for combination of features (LPCC, △,

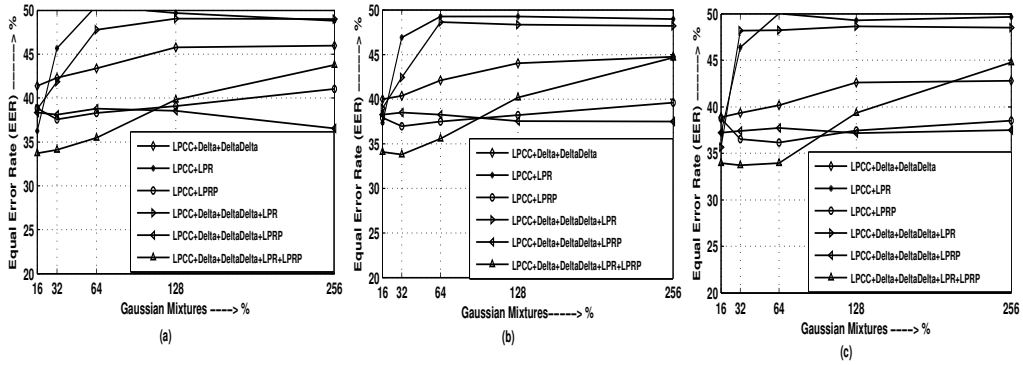

Figure 10. Performance of speaker verification system for LPCC and different combined system using (a) 3s-3s, (b) 4s-4s and (c) 5s-5s data and modeling using GMM.

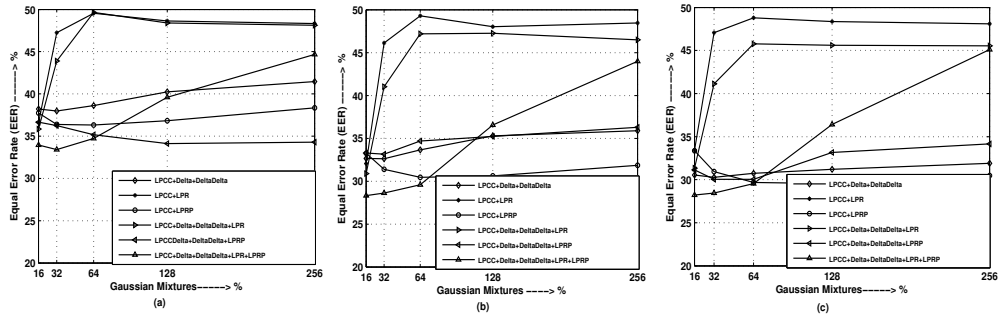

Figure 11. Performance of speaker verification system for LPCC and different combined system using (a) 6s-6s, (b) 9s-9s and (c) 12s-12s data and modeling using GMM.

△△, LPR and LPRP). The modeling is done by GMM. Consider 3s-3s data, the following experimental results are observed from Fig. 10(a).

The combination of features (LPCC+△+ △△) is providing minimum EER of 41.37% for Gaussian mixture of 16 and the individual features LPCC, △ and △△ are providing minimum EER of 43.08%, 44.89% and 44.76% respectively for the Gaussian mixture of 16. The (LPCC+ △+△△) provides a reduced EER which is less by 1.71%, 3.52% and 3.39% of LPCC, △ and △△ respectively. The performance of LPCC and its derivatives (LPCC+△+ △△) is better than individual performance of LPCC, △LPCC, △△LPCC. This is due to combination of both static and dynamic characteristics of speech data in training and testing.

The (LPCC+LPR) is providing minimum EER of 36.26% for Gaussian mixture of 16. The (LPCC+LPR) provides a reduction in EER which is less by 6.48% of LPR. The (LPCC+LPR) provides a reduction in EER which is less by 5.11% of (LPCC+△+ △△). The combination of (LPCC+LPR) performance is better than (LPCC+△+△△). The (LPCC+LPRP) is having minimum EER of 37.57% for the Gaussian mixture of 32. The individual features LPRP is provides a reduced EER which is more by 9.59% (LPCC+LPR). The (LPCC+LPRP) provides a reduction in EER which is less by 1.31% of (LPCC+LPR). The combination of (LPCC+LPRP) performance is better than (LPCC+LPR). This is because LPR and LPRP contains different speaker-specific information.

The (LPCC+△+△△+LPR) and (LPCC+△+△△+LPRP) is having minimum EER of 36.12% and 37.54% for the Gaussian mixture of 16 respectively and provids reduced EER which is less by 0.14% and 0.03% of (LPCC+LPR) and (LPCC+LPRP) respectively.

The combination of (LPCC+△+ △△+LPR+LPRP) provide minimum EER of 33.69% for Gaussian mixture of 16. The reduced EER which is less by (LPCC+△+ △△+LPR+LPRP) is 7.68%, 2.57%, 3.88% , 2.73% and 3.85% of (LPCC+△+ △△), (LPCC+LPR), (LPCC+LPRP), (LPCC+△+ △△+LPR) and (LPCC+△+ △△+LPRP) respectively. The combined (LPCC+△+ △△+LPR+LPRP) system perform better

Table 3. Comparison of minimum EER(%) for different combinations of features using different amount of training and testing data for GMM

Individual Features	Training/Testing data					
	3s-3s	4s-4s	5s-5s	6s-6s	9s-9s	12s-12s
MFCC+△+△	44.35	44.12	41.86	41.41	38.07	32.61
MFCC+LPR	37.75	37.57	37.48	37.98	33.55	32.06
MFCC+LPRP	37.62	36.99	36.22	36.54	32.07	31.44
MFCC+△+△+LPR	37.63	37.52	37.34	36.40	31.65	31.03
MFCC+△+△+LPRP	37.03	36.35	36.12	36.31	31.33	31.23
MFCC+△+△+LPR+LPRP	34.32	33.96	33.83	33.46	28.95	28.31
LPCC+△+△	41.37	39.97	38.88	37.98	32.61	30.26
LPCC+LPR	36.26	37.86	37.48	36.58	32.06	33.42
LPCC+LPRP	37.57	36.94	36.17	36.31	30.44	29.53
LPCC+△+△+LPR	36.12	36.85	35.64	35.86	31.32	30.89
LPCC+△+△+LPRP	37.54	36.48	36.12	34.13	30.15	29.12
LPCC+△+△+LPR+LPRP	33.69	33.78	33.73	33.42	28.31	28.22

as compared to other combined systems performance for all training and testing data. This is because, the combination of (LPCC+△+△△+LPR+LPRP) contains static, transitional characteristics and excitation source information. The same trend is observed for remaining data sizes as shown in Figure. 10 and Figure. 11.

Table 3. provides the comparison of different combined systems for different amount of training and testing data. The EER of (LPCC+△+△△) is less by 2.98%, 4.15%, 2.98%, 3.43%, 5.46% and 5.46% of (MFCC+△+△△) for 3s-3s, 4s-4s, 5s-5s, 6s-6s, 9s-9s and 12s-12s data respectively. The same trend has been observed for remaining combinations. In this experimental study, we observed that when both training and testing data are limited, the (LPCC+△+△△+LPR+LPRP) is having minimum EER compared to all other combination in case of GMM modeling. This is because LPCC and its derivatives along with excitation source features are able to capture more speaker-specific information from speech data, this will create different characteristics between speakers [25].

To study the significance of GMM-UBM for combination of features the following experiments are analysied. The performance of speaker verification system for combination of features (MFCC, LPCC, △, △△, LPR and LPRP) for different data sizes using GMM-UBM as a modeling technique is shown in Figure. 12 to Figure. 15. Further, The minimum EER of varies Gaussian mixtures of each speech data are tabulated in TABLE IV. Consider 3s-3s data, the following points are observed in this experimental setup as shown in Figure. 12 and Figure. 14.

The combination of features (MFCC+△+ △△) and (LPCC+△+ △△) is having minimum EER of 38.84% and 36.44% respectively. The (LPCC+ △+△△) is providing reduced EER which is less by 2.4% of (MFCC+△+ △△).

The combination of features (MFCC+ LPR) and (MFCC+LPRP) is having minimum EER of 38.3% and 36.54% respectively. Further, the minimum EER of (LPCC+LPR) and (LPCC+LPRP) is 36.94% and 34.55% respectively. The (LPCC+LPR) and LPCC+LPRP) is providing reduced EER of 1.36% and 1.99% less in EER as compared to (MFCC+LPR) and (MFCC+LPRP) respectively.

The combination of features (MFCC+△+△△+LPR) and (MFCC+△+ △△+LPRP)is having minimum EER of 34.73% and 34.74% respectively. Further, the minimum EER of (LPCC+△+ △△+LPR) and (LPCC+△+ △△+LPRP) is 34.12% and 33.93% respectively. The (LPCC+△+ △△+LPR) and (LPCC+△+ △△+LPRP) is providing reduced EER which is less by 0.61% and 0.81% of (MFCC+△+ △△+LPR) and (MFCC+△+△△+LPRP) respectively.

The combination of features (MFCC+△+ △△+LPR+LPRP) is providing reduced EER which less by 6.19%, 4.29%, 2.08%, 3.89% and 2.09% of (MFCC+ △+△△), (MFCC+LPR), (MFCC+LPRP), (MFCC+△+△△+LPR) and (MFCC+△+△△+LPRP) respectively.

The combination of features (LPCC+ △+△△+LPR+LPRP) is providing reduced EER which is less by 2.61%, 4.47%, 0.72%, 2.29% and 0.1% of (LPCC+ △+△△), (LPCC+LPR), (LPCC+LPRP), (LPCC+△+△△+LPR) and (LPCC+△+ △△+LPRP) respectively.

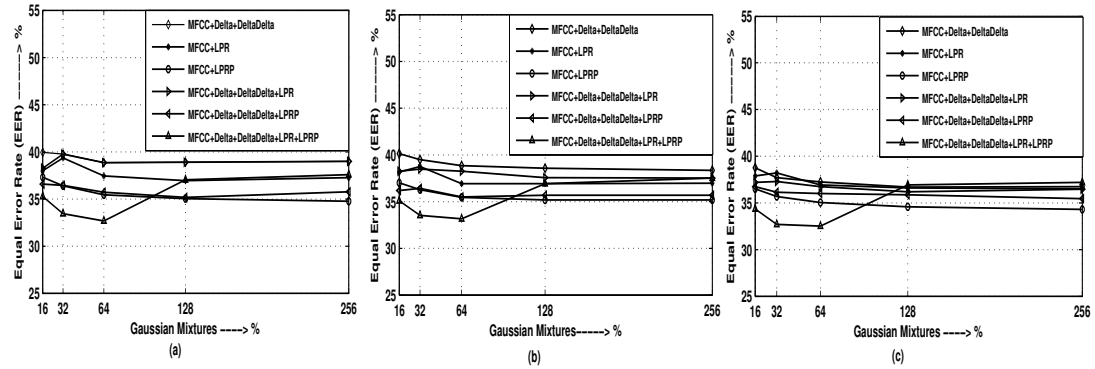

Figure 12. Performance of speaker verification system for MFCC and different combined system using (a) 3s-3s, (b) 4s-4s and (c) 5s-5s data and modeling using GMM-UBM.

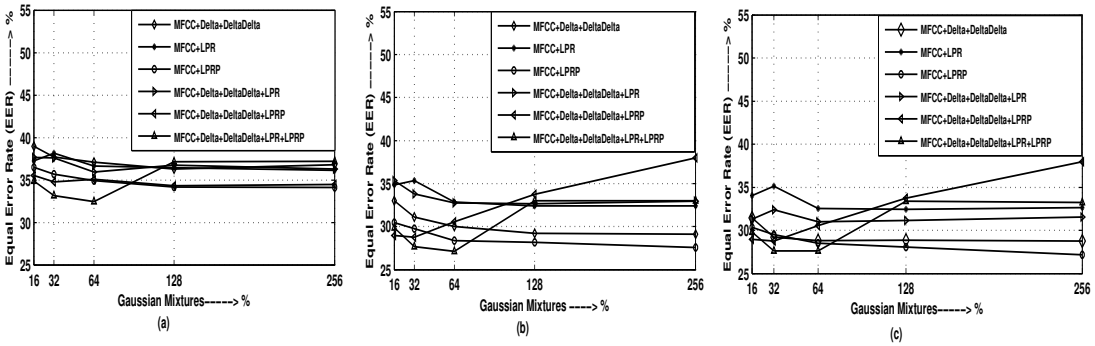

Figure 13. Performance of speaker verification system for MFCC and different combined system using (a) 6s-6s, (b) 9s-9s and (c) 12s-12s data and modeling using GMM-UBM.

The combination of features (MFCC+△+ △△+LPR+LPRP) is providing reduced EER which is less by 1.18% of (LPCC+△+△△) respectively. The same trend is not observed for remaining data sizes in all combinations and clear explanation is given below.

Consider Table 4., the table highlights some of the important points. First point is, from the experimental results it was observed that, the combinations of features (LPCC+△+ △△), (LPCC+LPR) and (LPCC+LPRP) is having minimum EER as compared to (MFCC+△+△△), (MFCC+LPR) and (MFCC+LPRP) respectively for all data sizes.

Second point is, the combination of (LPCC+△+ △△+LPR) and (LPCC+△+ △△+LPRP) is having minimum EER as compared to (MFCC+ △+ △△+LPR) and (MFCC+△+ △△+LPRP) only for 3s-3s, 4s-4s,5s-5s and 6s-6s data sizes. As data size increase (9s-9s and 12s-12s) MFCC and its combination is having minimum EER as compared to LPCC and its combination.

Third point is, the combination of (MFCC+ △+△△+LPR+LPRP) is having minimum EER as compared to (LPCC+△+△△+LPR+LPRP)respectively for all data sizes.

An important observation has been made to the end when the amount of training and testing data is very less, LPCC and its combination gives better verification performance as compared with MFCC and its combination. Further, when considering multiple features like vocal tract and its derivatives along with excitation source features or increasing training and testing data, MFCC and its combination gives better performance as compared with LPCC and its combination.

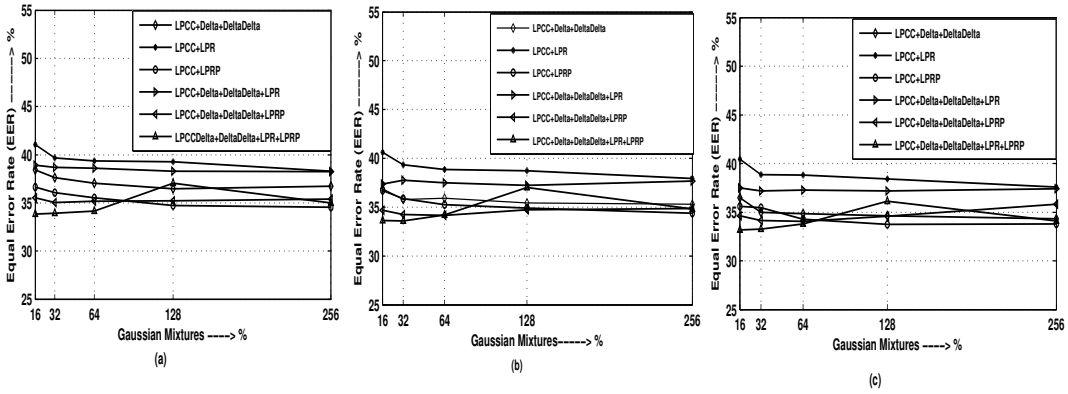

Figure 14. Performance of speaker verification system for LPCC and different combined system using (a) 3s-3s (b) 4s-4s and (c) 5s-5s data and modeling using GMM-UBM.

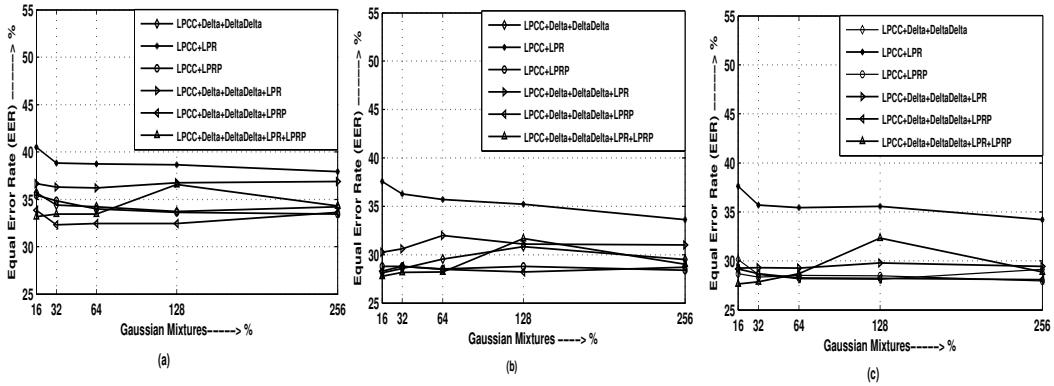

Figure 15. Performance of speaker verification system for LPCC and different combined system using (a) 6s-6s (b) 9s-9s and (c) 12s-12s data and modeling using GMM-UBM.

5. CONCLUSIONS

In this paper, we demonstrated the significance of different features and their combinations to improve the speaker verification perform under limited data condition. First, we studied the working principles of individual feature extraction techniques. Then, we combined MFCC and its derivatives, (MFCC+LPR), (MFCC+LPRP), (MFCC+△+ △△+LPR), (MFCC+△+△△+LPRP) and (MFCC+△+△△+LPR+LPRP). The same procedure carried out for LPCC and its combination and modeled using GMM. It was observed that, the LPCC and its combination outperforms than MFCC and its combination of features.

Further, the same procedure was carried out in case of GMM-UBM modeling. The combination of LPCC and its derivatives, (LPCC+LPR) and (LPCC+LPRP) system yields good performance compared to the MFCC and its derivatives, (MFCC+LPR) and (MFCC+LPRP). Also, (LPCC+△+ △△+LPR) and (LPCC+△+ △△+LPR) performance is better than (MFCC+△+△△+LPR) and (MFCC+△+ △△+LPRP) only in case of 3s-3s, 4s-4s, 5s-5s and 6s-6s data. Further, as training and testing data increases to 9s-9s and 12s-12s, MFCC, its derivatives and LPR or LPRP systems combinations yields good performance compared to the LPCC, its derivatives and LPR or LPRP. Also (MFCC+△+ △△+LPR+LPRP) performance is better than (LPCC+△+△△+LPR+LPRP) for all data sizes. Therefore, we suggest that when training and testing data are limited, the combination of LPCC and other features with GMM-UBM modeling can be used as system for

Table 4. Comparison of minimum EER(%) for different combinations of features using different amount of training and testing data for GMM-UBM

Individual Features	Training/Testing data					
	3s-3s	4s-4s	5s-5s	6s-6s	9s-9s	12s-12s
MFCC+△+△	38.84	38.3	36.67	36.31	29.1	28.74
MFCC+LPR	38.3	37.94	37.57	37.94	33.6	34.19
MFCC+LPRP	36.54	36.12	36.04	35.95	28.75	28.12
MFCC+△+△+LPR	34.73	35.74	34.32	34.1	27.55	27.14
MFCC+△+△+LPRP	34.74	34.23	34.05	33.98	27.35	27.08
MFCC+△+△+LPR+LPRP	32.65	33.15	32.52	32.47	27.14	27.59
LPCC+△+△	36.44	35.3	34.3	33.73	28.13	28.12
LPCC+LPR	36.94	36.9	36.49	36.12	32.38	32.43
LPCC+LPRP	34.55	34.37	33.73	33.46	28.41	27.97
LPCC+△+△+LPR	34.12	35.14	34.04	33.12	28.11	27.85
LPCC+△+△+LPRP	33.93	33.55	32.84	32.52	28.21	27.36
LPCC+△+△+LPR+LPRP	33.83	33.6	33.19	33.19	27.77	27.65

speaker verification.

REFERENCES

[1] A. K. Jain and A. Ross and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 14, no. 1, pp. 4–20, 2004.

[2] G. Pradhan and S. R. Prasanna, "Speaker verification under degraded condition: a perceptual study," *International Journal of Speech Technology.*, pp. 405–417, 2011.

[3] J. P. Campbell, "Speaker recognition: A tutorial." *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[4] H. S. Jayanna, "Limited data speaker recognition," in *Ph.D dissertation, Indian Institute of Technology*, (2009).

[5] L. R. Rabiner and B. H. Juang, "Fundamentals of speech recognition," *Pearson Education, Singapore.*, (1993).

[6] S. R. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Commun.*, vol. 48, pp. 1243–1261, 2006.

[7] G. Doddington, "Speaker recognition based on idiolectal differences between speakers." *Eurospeech*, pp. 2521–2524, 2001.

[8] G. Pradhan, "Speaker verification under degraded conditions using vowel-like and nonvowel-like regions," *Ph.D dissertation, Indian Institute of Technology.*, (2014).

[9] F. Bimbot, J. Bonastreand, and et al, "A tutorial on text-independent speaker verification," *EURASIP Journal on applied signal processing.*, vol. 4, pp. 430–451, 2004.

[10] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Acoust., Speech, Signal Processing.*, vol. 2, pp. 639–643, 1994.

[11] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304–1312, 1974.

[12] R. Murty and Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognmition." *IEEE Signal Processing Letters*, vol. 13, pp. 52–55, 2006.

[13] A. Salman, E. Muhammad, and K. Khurshid, "Speaker verification using boosted cepstral features with gaussian distributions." pp. 1–5.

[14] Hsu, Wei-Chih and Lai, Wen-Hsing and Hong, Wei-Ping, "Usefulness of Residual-Based Features in Speaker Verification and Their Combination Way with Linear Prediction Coefficients," in *Multimedia Workshops, 2007. ISMW'07. Ninth IEEE International Symposium on*, 2007, pp. 246–251.

[15] S. Furui, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, no. 3, pp. 342–350, 1981.

[16] Doddington, George R, "Speaker recognition?identifying people by their voices," *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1651–1664, 1985.

[17] G. L. Sarada, T. Nagarajan, N. Hemalatha, and A. Murthy, "Automatic transcription of continuous speech

using unsupervised and incremental training," *Int. Conf. Spoken Language Process.*, vol. 18, pp. 405–408, 2004.
[18] H. S. Jayanna and S. R. M. Prasanna, "Limited data speaker identification," *Sadhana 35.5*, pp. 525–546, 2010.
[19] Prakash, Vinod, and J. H. Hansen, "In-set / out-set speaker recognition under sparse enrollement," *IEE Trans. Audio Speech Lang. Process.*, no. 7, pp. 2044–2052, 2007.
[20] A.E. Rosenberg, "Automatic speaker verification: A review," in *Proc. IEEE*, vol. 64, no. 4, 1976, pp. 475–487.
[21] Taufiq Hasan and John H. L. Hansen, "A study on universal background model training in speaker verification," *IEEE Tractions on audio, speech and language processing*, vol. 19, Sept. 2011.
[22] "Nist2003,http://www.itl.nist.gov/iad/mig//tests/sre/2003/2003-spkrec-evalplan-v2.2.pdf[online]."
[23] J. Makhoul, "Linear prediction: a tutorial review," *Image and Vision Computing*, vol. 28, no. 1, pp. 55–63, 2010.
[24] H.S. Jayanna, "Limited data speaker recognition," Ph.D. dissertation, Indian Institute of Technology Guwahati, Dept. of of Electronics & Communication Engg., Guwahati, India, Mar. 2009.
[25] Wong, Eddie and Sridharan, Sridha, "Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification," in *Proc. Int. Symp. Intelligent Multimedia, Video and Speech Processing*, 2001, pp. 95–98.