T.R. Jayanthi Kumari* and H.S. Jayanna

i-Vector-Based Speaker Verification on Limited Data Using Fusion Techniques

https://doi.org/10.1515/jisys-2017-0047
Received November 30, 2016; previously published online May 3, 2018.

Abstract: In many biometric applications, limited data speaker verification plays a significant role in practical-oriented systems to verify the speaker. The performance of the speaker verification system needs to be improved by applying suitable techniques to limited data condition. The limited data represent both train and test data duration in terms of few seconds. This article shows the importance of the speaker verification system under limited data condition using feature- and score-level fusion techniques. The baseline speaker verification system uses vocal tract features like mel-frequency cepstral coefficients, linear predictive cepstral coefficients and excitation source features like linear prediction residual and linear prediction residual phase as features along with i-vector modeling techniques using the NIST 2003 data set. In feature-level fusion, the vocal tract features are fused with excitation source features. As a result, on average, equal error rate (EER) is approximately equal to 4% compared to individual feature performance. Further in this work, two different types of score-level fusion are demonstrated. In the first case, fusing the scores of vocal tract features and excitation source features at score-level-maintaining modeling technique remains the same, which provides an average reduction approximately equal to 2% EER compared to feature-level fusion performance. In the second case, scores of the different modeling techniques are combined, which has resulted in EER reduction approximately equal to 4.5% compared with score-level fusion of different features.

Keywords: Mel-frequency cepstral coefficient (MFCC), linear prediction cepstral coefficients (LPCC), linear prediction residual (LPR), linear prediction residual phase (LPRP), i-vector, feature-level fusion, score-level fusion.

1 Introduction

The last two decades, biometric technologies plays important role in recognizing a person for various applications [11, 18]. Person authentication using biometric application includes fingerprint, face, iris, ear, retina, DNA and speech. These biometric features are used to develop practical system for various applications. Among these biometric features, speech is considered as one of the features to recognize a person, and it is called speaker recognition. The subcategories of speaker recognition are speaker identification and speaker verification [32]. In identification, the registered test speech data match with all the speaker models to decide the current speaker of the test speech data [32]. Further, speaker verification uses the claimed speaker data to verify against the claimed model. Based on speech data, speaker verification system can be divided into text-dependent and text-independent systems. The same set of text data is used in text-dependent system, and for text-independent system, different text data are used to train and test the speakers [13].

In the present situation, speaker verification performs well in case of sufficient data. The sufficient data contains speech data of a few minutes (<1 min). Existing techniques like speech analysis, feature extraction, modeling and testing perform better under sufficient data condition. On the other hand, limited

***Corresponding author: T.R. Jayanthi Kumari,** Department of Electronics and Communication Engineering, Siddaganga Institute of Technology, Bengaluru 560077, Karnataka, India, e-mail: trjayanthikumari@gmail.com
H.S. Jayanna: Department of Information Science and Engineering, Siddaganga Institute of Technology, Tumkur 572103, Karnataka, India

data means speech data in terms of a few seconds (≤15 s). Many of the biometric applications have to be done using less amount of data to verify a speaker. Speech data can be analyzed in different techniques. The different analysis techniques in state-of-the-art speaker verification systems are segmental, subsegmental and suprasegmental analysis [10]. During segmental analysis, the frame size (FS) and frame rate (FR) in the range of 10–30 ms is used for speech analysis in order to extract the vocal tract information which is called single frame size and single frame rate (SFSR) analysis [19]. For subsegmental analysis, due to rapid variations in the excitation source information as compared to that of vocal tract information. In this case, the analysis of speech is done using FS and FR in the range of 3–5 ms [27]. The behavioral aspects of the speaker is captured by analyzing the FS and FR in the range of 100–300 ms during suprasegmental analysis [12].

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 [4, 5]. The mel-frequency cepstral coefficients (MFCC) [28] and linear prediction cepstral coefficients (LPCC) [2] are used to extract the information of vocal tract. The speech signal contains both static and dynamic characteristics. The MFCC and LPCC feature set contain only static characteristics. The dynamic characteristic represented by Delta (Δ) and Delta-Delta (ΔΔ) contains some more speaker information, which is useful in speaker verification [2]. The excitation source features are extracted using prediction residual (LPR) and linear prediction residual phase (LPRP) [24].

Reynolds [28] compared the vocal tract features for speaker recognition and reported that MFCC and LPCC give a better performance compared with the other features. The reasons may be less intra-speaker variability and the availability of rich spectral analysis tools. Speaker recognition performance can be increased by combining vocal tract features and excitation source features [24]. In this study, the training and testing data are limited to 3 s. The feature used as MFCC and its derivatives and either LPR or LPRP. They considered frame size and frame rate in the range of 10–30 and 3–5 ms for MFCC and its derivatives and LPR or LPRP, respectively. The study showed that the combination of MFCC and its derivatives along with either LPR or LPRP features gives better performance compared with individual performance. The author reported that the combined features give better performance compared with individual features in speaker recognition.

Das et al. [9] reported on the effort made to develop speech-based person authentication system involving three different modules of speaker verification under low security applications. They used three different modules of the speaker verification system. The modules are voice-password, text-dependent and text-independent speaker verification. The combination of these system is called multi-level system. The author reported that a multi-level system performs better compared to individual modules for speaker verification. Also, the functionality of each module can be moderated according to the type of application for which the system is designed.

Das et al. [8] concentrated on highlighting the requirement of phonetic match in a text-independent speaker verification framework from the perspective of having practical deployable systems. The authors considered sufficient train data and short test data, and i-vector modeling is used. Three minutes of read speech is considered for training the speaker models. The chosen phrase and a text-constrained phrase are used for testing, and the database used is NIST 2003. The EER obtained is 23%.

Pandey et al. [26] conducted experiments for sufficient training data and limited test data. Limited test data are created by truncating the test data of NIST SRE 2003 database. Four different cases of limited test data duration of 10, 5, 3 and 2 s are considered, and the EER obtained is 5.87%, 10.52%, 16.94% and 22.31%, respectively.

The speaker verification system contains different types of pattern-matching techniques like template matching, probabilistic model and artificial neural network. The nearest-neighbor vector quantization (VQ) belongs to template matching and probabilistic model contains Gaussian mixture model (GMM), GMM–universal background model (UBM), joint factor analysis and i-vector. Further, time delay neural work and decision tree belong to artificial neural network. Among these, i-vector is used for modeling in the present work. The state-of-the-art speaker verification system prefers i-vector-based speaker modeling technique over the

conventional modeling approaches due to its compact representation and compatible channel/session compensation techniques [10]. The i-vectors are the low-dimensional representation of GMM mean supervector-derived using factor analysis. The recent works in this area address the problem of mismatch in sensors, environment and language, and changes across sessions [15]. To address these problems, different techniques like score/handset/test normalization [3], within class covariance normalization (WCCN) [16], linear discriminant analysis (LDA) and joint factor analysis (JFA) [36] are available.

Li et al. [23] proposed a new way of applying PLDA mixture models for robust speaker verification. The key idea is to use a classifier to guide the training of PLDA mixture models so that each mixture component precisely models one cluster in the i-vector space. In the testing stage, the verification scores are computed by combining the PLDA scores with dynamic weights depending on the posterior probabilities given by the classifier. The proposed method was compared against state-of-the-art models on the NIST SRE 2012 data set. It achieves much better performance than PLDA and conventional mixture of PLDA under SNR-level variability and channel-type variability.

Al-Ali et al. [1] introduced the use of DWT-based MFCC features and their combination with traditional MFCC features for forensic speaker verification. The i-vector and PLDA classifier are used in this work. Experimental results indicate that the fusion feature warping DWT-MFCC and feature-warped MFCC approach achieved better performance under most environmental noise, reverberation, and noisy and reverberation environments. The robustness in the performance of the fusion feature approach could be used in forensic applications.

Kanagasundaram et al. [21] provided experimental analysis of text-independent speaker verification system by varying the amount of train and test data. The analysis given by Das et al. [7] for very less amount of test data (<10 s) shows that the performance drops significantly even though sufficient speech data are used during training. This trend of downfall in performance for limited test data motivates us to consider a source feature that captures complementary speaker information with limited data. Das et al. [7] demonstrated speaker verification for limited data condition using i-vector modeling. They used 150-dimensional LDA and a full-dimensional WCCN matrix. The study reported that different source features along with MFCC gives better performance than individual MFCC. Dey et al. [11] demonstrated the performance of speaker verification using i-vector as modeling techniques. The study reported that i-vector modeling gives better performance than GMM–UBM, and the performance of raw i-vector-based system was further improved by the use of LDA and WCCN followed by T-norm.

The literature in the works of Gudnason and Brookes [14] and Murty and Yegnanarayana [24] shows that better performance can be achieved by fusing the vocal tract with source excitation features. Studies by Prasanna et al. [27] and Chan et al. [6] show that while dealing with voice source features limited, the amount of train and test data can be used as compared with vocal tract features. This is because voice source features dependence on phonetic content is very less, where as in vocal tract features requires more amount of phonetic content to be captured for speaker modeling. For speaker modeling, a sufficient amount of phonetic content has to be captured in vocal tract features, whereas in excitation source features, the phonetic content dependency is very low. In case of limited data, the amount of data available is very less. The extracted features are insufficient to model well. Each modeling technique uses its own representation of the input pattern. Further, the testing method is used for verifying speaker is different for each modeling technique. Combining scores of different modeling techniques using score-level fusion may give better verification performance in limited data. This motivates us to use score-level fusion using different modeling techniques in case of limited data speaker verification.

In the testing phase, an unknown test speech is represented by channel/session-compensated feature vectors and compared against the claimed model to obtain similarity score. The similarity measure is done based on the employed modeling method. For instance, Euclidean distance [35], log likelihood score (LLS) [29, 30] and log likelihood score ratio (LLSR) [31] are used as the similarity scores for VQ and GMM and GMM–UBM modeling technique, respectively. In an i-vector-based speaker verification system, the test speech is represented as the channel/session-compensated i-vector, and the cosine kernel between the claimed and test i-vectors is used as the similarity measure [10].

The remaining structure of the article is arranged as follows: The database and experimental setup for the present work is explained in Section 2. The development of i-vector-based speaker verification system using vocal tract and source features is explained in Section 3. Section 4 provides the experimental results and discussion of the speaker verification system. Section 5 includes the summary and conclusion for the present work.

2 Database and Experimental Setup for the Work

In the present work, speaker verification is done using the NIST 2003 database [25]. This standard data set contains 2915 speakers. In 356 train speakers, 207 were female speakers and 149 were male speakers. The NIST SRE 2003 evaluation plan contains 2559 test speakers for verifying the speakers. Apart from test and train speech samples, the database also contains development data set to train the UBM [31] and T-matrix. We have used Switchboard Corpus II cellular data of 1872 utterances as development data. A small portion of the development data contains of 251 female and 251 male speakers of roughly 10 h of duration are required to train two gender dependent UBM separately of 8, 16, 32, 64, and 128 Gaussian mixtures. These are combined to form a gender-independent UBM having 16, 32, 64, 128 and 256 Gaussian mixtures. Experiments are conducted for limited data, and we considered the train and test data of durations 3–3, 6–6, 9–9, 12–12 and 15–15 s each, and for better performance, we increased test data by keeping the train data at 15 s (15–20 and 15–25 s). The maximum Gaussian mixtures limited for UBM is 256.

3 Development of i-Vector-Based Speaker Verification System: Vocal Tract and Source Features

The speaker-specific information can be extracted from feature extraction techniques at a reduced data rate [33]. These feature vectors contain vocal tract, excitation source and behavioral traits of speaker-specific information [19]. 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. These features are modeled using i-vector modeling technique.

3.1 Vocal Tract Features

MFCC and LPCC feature extraction techniques are used to extract vocal tract features. The technique used to extract features of MFCC and LPCC are different, and the performance of these features also varies.

In the case of MFCC, the spectral distortion is minimized using hamming window. The magnitude frequency responses are 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 (DCT) 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. The computing method using LPCC is a recursion from LPC parameter to LPC cepstrum according to all-pole model. The coefficients of the all-pole filter form the LPC. It is equivalent to the smoothened envelope of the log spectrum of the speech. The part of the speech which has been windowed is used to calculate the LPC by either autocorrelation or covariance methods. The discrete Fourier transform (DFT) and inverse DFT can be avoided while calculating LPCC using the Durbin recursive method because those methods are time consuming and complex [17].

3.2 Excitation Source Features

The spectral features extracted from THE vocal tract are in the range of 10–30 ms. Some of the speaker-specific information such as the linear prediction (LP) residual and LP residual phase are ignored by the spectral features which can be utilized for speaker verification [27]. The following procedure is carried out to calculate the LP residual. First the LP analysis is used to predict the vocal tract information from the speech data, followed by suppressing the same from the speech data using the inverse filter formation [27, 34]. To calculate LPRP, the LP residual is divided by the Hilbert envelope [34]. The LPR contains information which is obtained by the excitation source mainly glottal closure instants (GCIS) and the speaker-specific information is contained by LPRP [20]. The LPR and LPRP features contain speaker-specific excitation source information which is dissimilar in characteristic. The advantage can be improved by combining these two features.

3.3 i-Vector

The total variability i-vector representation of speech utterances forms the basis for all state-of-the-art speaker verification systems [10]. The GMM mean supervectors of each utterance are projected to a low rank matrix to get a reduced dimension representation is called i-vector. The concatenation of mean vectors of adapted GMM produces GMM mean supervector for speaker utterance. Figure 1 shows the block diagram of the i-vector-based speaker verification system.

The total variability matrix is the channel variability and dominant speaker which are simultaneously represented in the low-rank matrix. If T is the total variability matrix, the i-vector w and adapted GMM mean supervector M_s can be related by the following equation,

$$M_s = m + Tw, \tag{1}$$

where m represents the speaker and channel-independent supervector (UBM mean supervector). The UBM represented by a weighted sum of C component Gaussian densities as $U=\{\mu_c, \lambda_c, \eta_c\}$ $c=1, 2,\ldots, C$ where μ_c, λ_c and η_c represents the mean vector, covariance matrix and weight associated with mixture component, respectively. It is assumed that L speech feature vectors $\{X_1, X_2,\ldots, X_L\}$ each having a dimension of F extracted from the speech signal. The mean of 0th-order which is the weight of the mixture component and mean of the first-order F_c centralized with respect to UBM is given by

$$N_c = \sum_{i=1}^{L} P(c \mid x_t, U) \tag{2}$$

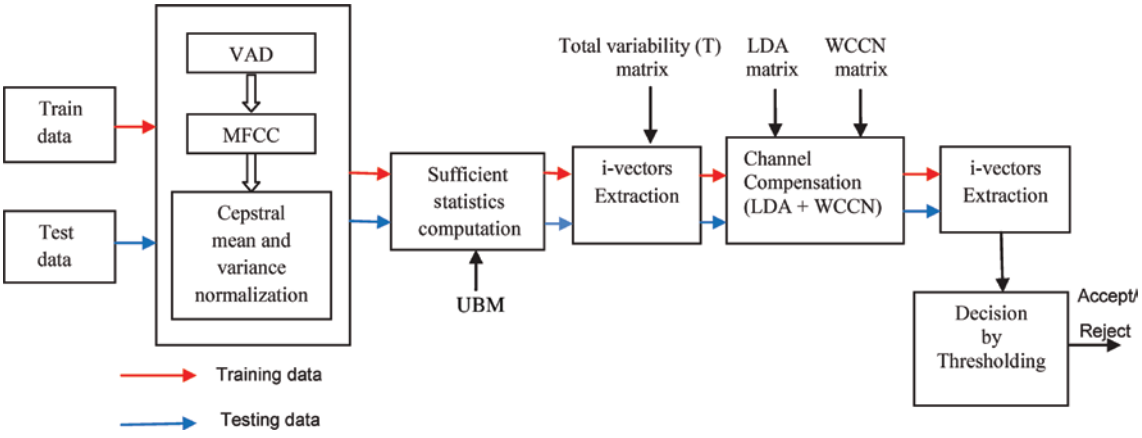

Figure 1: Block Diagram Representation of the Speaker Verification System Using i-Vector Modeling [21].

$$F_c = \sum_{i=1}^{L} P(c \mid x_t, U)(x_t - \mu_t), \tag{3}$$

where $c = 1, 2, \ldots C$ represent the components of the UBM, $P(c \mid x_t, U)$ is the posterior probability of the mixture component c generating the feature vector and is the mean of UBM component c.

The learning of the total variability matrix T from the development data is generally done using a variant of probabilistic principal component analysis modified to operate on the Baum-welch statistics of the speech data computed using UBM. The estimated i-vector W is computed by the following equation

$$W = \left(I + T' \overset{-1}{\Sigma} N(u) T\right)^{-1} T' \overset{-1}{\Sigma} F(u) \tag{4}$$

where $N(u)$ and Σ are the diagonal matrix of dimension $CF \times CF$ with diagonal blocks being NcI and Σc, respectively. $F(u)$ is supervector of dimension $CF \times 1$ generated by concatenating all first order Baum–Welch statistics (F_c) for a given utterance u. The speech utterances are represented in the form of i-vectors during training and testing phase. Computing the cosine kernel score of the two i-vectors $\hat{y}_{\text{clm}}$ and $\hat{y}_{\text{tst}}$ as given below, where $\hat{y}_{\text{clm}}$ is the i-vectors of the claimed which was obtained in the training phase and $\hat{y}_{\text{tst}}$ is the test speaker utterances:

$$\text{Score} = \frac{\langle \hat{y}_{\text{clm}}, \hat{y}_{\text{tst}} \rangle}{\| \hat{y}_{\text{clm}} \| \hat{y}_{\text{tst}} \|} \tag{5}$$

3.3.1 Session/Channel Compensation in i-Vector-Based Speaker Modeling

The i-vector extracted from the speech utterance contains both channel and speaker variabilities. Implementing channel/session compensation methods help improves the performance of the speaker verification system which is based on i-vector. The different compensation method of session/channel variability to improve the performance of the speaker verification system which is based on i-vector is given below:

(i) Linear discriminant analysis in linear discriminant analysis (LDA): the feature vectors are projected down to a set of new orthogonal axes where the intra-class variance caused by the channel is minimized and inter-class variance is increased [7, 10]. The projection matrix is composed of the eigen vectors corresponding to the best eigen values of the eigen analysis equation as

$$(W_c^{-1} B_c) v = \lambda v, \tag{6}$$

where W_c is the within-class covariance matrix, B_c is the between-class covarience matrix, v is an arbitrary vector and λ is the diagonal matrix of eigen values [17].

(ii) Within-class covariance normalization: Compensation of the effects of channel/session mismatch in i-vector and super vector representations can be done using within-class covariance normalization (WCCN) which is a type of linear transformation [10, 16]. The transformation minimizes the upper bounds on the classification error metric and hence minimizes the classification error. The transformation matrix B is obtained by Cholesky decomposition of the inverse of the within-class covariance matrix W as,

$$W^{-1} = BB^t \tag{7}$$

4 Experimental Results and Discussion

In this work, the experiments are conducted using different features for speaker verification system. The verification performance of the system can be calculated by using equal error rate (EER). It is an operating

point where false rejection rate (FRR) equals to false acceptance rate (FAR) [27]. The FRR is defined as ratio of number of rejective true speaker to total number of true speaker. The definition of FAR is ratio of number of accepted impostor to total number of impostor. For the present work, we extracted vocal tract and excitation source features. The vocal tract features are MFCC and LPCC, excitation source features are LPR and LPRP. 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 get good verification performance. The vocal tract and excitation source information are combined for improving the performance of the speaker verification system under limited data condition. All these features are in the dimension of 13 and 39.

First, we conducted experiments for 13 dimension features. The features of MFCC and LPCC are extracted by considering FS of 20 ms and FR of 10 ms. The LPR and LPRP features are extracted by considering FS of 12 ms and FR of 6 ms. The i-vectors are used as classifier. The Gaussian mixtures of 16, 32, 64, 128 and 256 are considered to model the speakers. The 0-th first-order statistics (GMM mean supervectors) of all speech data of train, test and development data are computed. The channel and session compensation can be computed using LDA analysis followed by WCCN.

Table 1 shows the performance of vocal tract and excitation source feature extraction techniques using different dimensions of i-vector and LDA in terms of EER and decision cost function (DCF). The experiments conducted for the 3–3 second training and testing data for different feature extraction techniques using different dimensions of i-vectors and LDA combination. The best performance is obtained for 100–50 and 50–50 compared to other combinations. For the present work, we considered dimension of i-vector and LDA is 100–50 and 50–50, respectively. The reason for considering this dimension is that, extracted features are in 13 dimensions, and data are limited. Therefore, for 6–6, 9–9, 12–12, 15–15, 15–20 and 15–25 s, this combination is used to model the speaker different amount of train, test and development data. The DCF also follows the same trend in improvement like EER, when EER decreases there are decrease in DCT also.

The performance of the i-vector-based speaker verification system developed using vocal tract features (MFCC and LPCC) and excitation source features (LPR and LPRP) are evaluated for different duration of train and test data for dimensions of 100–50 combination is shown in Table 2. Since the study is for limited data, we evaluated system performance until 15 s. Consider 3–3 second data, the minimum EER is 42.68% and 41.55% is obtained for MFCC and LPCC for Gaussian mixtures of 32 and 128, respectively, and the minimum EER of 39.92% and 40.28% is obtained for LPR and LPRP for the Gaussian mixtures of 64 and 32, respectively. The reduction in EER of LPCC is 1.13% less as compared with reduction in EER of MFCC and LPR is 0.36% less in reduction in EER as compared with LPRP. Further, Table 2 clearly shows that, when train and test data is increased performance will also increases in all feature extraction techniques. The similar trend is observed from Table 3. The minimum EER for 3–3 second data is 40.83% and 40.42% is obtained for MFCC and LPCC features for Gaussian mixtures of 128 in both cases. The LPCC is having reduced EER of 0.41% less as compared with reduced EER of MFCC. The LPR and LPRP is having minimum EER of 40.46% and 40.24% for Gaussian mixture of 128. The LPR is having 0.22% less in reduction as compared with LPRP. As we observed from these two tables, the dimensions of i-vector and LDA combinations 50–50 performance better

Table 1: EER of Speaker Verification for Different Dimensions of i-Vector and LDA Using the NIST 2003 Database for the 3–3 second Train/Test Data and Gaussian Mixture of 16, Features are MFCC, LPCC, LPR and LPRP (13 Dimensions).

Dimensions of (i-vector/LDA)	MFCC		LPCC		LPR		LPRP	
	EER%	DCT	EER%	DCT	EER%	DCT	EER%	DCT
400–150	49.86	0.8991	48.10	0.8994	49.10	0.8891	49.06	0.8873
200–150	49.41	0.8981	43.45	0.8152	49.63	0.8983	50.00	0.9000
100–100	43.63	0.8252	42.90	0.8060	45.24	0.8523	44.32	0.8342
100–50	43.36	0.8099	41.64	0.7860	40.51	0.7628	41.86	0.7927
50–50	43.58	0.8247	41.37	0.7812	40.42	0.7643	41.50	0.7821

Table 2: EER of Speaker Verification for Using NIST 2003 Database for Different Feature Extraction Techniques (13 Dimensions) and Modeling Done by i-Vectors.

Train/test data (s)	Feature extraction techniques	No. of dimension: i-vectors = 100; LDA = 50									
		16		32		64		128		256	
		EER%	DCT	EER%	DCT	EER%	DCT	EER%	DCT	EER%	DCT
3–3	MFCC	43.36	0.8099	**42.68**	0.8051	42.77	0.8011	42.68	0.8046	43.81	0.8221
	LPCC	41.64	0.7812	42.77	0.8028	42.09	0.7979	**41.55**	0.7851	41.59	0.7820
	LPR	41.86	0.7927	40.24	0.7593	**39.92**	0.7554	40.92	0.7734	40.46	0.7639
	LPRP	40.51	0.7628	**40.28**	0.7520	40.42	0.7663	41.37	0.7848	42.63	0.8063
6–6	MFCC	38.70	0.7306	38.16	0.7199	**37.08**	0.7017	39.43	0.7409	39.97	0.7496
	LPCC	39.70	0.7487	38.75	0.7333	39.47	0.7429	37.17	0.6998	**35.77**	0.6757
	LPR	38.79	0.7262	38.88	0.7252	**38.07**	0.7127	38.34	0.7250	38.79	0.7308
	LPRP	38.88	0.7306	**38.12**	0.7131	39.74	0.7503	39.83	0.7384	39.61	0.7469
9–9	MFCC	32.56	0.6180	31.12	0.5872	32.20	0.6103	29.94	0.5657	**29.85**	0.5641
	LPCC	31.12	0.5903	31.57	0.5953	27.68	0.5191	26.64	0.5028	**26.01**	0.4910
	LPR	35.95	0.6700	34.10	0.6315	**33.46**	0.6189	33.60	0.6289	33.92	0.6367
	LPRP	32.92	0.6133	**32.06**	0.6013	33.15	0.6262	34.95	0.6575	34.23	0.6449
12–12	MFCC	31.75	0.5993	30.71	0.5823	**28.68**	0.5415	30.98	0.5800	29.31	0.5541
	LPCC	28.36	0.5339	26.19	0.4971	25.11	0.4749	25.20	0.4768	**22.31**	0.4206
	LPR	34.10	0.6380	32.97	0.6230	32.24	0.5864	31.75	0.5888	**31.02**	0.5798
	LPRP	32.38	0.6042	**31.57**	0.5893	32.83	0.6143	33.24	0.6309	34.10	0.6406
15–15	MFCC	29.43	0.5608	29.11	0.5338	28.61	0.5413	27.35	0.5321	**26.21**	0.4985
	LPCC	21.64	0.4023	21.32	0.4164	20.87	0.3994	20.36	0.3852	**19.24**	0.3627
	LPR	31.79	0.5953	31.57	0.5893	30.98	0.5872	30.68	0.5823	**29.64**	0.5612
	LPRP	32.94	0.6814	32.24	0.6309	31.75	0.5972	31.39	0.5888	**30.30**	0.5926
15–20	MFCC	28.45	0.5438	28.32	0.5418	27.64	0.5333	27.91	0.5365	**25.32**	0.4775
	LPCC	21.35	0.4012	20.84	0.3965	20.62	0.3964	20.41	0.3832	**18.43**	0.3512
	LPR	29.34	0.5543	29.18	0.5523	28.84	0.5447	28.52	0.5432	**28.32**	0.5418
	LPRP	29.64	0.5578	29.34	0.5542	28.74	0.5441	28.85	0.5453	**28.84**	0.5543
15–25	MFCC	24.54	0.5554	24.32	0.5532	23.33	0.5476	23.21	0.5365	**22.82**	0.5275
	LPCC	18.32	0.3045	18.21	0.3022	17.64	0.2987	17.31	0.2943	**16.35**	0.2812
	LPR	28.11	0.5443	28.54	0.5423	27.72	0.5337	27.31	0.5323	**26.61**	0.5234
	LPRP	28.72	0.5478	28.61	0.5442	27.81	0.5351	27.61	0.5343	**26.74**	0.5243

The bold values represent minimum EER of particular feature extraction techniques.

as compared to 100–50 combination under limited data condition and also performance of LPCC features gives minimum EER as compared to minimum EER of MFCC features. Therefore, future experiments in this work 50–50 combination of i-vector and LDA is used.

In our earlier work [22], we evaluated the speaker verification system using GMM–UBM modeling for the NIST 2003 data set. The extracted features are MFCC and LPCC with FS of 20 ms and FR of 10 ms. The features are in the dimensions of 13. It was observed that GMM–UBM modeling works well under limited data [22]. The performance using vocal tract features for limited data conditions using GMM–UBM modeling is mentioned in Table 4. The performance of GMM–UBM modeling using 3–3 second data, the minimum EER of MFCC and LPCC is 40.10% and 39.06% respectively. Comparing the results of GMM–UBM and i-vectors modeling as mentioned in Tables 3 and 4. The reduction in EER is 0.73% and 0.36% for MFCC and LPCC in case of 3–3 second data compared to i-vector modeling. The same trend is not continued in 6–6, 9–9, 12–12, 15–15, 15–20 and 15–25-second data size. However, 6–6, 9–9, 12–12, 15–15, 15–20 and 15–25-second data, i-vector performs better than GMM–UBM. This is because, i-vector-based modeling is an advanced technique than GMM–UBM modeling. The i-vector extracted from speech data contains both channel/session variabilities and these variabilities can be compensated by various techniques to improve the performance of i-vector-based speaker verification system and these compensation techniques not present in GMM–UBM.

Further, it was observed that EER of LPCC is less than all other features in Table 3. The minimum EER obtained is 40.12%, 34.68%, 25.88%, 22.53%, 18.18%, 16.45% and 14.65% for LPCC features for 3–3, 6–6, 9–9, 2–12, 15–15, 15–20 and 15–25-second data compared to MFCC, LPR and LPRP feature extraction techniques.

Table 3: EER of Speaker Verification for Using NIST 2003 Database for Different Feature Extraction Techniques (13 Dimensions) and Modeling Done by i-Vectors.

Train/test data (s)	Feature extraction techniques	No. of dimension: i-vectors = 50; LDA = 50									
		16		32		64		128		256	
		EER%	DCT	EER%	DCT	EER%	DCT	EER%	DCT	EER%	DCT
3–3	MFCC	43.58	0.8247	42.72	0.8089	41.96	0.7943	41.77	0.7905	**40.83**	0.7629
	LPCC	41.37	0.7812	42.68	0.8093	42.05	0.7944	40.65	0.7703	**40.12**	0.7635
	LPR	41.50	0.7821	40.92	0.7743	40.65	0.7652	**40.46**	0.7640	41.01	0.7748
	LPRP	40.42	0.7643	40.37	0.7647	40.46	0.7670	**40.24**	0.7586	41.50	0.7806
6–6	MFCC	39.47	0.7412	38.93	0.7261	38.70	0.7292	37.85	0.7133	**36.62**	0.7088
	LPCC	39.74	0.7402	38.88	0.7322	37.26	0.7022	36.85	0.6975	**34.68**	0.6541
	LPR	39.38	0.7349	39.97	0.7484	**38.25**	0.7110	38.43	0.7199	38.52	0.7223
	LPRP	38.70	0.7300	38.66	0.7308	39.25	0.7421	**38.61**	0.7309	38.66	0.7309
9–9	MFCC	33.33	0.6280	31.61	0.6002	32.24	0.6035	32.56	0.6124	**28.13**	0.5295
	LPCC	30.89	0.5868	29.81	0.5668	29.22	0.5338	29.53	0.5608	**25.88**	0.5338
	LPR	35.59	0.6618	35.27	0.6515	33.73	0.6264	32.83	0.6164	**32.56**	0.6125
	LPRP	33.73	0.6300	**31.88**	0.5979	33.15	0.6279	33.78	0.6362	33.96	0.6372
12–12	MFCC	31.39	0.5929	31.02	0.5850	**27.04**	0.5300	32.06	0.6074	31.25	0.5914
	LPCC	28.86	0.5416	26.24	0.4985	25.38	0.4778	24.11	0.4555	**22.53**	0.4255
	LPR	33.55	0.6326	34.41	0.6399	**32.06**	0.5990	32.11	0.5851	33.19	0.5798
	LPRP	32.20	0.6032	31.43	0.5831	**32.06**	0.6054	32.33	0.6046	32.29	0.6067
15–15	MFCC	27.64	0.5374	27.23	0.5086	26.01	0.4922	25.68	0.4986	**23.21**	0.3717
	LPCC	20.68	0.3964	20.41	0.3852	19.01	0.3575	18.54	0.3214	**18.18**	0.3061
	LPR	30.39	0.5738	29.04	0.5662	28.31	0.5364	29.99	0.5626	**28.16**	0.5213
	LPRP	31.30	0.5926	31.25	0.5888	30.62	0.5746	29.10	0.5682	**28.31**	0.5364
15–20	MFCC	22.77	0.3774	22.54	0.3786	21.59	0.3612	**20.64**	0.3586	20.78	0.3537
	LPCC	18.45	0.3474	18.32	0.3432	17.34	0.3325	17.11	0.3314	**16.45**	0.3241
	LPR	27.33	0.5768	27.14	0.5732	26.54	0.5654	26.61	0.5638	**25.34**	0.5513
	LPRP	27.64	0.5716	27.42	0.5748	26.66	0.5616	26.78	0.5682	**25.63**	0.5364
15–25	MFCC	21.78	0.3632	20.64	0.3576	19.32	0.3434	18.63	0.3445	**17.32**	0.3476
	LPCC	16.35	0.3643	15.64	0.3552	15.32	0.3575	14.78	0.3234	**14.65**	0.3145
	LPR	25.33	0.5358	24.45	0.4472	23.31	0.4364	23.15	0.4326	**22.32**	0.4213
	LPRP	25.45	0.5396	24.78	0.4468	23.51	0.4346	23.32	0.4382	**22.45**	0.4264

The bold values represent minimum EER of particular feature extraction techniques.

Table 4: EER of the Speaker Verification System Using MFCC and LPCC Features (13 Dimensions) and GMM–UBM Modeling for the NIST 2003 Data Set.

Train/test data (s)	Feature extraction techniques	Gaussian mixtures				
		16	32	64	128	256
3–3	MFCC	41.32	40.15	**40.10**	40.19	40.37
	LPCC	40.01	39.79	39.11	39.25	**39.06**
6–6	MFCC	38.16	36.94	**36.67**	37.12	37.57
	LPCC	37.48	36.54	36.49	36.04	**35.63**
9–9	MFCC	32.47	31.02	30.21	**29.62**	30.26
	LPCC	28.99	**28.54**	28.68	29.17	28.54
12–12	MFCC	30.57	28.09	27.46	**27.19**	27.59
	LPCC	27.28	26.24	26.42	26.28	**25.38**
15–15	MFCC	28.64	27.32	26.49	24.33	**25.32**
	LPCC	26.54	25.34	23.23	22.64	**21.37**
15–20	MFCC	25.32	25.14	24.68	23.84	**23.64**
	LPCC	21.79	21.64	20.71	20.41	**19.64**
15–25	MFCC	22.44	22.32	21.78	21.46	**20.72**
	LPCC	20.45	20.28	19.64	19.32	**18.32**

The bold values represent minimum EER of particular feature extraction techniques.

The same set of experiments are conducted for 39 dimension features. Tables 5 and 6 represents performance of the speaker verification system using i-vector and GMM–UBM modeling, respectively. In both modeling techniques, LPCC performance is better than MFCC and also i-vector modeling gives better performance than GMM–UBM for all data sizes. Because of first- and second-order derivatives, 39 dimension features gives better performance than 13 dimension eatures in both modeling techniques.

The literature survey shows that speaker verification under limited data condition is widely used in security, controlled access, authentication of remote transactions, criminal and forensic investigations etc. In almost all these applications, the speech data may be limited (criminal may speak for only a few seconds). However, when the speech data are less, the speaker-specific information obtained is also less. The speaker

Table 5: EER of Speaker Verification Using the NIST 2003 database for Different Feature Extraction Techniques (39 Dimensions) and Modeling Done by i-Vectors.

Train/test data (s)	Feature extraction techniques	No. of dimension: i-vectors = 50; LDA = 50									
		16		32		64		128		256	
		EER%	DCT	EER%	DCT	EER%	DCT	EER%	DCT	EER%	DCT
3–3	MFCC	40.10	0.7591	40.42	0.7630	40.24	0.7594	39.92	0.7550	**39.83**	0.7548
	LPCC	39.20	0.7426	39.11	0.7397	39.47	0.7412	38.70	0.7292	**38.86**	0.7199
6–6	MFCC	38.88	0.7322	37.26	0.7022	36.78	0.6799	34.77	0.6575	**33.42**	0.6305
	LPCC	37.08	0.7018	37.18	0.7150	36.13	0.6799	34.77	0.6575	**33.42**	0.6305
9–9	MFCC	31.25	0.5891	31.43	0.5671	29.94	0.5650	29.04	0.5485	**26.64**	0.5210
	LPCC	28.00	0.5224	27.68	0.5124	26.87	0.5045	**25.11**	0.4749	25.38	0.4749
12–12	MFCC	30.66	0.5811	30.35	0.5677	30.26	0.5683	28.31	0.5364	**24.42**	0.5274
	LPCC	27.42	0.5086	26.78	0.4991	25.20	0.4768	24.57	0.4651	**20.41**	0.385
15–15	MFCC	26.34	0.5711	25.34	0.5307	25.17	0.5373	25.32	0.4964	**20.12**	0.3254
	LPCC	19.43	0.3273	18.31	0.3291	18.03	0.3205	17.34	0.3134	**16.68**	0.2450
15–20	MFCC	19.84	0.3254	19.32	0.3243	18.64	0.3223	18.52	0.3214	**18.35**	0.3204
	LPCC	17.64	0.3173	17.32	0.3151	17.18	0.3135	16.45	0.3034	**16.32**	0.3050
15–25	MFCC	18.32	0.3246	18.45	0.3257	17.41	0.3173	17.31	0.3164	**16.64**	0.3045
	LPCC	16.78	0.3073	16.64	0.3061	15.32	0.2965	15.11	0.2934	**14.86**	0.2850

The bold values represent minimum EER of particular feature extraction techniques.

Table 6: EER of the Speaker Verification System Using MFCC and LPCC Features (39 Dimensions) and GMM–UBM Modeling for the NIST 2003 Data Set.

Train/test data (s)	Feature extraction techniques	Gaussian mixtures				
		16	32	64	128	256
3–3	MFCC	39.70	39.02	**38.84**	39.15	39.02
	LPCC	39.61	39.15	**38.54**	38.66	38.70
6–6	MFCC	38.21	37.08	**36.22**	36.76	36.94
	LPCC	36.44	35.45	34.73	34.73	**34.28**
9–9	MFCC	28.95	27.59	27.23	**27.05**	27.14
	LPCC	28.13	27.32	27.05	**26.73**	27.95
12–12	MFCC	26.01	**24.66**	24.79	24.48	24.84
	LPCC	25.70	25.15	**23.71**	23.84	24.62
15–15	MFCC	24.32	23.64	23.68	22.78	**21.16**
	LPCC	24.13	22.77	22.32	21.94	**20.64**
15–20	MFCC	22.32	22.14	21.64	21.48	**19.86**
	LPCC	19.64	19.32	18.74	18.34	**17.64**
15–25	MFCC	19.84	19.32	18.24	18.11	**17.32**
	LPCC	17.64	17.32	16.62	16.12	**15.32**

The bold values represent minimum EER of particular feature extraction techniques.

verification performance can be improved by using feature level and score-level fusion. To study the effect of this on limited data the following experiments are conducted.

4.1 Feature-Level Fusion

The feature-level fusion is accomplished by a simple concatenation of the feature sets obtained by different feature extraction techniques. In our experiments, we fused vocal tract feature (system information) with excitation source features (source features). For instance, let $X = \{x_1, x_2, x_3, \ldots, x_m\}$ denotes vocal tract features

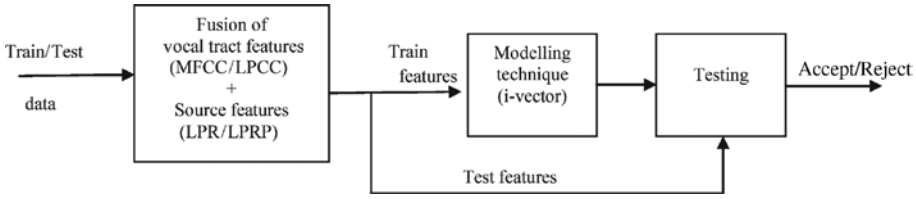

Figure 2: Block Diagram of Feature-Level Fusion.

Table 7: Results of the i-Vectors System Using Feature-Level Fusion of the Vocal Tract and Excitation Source Features (13 Dimension) for Limited Data Condition on NIST 2003 Database.

Train/test data (s)	Scores of vocal tract and excitation source features	No. of dimension: i-vectors = 50; LDA = 50									
		16		32		64		128		256	
		EER%	DCT	EER%	DCT	EER%	DCT	EER%	DCT	EER%	DCT
3–3	MFCC + LPR	**35.99**	0.6792	37.48	0.7017	36.90	0.6950	37.03	0.6991	38.34	0.7251
	MFCC + LPRP	41.37	0.7659	40.65	0.7642	41.10	0.7757	40.65	0.7642	**40.37**	0.7659
	LPCC + LPR	**35.32**	0.6669	36.22	0.6822	37.12	0.6990	36.31	0.6891	37.39	0.7035
	LPCC + LPRP	41.18	0.7696	41.32	0.7782	**39.79**	0.7515	40.24	0.7592	40.87	0.7696
6–6	MFCC + LPR	**34.46**	0.6512	35.09	0.6628	34.82	0.6567	35.68	0.6696	36.35	0.6745
	MFCC + LPRP	39.74	0.7420	38.52	0.7312	38.79	0.7328	**38.75**	0.7218	38.79	0.7358
	LPCC + LPR	34.68	0.6543	35.00	0.6588	34.64	0.6457	**34.55**	0.6525	34.91	0.6488
	LPCC + LPRP	39.52	0.7404	39.47	0.7471	37.80	0.7038	**37.48**	0.7066	38.03	0.7140
9–9	MFCC + LPR	28.45	0.5368	29.53	0.5486	**28.04**	0.5300	28.54	0.5613	30.17	0.5613
	MFCC + LPRP	33.69	0.6318	32.47	0.6126	33.46	0.6325	32.33	0.6069	**33.24**	0.6255
	LPCC + LPR	28.99	0.5487	28.99	0.5460	27.77	0.5255	28.22	0.5255	**25.15**	0.4675
	LPCC + LPRP	35.18	0.6588	34.77	0.6386	32.52	0.6139	**31.88**	0.5942	31.97	0.5972
12–12	MFCC + LPR	28.22	0.5306	28.13	0.5219	28.00	0.5223	**27.55**	0.5176	28.22	0.5311
	MFCC + LPRP	32.61	0.6030	32.02	0.6087	32.52	0.6087	**31.39**	0.5915	32.83	0.6175
	LPCC + LPR	28.41	0.5313	27.32	0.5043	27.32	0.5080	25.73	0.4958	**22.33**	0.4288
	LPCC + LPRP	33.46	0.6198	32.65	0.6079	**30.21**	0.5618	30.57	0.5724	30.39	0.5637
15–15	MFCC + LPR	24.64	0.5366	24.18	0.5309	24.34	0.5123	23.22	0.4176	**23.11**	0.4121
	MFCC + LPRP	24.84	0.4930	24.54	0.4987	24.94	0.4687	23.74	0.4615	**23.64**	0.4675
	LPCC + LPR	20.42	0.3113	19.32	0.3843	19.81	0.3280	18.24	0.3055	**17.32**	0.2134
	LPCC + LPRP	20.64	0.3698	20.24	0.3679	19.54	0.3021	**17.63**	0.2564	18.32	0.2837
15–20	MFCC + LPR	20.84	0.3696	**20.34**	0.3609	21.54	0.3723	21.64	0.3776	21.32	0.3721
	MFCC + LPRP	21.64	0.3730	**20.54**	0.3687	21.72	0.3687	21.32	0.3615	21.54	0.3675
	LPCC + LPR	16.84	0.2413	**15.24**	0.2343	15.34	0.2380	16.32	0.2455	16.14	0.2134
	LPCC + LPRP	17.35	0.2598	**15.44**	0.2379	15.64	0.2321	16.55	0.2564	16.28	0.2637
15–25	MFCC + LPR	19.44	0.3696	**19.32**	0.3609	19.45	0.3723	19.64	0.3776	20.32	0.3721
	MFCC + LPRP	19.72	0.3730	**19.64**	0.3687	19.82	0.3687	20.32	0.3615	20.54	0.3675
	LPCC + LPR	14.54	0.2413	**15.32**	0.2343	15.54	0.2380	15.62	0.2455	15.44	0.2134
	LPCC + LPRP	15.64	0.2368	**14.78**	0.2169	15.68	0.2381	15.92	0.2394	16.11	0.2547

The bold values represent minimum EER of particular feature extraction techniques.

(MFCC or LPCC) and $Y=\{y_1, y_2, y_3 \ldots y_n\}$ represents excitation source features (LPR or LPRP). In our experiments, we concatenated the both feature set to form $Z=\{x_1, x_2, x_3, \ldots, x_m, y_1, y_2, y_3, \ldots, y_n\}$. This new feature set Z is used for both training and testing. Figure 2 shows the block diagram of feature-level fusion. The features are fused on the frame-level by concatenating vocal tract features with excitation source features.

The individual EER of MFCC, LPCC, LPR and LPRP is 40.83%, 40.42%, 40.46% and 40.24%, respectively, as shown in Table 3. The results obtained for limited data by combining the features of vocal tract and excitation source are shown in Table 7. The minimum EER of feature-level fusion of MFCC + LPR, MFCC + LPRP, LPCC + LPR and LPCC + LPRP is 35.99%, 40.37%, 35.32% and 39.79% respectively. The fusion of MFCC + LPR and MFCC + LPRP is having reduction of 4.84% and 0.46% less in EER as compared to individual performance of MFCC. Similarly, the fusion of LPCC + LPR and LPCC + LPRP is having reduction in EER of 5.1% and 0.63% less as compared to LPCC. There is an improvement in the performance of EER using feature-level fusion, the combination of vocal tract and excitation source are having dissimilar features relatively improves the performance compared to all individual features. The important point noticed in feature-level fusion is that, fusion of MFCC + LPR and LPCC + LPR gives better performance as compared to fusion of MFCC + LPRP and LPCC + LPRP. This may be due to, LPR contains information obtained from excitation source mainly to glottal closure an instant (GCIS) [20]. Further, the fusion of LPCC + LPR performance is better compared to MFCC + LPR for limited data. Almost similar trend is observed for other data sizes except fusion of MFCC + LPRP and LPCC + LPRP for 9–9, 12–12, 15–15, 15–20 and 15–25-second data.

4.2 Score-Level Fusion

In case of score-level fusion, fuse the sores of individual system at verification level using the following equation:

$$S_{\text{total}} = \alpha S_1 + (1-\alpha) S_2, \tag{8}$$

where S_1 and S_2 represent the scores obtained using individual systems and S_{total} represents the fused scores of S_1 and S_2; α represents the optimal value of which is chosen for fusion of the two scores to give and it is a scalar between the value 0 and 1.

In this work, we conducted two types of score-level fusion to improve the performance of the speaker verification system. The first one is to fuse the scores of vocal tract and source excitation features, keeping the modeling techniques same. The second one is fusing the scores of different modeling techniques by maintaining the same feature extraction technique.

The score-level fusion of different feature extraction techniques is shown in Figure 3. The weight factor α becomes optimal for the value of 0.05 in case of fusion of scores of two different feature extraction techniques.

From the previous section, we already proved that feature-level fusion gives better performance compared to individual feature. The performance of score-level fusion of vocal tract and excitation source features for limited data using NIST 2003 data set is shown in Table 8. Further, the minimum EER of score-level

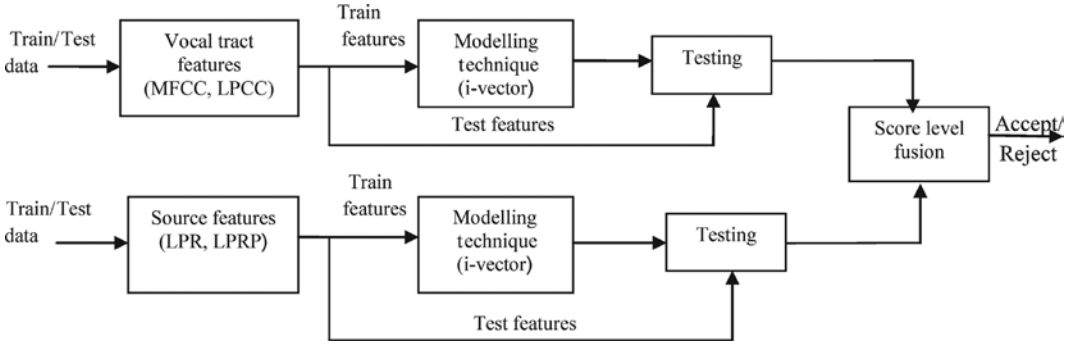

Figure 3: Block Diagram of Score-Level Fusion for Different Feature Extraction Techniques.

Table 8: Results of the i-Vectors System using Score-Level Fusion of the Vocal Tract and Excitation Source Features (13 Dimension) for Limited Data Condition on NIST 2003 Database.

Train/test data (s)	Scores of feature extracted signal	Score-level fusion (no. of dimension: i-vectors = 50; LDA = 50)				
		Gaussian mixtures				
		16	32	64	128	256
3–3	MFCC + LPR	**34.91**	35.09	35.63	36.81	36.92
	MFCC + LPRP	39.43	39.47	38.75	**38.12**	39.47
	LPCC + LPR	**34.82**	35.04	35.68	35.99	35.81
	LPCC + LPRP	39.34	40.06	39.47	**39.15**	39.83
6–6	MFCC + LPR	32.79	33.15	**32.74**	33.55	33.64
	MFCC + LPRP	36.58	36.81	36.94	**35.72**	36.49
	LPCC + LPR	**33.46**	33.92	33.64	35.09	34.37
	LPCC + LPRP	38.07	38.21	**36.49**	36.58	37.17
9–9	MFCC + LPR	26.64	25.92	25.20	**25.02**	26.28
	MFCC + LPRP	28.31	**28.09**	28.54	28.27	29.17
	LPCC + LPR	26.87	26.60	27.10	26.64	**25.15**
	LPCC + LPRP	31.57	31.97	30.84	**29.04**	30.44
12–12	MFCC + LPR	25.88	24.52	**24.48**	26.15	25.70
	MFCC + LPRP	28.68	**26.91**	27.00	29.31	28.50
	LPCC + LPR	26.15	25.92	25.79	26.01	**25.42**
	LPCC + LPRP	29.58	29.99	28.41	**27.82**	30.53
15–15	MFCC + LPR	23.98	23.64	23.32	**22.16**	22.71
	MFCC + LPRP	24.54	24.32	23.94	**22.54**	22.67
	LPCC + LPR	19.34	19.31	18.13	17.72	**16.34**
	LPCC + LPRP	19.49	19.64	18.43	17.49	**16.54**
15–20	MFCC + LPR	20.64	**19.32**	20.32	20.16	20.32
	MFCC + LPRP	20.82	**19.82**	20.54	20.31	20.24
	LPCC + LPR	**14.54**	14.64	15.32	15.84	15.15
	LPCC + LPRP	15.34	**14.78**	15.64	15.92	15.45
15–25	MFCC + LPR	16.48	16.54	**16.13**	16.98	16.74
	MFCC + LPRP	**16.34**	16.64	17.32	17.13	17.11
	LPCC + LPR	**12.34**	12.64	13.32	13.64	13.51
	LPCC + LPRP	13.51	13.64	**12.62**	12.82	12.87

The bold values represent minimum EER of particular feature extraction techniques.

fusion of MFCC + LPR, MFCC + LPRP, LPCC + LPR and LPCC + LPRP is 34.91%, 38.12%, 34.82% and 39.15% respectively. The score-level fusion of MFCC + LPR and MFCC + LPRP is having reduction of 1.08% and 2.25% less in EER as compared to feature-level fusion of MFCC + LPR and MFCC + LPRP, respectively. Similarly score-level fusion of LPCC + LPR and LPCC + LPRP is 0.5% and 0.64% less in reduction in EER as compared to feature-level fusion of LPCC + LPR and LPCC + LPRP, respectively. The trend in EER reduction remains same for other data sizes.

Over the last two decades, speaker verification system facing two main problems such as session variability and channel mismatch. The main reason for these problems is emotional state of the speaker, environmental conditions, recording devices, different transmission channels,...etc. Due to this variability the system performance decreases drastically. The GMM–UBM modeling system facing these problems [31].

In case of GMM–UBM, the feature is extracted by frame and number of features are unfixed. Gaussian Mixtures are used to fit all the features. In GMM mapping, the feature is calculated by frame with MAP. The difference of the likelihood ratio from the GMM to the UBM is used to describe the result. These entire problems can be overcome by using i-vector. The i-vector-based modeling is an advanced technique than GMM–UBM modeling. The i-vector extracted from the speech utterance contains both channel and speaker variabilities. Implementing channel/session compensation methods help improves the performance of the speaker verification system [10]. Due to these reasons, the experiments we conducted using i-vectors show an improvements in performance over GMM–UBM. The results are shown in Tables 3–6.

Figure 4 shows the proposed combined modeling technique in score-level fusion. The improvement in performance by combining the scores of i-vector and GMM–UBM modeling techniques at score-level are significant as shown in Table 9. The reason for improvement in performance may be due to i-vectors use cosine kernel whereas GMM–UBM uses log likelihood ratio test and the working principle of i-vector and GMM–UBM are different.

In the proposed system, the optimal value of α is 0.5. The performance of score-level fusion for different modeling techniques for 13 dimensions. The LPCC gives very good performance in all data sizes compared to MFCC. The LPCC is having minimum reduction in EER of $\approx 4\%$ as compared to MFCC in score-level fusion of different modeling techniques.

Table 10 shows comparision for different fusion techniques for 13 dimension features. From this table we observed few points. First point is, for 3–3 second data, feature-level fusion and score-level fusion of different features gives minimum EER as compared to score-level fusion of different modeling techniques. In this case combination of MFCC or LPCC with LPR gives better performance than MFCC or LPCC with LPRP. The second point is, when data size increases (6–6, 9–9, 12–12, 15–15, 15–20 and 15–25 s) score-level fusion for different modeling techniques gives better performance compared to other fusion techniques. From these observation it is clear that, for limited data score-level fusion for different modeling technique can be used to get better verification results. This results motivates us to conduct score-level fusion for 39 dimensions. Table 11 shows the score-level fusion of 39 dimensions for different modeling techniques. In this case, LPCC also performs better than MFCC for all data sizes. Compare Tables 9 and 11's score-level fusion for both dimensions. The results shows that there is drastic improvement in 39 dimensions compared with 13 dimensions.

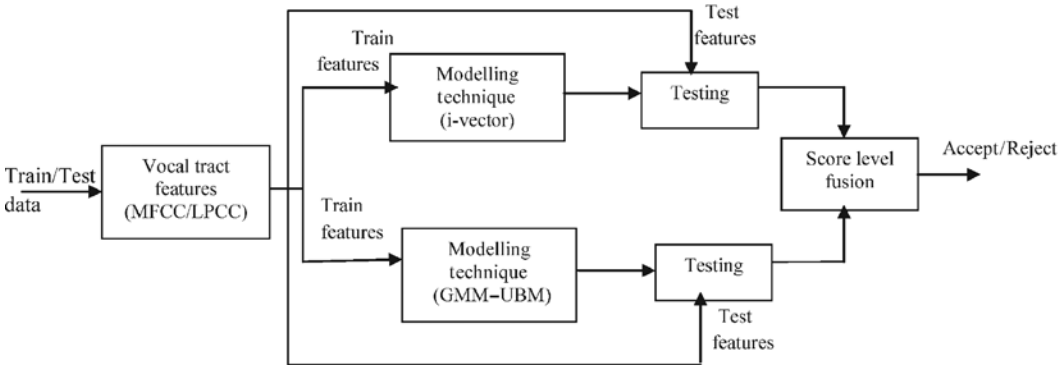

Figure 4: Block Diagram of Proposed Combined Modeling.

Table 9: Performance of the Speaker Verification System for Score-Level Fusion Using i-Vector and GMM–UBM Modeling Techniques for Limited Data Condition in the NIST 2003 Database.

Train/test data (s)	Score-level fusion for MFCC features (13 dimensions)					Score-level fusion for LPCC features (13 dimensions)				
	Gaussian mixtures					Gaussian mixtures				
	16	32	64	128	256	16	32	64	128	256
3–3	38.73	37.28	37.74	36.65	**36.06**	37.35	37.75	36.90	36.63	**35.99**
6–6	36.72	36.08	36.63	35.90	**32.55**	34.77	33.69	33.60	32.70	**32.15**
9–9	29.72	28.45	28.13	30.08	**25.00**	24.11	23.44	23.89	23.75	**23.30**
12–12	27.37	26.24	**23.80**	28.27	28.00	**20.09**	21.45	21.95	21.45	20.09
15–15	21.64	22.32	19.18	18.92	**17.64**	16.11	15.24	15.34	**14.52**	14.63
15–20	14.43	14.34	14.64	14.36	14.52	11.34	11.41	11.14	11.54	11.64
15–25	11.17	11.34	11.32	11.64	11.54	8.38	8.34	8.54	8.51	8.49

The bold values represent minimum EER of particular feature extraction techniques.

Table 10: Minimum EER of Different Fusion Techniques.

Train/test data (s)	Feature-level fusion				Score-level fusion for different features				Score-level fusion for different modeling techniques	
	MFCC+LPR	MFCC+LPRP	LPCC+LPR	LPCC+LPRP	MFCC+LPR	MFCC+LPRP	LPCC+LPR	LPCC+LPRP	MFCC	LPCC
3–3	35.99	40.37	35.32	39.79	34.91	38.12	34.82	39.15	36.06	35.99
6–6	34.46	38.52	34.55	37.48	32.74	35.72	33.46	36.49	32.55	32.15
9–9	28.04	32.33	27.15	31.88	25.02	28.27	25.15	29.04	25.00	23.30
12–12	27.55	32.02	26.33	30.21	24.48	26.91	25.42	27.82	23.80	20.09
15–15	23.11	23.64	17.32	17.52	22.16	22.64	16.34	16.54	17.64	14.52
15–20	20.34	20.54	15.24	15.64	19.32	19.82	14.54	14.78	14.34	11.14
15–25	19.32	19.64	14.54	14.78	16.13	16.14	12.34	12.62	11.17	8.34

Table 11: Performance of the Speaker Verification System for Score-Level Fusion Using i-Vector and GMM–UBM Modeling Techniques for Limited Data Condition on the NIST 2003 Database.

Train/test data (s)	Score-level fusion for MFCC features (39 dimensions)					Score-level fusion for LPCC features (39 dimensions)				
	Gaussian mixtures					Gaussian mixtures				
	16	32	64	128	256	16	32	64	128	256
3–3	36.20	36.11	35.70	**34.16**	38.93	34.73	33.73	33.69	31.97	**31.48**
6–6	33.19	33.42	33.46	**30.62**	30.84	29.26	27.95	28.90	**27.10**	27.59
9–9	27.82	27.23	25.42	24.61	**24.57**	21.09	**20.05**	22.94	22.71	21.34
12–12	25.38	24.11	23.75	22.94	**22.17**	19.28	19.20	18.15	**17.34**	19.41
15–15	17.49	16.34	16.51	**15.49**	15.63	15.32	15.64	14.34	13.21	**12.36**
15–20	11.69	10.84	9.48	8.64	**8.12**	9.14	8.34	6.34	6.23	**6.04**
15–25	5.65	5.84	5.82	5.12	**5.05**	4.32	4.64	4.12	**3.58**	3.98

The bold values represent minimum EER of particular feature extraction techniques.

5 Conclusion

In this article, we demonstrated the significance of performance of individual modeling technique and different fusion techniques for limited data condition. First, we studied the working principles of individual features using i-vector modeling technique. It was observed that i-vector modeling gives better EER compared with the GMM–UBM modeling technique. To increase the performance of the speaker verification system, we conducted experiments using feature- and score-level fusion for 13 dimensions. Here the vocal tract features are fused with excitation source features in feature-level fusion, the performance of feature-level fusion is better as compared to performance of individual feature extraction techniques. The two cases of score-level fusions are demonstrated in this work for 13 dimensions. In the first case, fusing the scores of vocal tract and excitation source features at score-level by maintaining same modeling technique. It was observed that, an average reduction of EER is approximately equal to 2% compared with feature-level fusion performance. In the second case, the different modeling scores are fused by keeping feature remain same. In the experimental results, it was observed that an average reduction in EER is approximately equal to 4.5% compared with score-level fusion of different features. Further it was observed that, score-level fusion for different modeling technique gives better performance compared to the other fusion techniques. It was also observed that, LPCC with source features combinations gives better performance as compared to MFCC + LPR and MFCC + LPRP under limited data condition. Therefore, we suggest that the LPCC with source features can be used as features along with score-level fusion for different modeling to improve the performance of speaker verification under limited data condition. Also we observed that score-level fusion for different modeling technique provides better performance than other two fusion techniques. With this movitation we conducted score-level fusion for 39 dimensions. The result shows that there is drastic improvement in EER using 39 dimensions compared with 13 dimensions. Therefore, we also suggest that 39 dimensions can also be used to improve the performance of the speaker verification system under limited data.

Bibliography

[1] A. K. H. Al-Ali, D. Dean, B. Senadji, V. Chandran and G. R. Naik, Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions, *IEEE Access* **5** (2017), 15400–15413.

[2] B. S. Atal, Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, *J. Acoust. Soc. Am.* **55** (1974), 1304–1312.

[3] R. Auckenthaler, M. Carey and H. Lloyd-Thomas, Score normalization for text-independent speaker verification systems, *Digit. Signal Process.* **10** (2000), 42–54.

[4] F. Bimbot, J. Bonastreand, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz and D. A. Reynolds, A tutorial on text-independent speaker verification, *EURASIP J. Appl. Signal Process.* **4** (2004), 430–451.
[5] J. P. Campbell, Speaker recognition: a tutorial, *Proc. IEEE* **85** (1997), 1437–1462.
[6] W. Chan, N. Zheng and T. Le, Discrimination power of vocal source and vocal tract related features for speaker segmentation, *IEEE Trans. Audio Speech Lang. Process.* **15** (2007), 1884–1892.
[7] R. K. Das, D. Pati and S. R. M. Prasanna, Different aspects of source information for limited data speaker verification, *Proc. 21st National Conference on Communications (NCC), 2015 Twenty First National Conference on,* Mumbai, pp. 1–6, 2015.
[8] R. K. Das, S. Jelil and S. R. M. Prasanna, Significance of constraining text in limited data text-independent speaker verification, in: *Signal Processing and Communications (SPCOM), 2016 International Conference on*, Bangalore, pp. 1–5, 2016.
[9] R. K. Das, S. Jelil and S. R. M. Prasanna, Development of multi-level speech based person authentication system, *J. Signal Process. Syst.* **88** (2016), 259–271.
[10] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, Front-end factor analysis for speaker verification, *IEEE Trans. Audio Speech Lang. Process.* **19** (2011), 788–798.
[11] 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, *National Conference on Communications (NCC)*, pp. 1–6, 2014.
[12] F. Farahani, P. G. Georgiou and S. S. Narayanan, Speaker identification using supra-segmental pitch pattern dynamics, *Proc. Int. Conf. Accous. Speech Signal Process* (May. 2004), Montreal, pp. 89–92.
[13] D. Garcia-Romero and A. McCree, Supervised domain adaptation for i-vector based speaker recognition, *Acoustics, speech and signal processing (ICASSP), 2014 IEEE International conference*, pp. 4047–4051, 2014.
[14] J. Gudnason and M. Brookes, Voice source cepstrum coefficients for speaker identification, *in Proc. ICASSP* (2008), Las Vegas, pp. 4821–4824.
[15] B. C. Haris, G. Pradhan, A. Misra, S. Shukla, R. Sinha and S. R. M. Prasanna, Multi-variability speech database for robust speaker recognition, in: *Communications (NCC), 2011 National Conference on*, Bangalore, pp. 1–5, 2011.
[16] A. O. Hatch, S. S. Kajarekar and A. Stolcke, Within-class covariance normalization for SVM-based speaker recognition, in: *Interspeech*, Pittsburgh, 2006.
[17] 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, *Multimedia workshops, 2007. ISMW'07. Ninth IEEE International Symposium*, pp. 246–251, 2007.
[18] A. K. Jain, A. Ross, and S. Prabhakar, An introduction to biometric recognition, *IEEE Trans. Circuit Syst. Video Technol.(Special Issue on Image and Video-Based Biometrices)* **14** (2004), 4–20.
[19] H. S. Jayanna and S. R. Prasanna, Analysis, feature extraction, modeling and testing techniques for speaker recognition, *IETE Tech. Rev.* **26** (2009), 181–190.
[20] H. S. Jayanna and S. R. M. Prasanna, Limited data speaker identification, *Sadhana* **35** (2010), 525–546.
[21] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan and M. W. Mason, I-vector based speaker recognition on short utterances, *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, pp. 2341–2344, 2011.
[22] T. R. J. Kumari and H. S. Jayanna, Comparison of LPCC and MFCC features and GMM and GMM–UBM modeling for limited data speaker verification, *IEEE Proc. ICCIC 2014* (2014), 1–6.
[23] N. Li, M.-W. Mak and J.-T. Chien, DNN-driven mixture of PLDA for robust speaker verification, *IEEE/ACM Trans. Audio Speech Lang. Process.* **25** (2017), 1371–1383.
[24] K. S. R. Murty and B. Yegnanarayana, Combining evidence from residual phase and MFCC features for speaker recognition, *Signal Process. Lett. IEEE* **13** (2006), 52–55.
[25] NIST2003, http://www.itl.nist.gov/iad/mig//tests/sre/2003/2003-spkrec-evalplan-v2.2.pdf[online].
[26] A. Pandey, R. K. Das, N. Adiga, N. Gupta and S. R. M. Prasanna, Significance of glottal activity detection for speaker verification in degraded and limited data condition, in: *TENCON 2015–2015 IEEE Region 10 Conference*, Macao, pp. 1–6, 2015.
[27] S. R. M. Prasanna, C. S. Gupta and B. Yegnanarayana, Extraction of speaker-specific excitation information from linear prediction residual of speech, *Speech Commun.* **48** (2006), 1243–1261.
[28] D. A. Reynolds, Experimental evaluation of features for robust speaker identification, *IEEE Trans. Acoust. Speech Signal Process.* **2** (1994), 639–643.
[29] D. A. Reynolds, Speaker identification and verification using Gaussian mixture speaker models, *Speech Commun.* **17** (1995), 91–108.
[30] D. A. Reynolds and R. C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, *IEEE Trans. Speech Audio Process.* **3** (1995), 72–83.
[31] D. Reynolds, T. Quatieri and R. Dunn, Speaker verification using adapted Gaussian mixture models, *Digital Signal Process.* **10** (2000), 19–41.
[32] A. E. Rosenberg, Automatic speaker verification: a review, *Proc. IEEE* **64** (1976), 475–487.

[33] A. Salman, E. Muhammad and K. Khurshid, Speaker verification using boosted cepstral features with Gaussian distributions, in: *Multitopic Conference, 2007. INMIC 2007. IEEE International*, Lahore, pp. 1–5, 2007.
[34] G. L. Sarada, N. Hemalatha, T. Nagarajan and H. A. Murthy, Automatic transcription of continuous speech using unsupervised and incremental training, in: *Proceedings of Interspeech*, Jeju Island, pp. 405–408, 2004.
[35] F. K. Soong and A. E. Rosenberg, On the use of instantaneous and transitional spectral information in speaker recognition, *IEEE Trans. Acoust. Speech Signal Process.* **36** (1988), 871–879.
[36] S. C. Yin, R. Rose and P. Kenny, A joint factor analysis approach to progressive model adaptation in text-independent speaker verification, *IEEE Trans. Audio Speech Lang. Process.* **15** (2007), 1999–2010.