International Journal of Speech Technology (2018) 21:1057–1070
https://doi.org/10.1007/s10772-018-09568-4

Effective use of combined excitation source and vocal-tract information for speaker recognition tasks

Krishna Dutta[1] · Jagabandhu Mishra[1] · Debadatta Pati[1]

Received: 11 August 2018 / Accepted: 13 October 2018 / Published online: 29 October 2018

Abstract

In automatic speaker recognition (SR) tasks the widely used score level combination scheme derives a general consensus from the independent opinions of individual evidences. Instead, we conjecture that collectively contributed decisions may be more effective. Based on this idea this work proposes an effective combination scheme, where the vocal-tract and excitation source information take decisions collectively, resulting higher improvements in SR accuracy. In the proposed scheme, independently made feature-specific models are padded for building resultant models. While testing, feature-specific test features are padded in similar fashion, and then used for comparison with resultant models. The main advantage of this proposed scheme is that it does not require any ground truth information for combined use of multiple evidences like in score level combination scheme. The potential of the proposed scheme is experimentally demonstrated by conducting different speaker recognition experiments in clean and noisy conditions, and also comparative studies with score level fusion scheme as reference. The TIMIT database is used for studies with clean case, and Indian Institute of Technology Guwahati Multi-Variability (IITG-MV) databases for noisy case. In clean case the proposed scheme provides relatively 1% of higher improvements in performance for GMM based speaker identification system and 8.5% for GMM–UBM based speaker verification system. In noisy case the corresponding parameters are 1% and 3%, respectively. The final evaluations on NIST-2003 database with GMM–UBM and i-vector based systems show relatively higher improvements in performance by 5.17% and 4.73%, respectively. The proposed scheme is observed to be statistically more significant than the commonly used score level fusion of multiple evidences.

Keywords Vocal-tract and excitation source information · Score level fusion · Gaussian mixtures model (GMM) · Gaussian mixtures model–universal background model (GMM–UBM) · I-vectors

1 Introduction

Automatic speaker recognition (SR) is a task of recognizing people based on the information available in their speech samples by machine. The speaker recognition tasks are classified as: automatic speaker identification (SI) and automatic speaker verification (SV). In case of SI task, the machine has to decide the identity of the speaker from the input test speech samples. In case of SV task the machine has to authenticate the claimed identity presented through his/her speech samples. Based on text mode, the SR tasks can also be classified as text-dependent (TD) and text-independent (TI). In case of TD mode, the test speakers are required to speak the same speech samples as in while their enrollment. There is no such constraint in TI mode. The SR system with TI mode is commonly preferred in real-time use.

The SR process is performed in two phases: training phase and testing phase (Campbell 1997). In training phase, the machine collects the speech samples from the known speakers (targets) and enroll them by using SR algorithms that involve suitable speaker-specific feature(s) extraction and modelling techniques. In testing phase, the machine computes the speaker-specific features from the test speech samples by using similar feature extraction algorithms

✉ Krishna Dutta
krishnadutta54@gmail.com

Jagabandhu Mishra
jagabandhumishrabcet@gmail.com

Debadatta Pati
debapati2003@yahoo.com

[1] Department of Electronics and Communication Engineering, National Institute of Technology Nagaland, Dimapur 797103, India

while enrollment, compares with the enrolled model(s) and provides matching score(s). State-of-the-art SR systems dominantly use probabilistic approaches like, i.e. Gaussian mixtures model (GMM), GMM universal background model (GMM–UBM) and i-vectors (Reynolds 1995; Dehak et al. 2011). In probabilistic approaches the matching scores are computed in the form of likelihood ratio (LLR) (Reynolds 1995). Accordingly, in case of SI tasks the model corresponds to maximum LLR is decided as the identified speaker and the performance of the machine is evaluated in terms of identification accuracy (%). In SV task the machine has to accept or reject the claimed identity by comparing the corresponding matching score with the pre-defined threshold. If the matching score is greater than or equal to the threshold, then the claimed identity is accepted else rejected. The machine may fails either by falsely rejecting a genuine speaker or falsely accepting an impostor. As such, the SV system performance is shown in terms of false acceptance rate (FAR) and false rejection rate (FRR). Ideally, both FAR and FRR should be as small as possible. A common evaluation measurement parameter is equal error rate (EER), where FAR = FRR (Wong and Russell 2001). A SI system with higher identification accuracy and SV system with lower EER are considered as efficient machines.

The SR system performance largely depends upon the discriminating ability of the speaker-specific features. These features reflect the speaker related variations in the speech signal, that caused in part due to the physiological differences of the vocal system and in part due to the differences in speaking habits of individuals (Atal 1976). The physiological aspects include the differences in vocal-tract and vocal folds structures and the behavioral part includes how the speaker is learnt to use his/her speech mechanism. The former is mostly used in SR tasks. The reasons may be two folds: First the behavioral aspects (like speaking style, traits and others) are highly variant even within speakers and also relatively more difficult for automatic extraction. On the other hand the physiological information is relatively more robust and easier for automatic estimation. Further, due to the availability of suitable signal processing tools, state-of-the-art SR systems dominantly use vocal-tract related information. For example, the linear prediction cepstral coefficients (LPCC) or mel-frequency cepstral coefficients (MFCC) features are popularly used to represent the vocal-tract related information in compact form (Reynolds 1995; Dehak et al. 2011; Campbell et al. 2006). The later is more popular, because the computational procedure follows perceptual characteristic of the speaker, and thereby provides relatively better performance (Hermansky and Morgan 1994). In last few years, extensive efforts able to achieve tremendous improvements in the recognition accuracy, particularly at channel/handset mismatch conditions. For example, state-of-the-art i-vectors based techniques able to achieve $< 2\%$ EER on the latest challenging NIST data, recorded under real telephone and microphone situations (Beigi 2011). The i-vectors based methods have mostly been applied to counter channel/handset mismatches but not specifically for additive background noise happen to be in real-time scenario. The cepstral features are representation of overall spectral characteristics. The spectral contents are corrupted with varying environmental conditions and mismatched channels, resulting degradation in the recognition performance (Reynolds 1995). Further, cepstral features based models are primarily phonetic in nature, differentiating speakers by characterization on pronunciation patterns (Atal 1976). This requires large amount of data to cover speaker's entire phonetic space and higher model complexity to cover that space. In Das and Prasanna (2016), it is shown that the i-vector based systems are highly dependent on the amount of speech data used. The quality and quantity of data remain an issue with state-of-the-art SR systems, encouraging researchers for the use of supplementary evidences (Mashao and Skosan 2006).

It is widely acceptable fact that humans use excitation characteristics like pitch, intonation and duration to recognize speakers (Pati and Prasanna 2011). Also, human listeners have proved themselves as robust speaker recognizers even in degraded speech and session variability conditions, indicating the robustness of excitation source information (Feustel et al. 1988). We may agree from our day-to-day experience that humans can easily recognize speakers by listening their few words. So, in terms of quality and quantity of data, the excitation source information is relatively more robust and effective in recognizing speakers. However, the lack of accurate estimation approaches discourage the research community from exploring excitation source features independently, but as a supplementary evidence to enhance the robustness of the state-of-the-art SR systems (Yegnanarayana et al. 2001). In that direction several studies have been made and available in (Mashao and Skosan 2006; Murty and Yegnanarayana 2006; Nakagawa et al. 2012). They have unanimously reported that the combined use of vocal-tract and excitation source information improves the performance and robustness of the state-of-the-art SR systems. The optimum benefit can be achieved by using suitable combination schemes. The present work deals with developing effective combination scheme for the combined use of vocal-tract and excitation source information in the context of SR tasks.

The rest of the paper is organized as follows: Sect. 2 provides the brief review about the existing combination schemes and motivation for the present work. The selection of suitable vocal-tract and excitation source features is described in Sect. 3. Section 4 describes the detail about the proposed combination scheme. The databases used for experimental studies are described in Sect. 5.

The experimental studies and discussions are reported in Sect. 6. The statistical significance of the proposed method is investigated in Sect. 7. The summary and future scopes of this work are reported in Sect. 8.

2 Combination schemes: a review

It is a widely acceptable fact that the combined use of multiple evidences leads to more accurate and robust recognition results. In speaker recognition tasks the combined use of multiple evidences is made either at feature level, or at model level, and/or at decision level. Accordingly, the combination methods can be categorized as *feature level*, *model level* and *decision level* combination schemes. A brief review on all these schemes is presented below:

2.1 Feature level combination scheme

The idea behind feature level combination scheme is that each feature may contain some aspect of the speaker information that might be missed by other. In this scheme, multiple speaker-specific features are concatenated for building models and comparisons. To the best of the authors knowledge Prof. Furui first employed the feature concatenation approach for joint use of cepstral features and their polynomial in the form of delta and delta delta coefficients (Furui 1981). As a result, the individual cepstral features error rate is reduced by a factor of three. Later such representation becomes very popular and widely used for speaker recognition tasks. The concatenation of excitation source information based spectral features with MFCC improves the performance of the later by 4.03% (Hosseinzadeh and Krishnan 2007). In Nakagawa et al. (2012), it is shown that the concatenation of phase information with MFCC improves the latter's individual performance. In another work, the authors jointly used statistical pH features and MFCC features by concatenation process and achieved improved performance in non-stationary noisy conditions (Venturini et al. 2014). All these works show that concatenation of multiple features helps in improving the individual performance. In concatenation approach the major concern is with the dimension of the resultant feature vectors. In general cepstral features with their delta information are higher dimensional. In addition the concatenation of other features further increases the dimension. The increased dimension suffers from the *curse of dimensionality* (Duda et al. 1973). The increased redundancy often creates confusions, resulting degradation in recognition accuracy. Thus, feature level combination schemes are limited to low dimensional feature vectors.

2.2 Model level combination scheme

Model level combination scheme refers to the use of different modelling techniques with a single features set. The motivation for using model level combination scheme is to capture the different aspects of a feature. For example, Gaussian mixtures model (GMM) and hidden markov model (HMM) capture the statistics about the probability distributions of a specific feature vectors set. On the other hand the vector quantization (VQ) and dynamic time warping (DTW) methods provide similarity measurements of the feature vectors during training and testing. The discriminant trained classifier, such as neural networks (NN) capture the difference between feature vectors for a target speaker and those of non-target speakers. Thus, different classes of modelling approaches capture different characteristics of the feature vector. In Altnccay and Demirekler (2003), it is shown that when all these approaches are evaluated on identical task, the errors are tend to be uncorrelated, indicating a chance of improving the performance by their joint use. In that direction several approaches have been applied for speaker recognition tasks. In Farrell et al. (1994), the combination of VQ and NN was employed for text-independent speaker identification and verification tasks. The distortion scores from VQ were converted to probabilities by logistics exponential equation and then combined with the NN scores, resulting improvements in the performance than the individuals for either task. In another case, NN and GMM were trained for each sub-word within a password for text-dependent SV task and achieved improved performance (Ramachandran et al. 2002). A hybrid GMM/SVM system that appropriately incorporates the individual advantages was proposed for text-independent speaker identification tasks (Djemili et al. 2007). The SVM was used to reduce the whole speaker's space into small subset and then GMM is employed for building models, resulting 50% improvement in the performance.

The outcomes of these works shows the benefit of using model level combination scheme, but less often used in speaker recognition task. It is mainly due to the difficulty arises in combination of different model parameters. The distortion based modelling that include VQ, GMM and HMM provides within class information. In contrast the discriminant trained model like NN provides between class information. In combining different model parameters we need suitable transformation or mapping approaches. Further, the combination of different modelling techniques with single feature often render redundant information. Because, any modelling technique (irrespective of the type) reflects the overall features relative information nearly in complete and compact manners. The combination of model specific information more likely to overlap in the speaker's space.

2.3 Decision level combination scheme

Decision level or score level combination scheme refers to deriving consensual decision from the opinions of multiple independent systems. The consensus is made based on the type of decision forms, such as opinion score, class labels or fuzzy. In present scenario, probabilistic modelling techniques are extensively used for speaker recognition tasks. These probabilistic approaches provide decisions in the form of opinion scores, where general consensus is made by opinion scores pooling, with an objective that the errors of one system are corrected by other and vice versa (Mashao and Skosan 2006). Because of simplicity score level combination scheme is very much popular in speech and speaker recognition tasks including other pattern recognition tasks. The opinion scores pooling follows either linear or logarithmic arithmetic operations (Ramachandran et al. 2002). The linear opinion pool is evaluated by linear weighted sum of the individual opinion scores and comparatively effective. In the context of speaker recognition task, the interested readers can found the benefit of using score level fusion scheme in Mashao and Skosan (2006), Altnccay and Demirekler (2000), Yegnanarayana et al. (2005) and Poh and Kittler (2008). The major drawbacks of opinion score pooling are two folded. If all systems make the same error then there is no benefit in combination. Also, score pooling scheme performs well with multiple evidences of extremely complimentary in nature (Pati and Prasanna 2013). The other one is in assigning the weighting factors. They are generally considered as the respective individual recognition performance which may not be possibly available a priori in real-time situations.

To summarize, the following points can be outlined from previous works:

- The joint use of multiple evidences helps in improving the speaker recognition performance.
- Multiple evidences can be jointly used at feature, model or at decision levels by various combination schemes.
- In feature level combination scheme the multiple evidences jointly take the decision. However, it's use is limited to low-dimensional feature vectors due to suffering from curse of dimensionality.
- In model and decision level combination schemes the multiple evidences take decisions independently and later build a consensus.
- Because of simplicity and easy implementation the decision or score level combination scheme is popular used in speaker recognition tasks.
- The decision or score level combination scheme is quite efficient, but requires a priori information about the classifiers.

The available combination schemes have their own merits and demerits. We aim to build a composite combination scheme by exploiting the merits of all existing combination schemes for speaker recognition tasks.

3 Selection of features

We need to use the best possible representation of joint vocal-tract and excitation source information for achieving higher benefits. Because of computational simplicity and performance, standard cepstral features (preferably MFCC) are widely used as the representation of vocal-tract information. But, the excitation source reflects various information and no single representative feature is available for complete representation. In this section we investigate and select the suitable excitation source information based feature, particularly in the context of using as complementary evidences for speaker recognition tasks.

The excitation source information based speaker-specific features are commonly derived either by using *explicit* or *implicit* approaches (Pati and Prasanna 2013). The explicit approaches categorically derive specific excitation parameters and use them for speaker recognition tasks. The interested readers can find a brief summary on those attempts in Pati and Prasanna (2010). The explicit approaches are successful but with smaller dataset. Because, excitation source parameters are highly intra-variant and corresponding representative feature vectors are too much overlapping in speakers' feature space. In addition, the parameters measurement accuracy highly rely on the accurate estimation of the excitation signal. Due to dynamic nature it is difficult to obtain a precise measurement of the excitation signal. On the other hand, implicit approaches use parametric representation of the excitation signal (preferably derived by inverse filtering) for speaker recognition tasks. The advantages over explicit approaches are at least three folded: (i) The implicit approaches mostly use approximate estimation of the excitation signal, like inverse filtering through linear prediction (LP) analysis, (ii) The parameterization process represents the feature vectors in compact form and thereby reduces the computational complexity, and (iii) As a complementary evidence the implicit based features perform comparatively well with conventional MFCC features even in larger and limited data conditions (Das and Prasanna 2016). Thus, we prefer to explore the implicit processing of the excitation source information.

In implicit approaches the LP residual signal is commonly used as the representation of the excitation signal (Prasanna et al. 2006). The LP residual based mel-wrapped power differences in subband spectra (MPDSS), residual mel-frequency cepstral coefficients (RMFCC), and very recently discrete cosine transform of the integrated

LP residual (DCTILPR) features are dominantly used for speaker recognition tasks (Pati and Prasanna 2011a, 2013b), Ramakrishnan et al. (2015). The mathematical expressions of these features are given in Eqs. 1, 2 and 3, where $R(k)$ is the Fourier transform of the LP residual signal $r(n)$.

$$MPDSS(m) = 1 - \frac{\left[\prod_{k=l_m}^{h_m} M_l(|R(k)|)\right]^{\frac{1}{N_m}}}{\frac{1}{N_m}\sum_{k=l_m}^{h_m} M_l(|R(k)|)} \tag{1}$$

$$RMFCC(c) = IDFT[M_l(log|R(k)|)] \tag{2}$$

$$DCTILPR(c) = \sum_{n=0}^{N-1} i_r(n)cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)c\right] \tag{3}$$

In Eq. 1, $N_m = h_m - l_m + 1$ is the total number of samples, lm and hm denoting the first and last sample of the sub-band in the mth filter. $M_l(|R(k)|)$ corresponds to the mel scale representation of the power of the kth sample of the sub-band. In Eq. 2, $M_l(log|R(k)|)$ denote the mel scale representation of the log magnitude spectrum of the LP residual signal $r(n)$. In Eq. 3, $i_r(n)$ represents integrated linear prediction residual (ILPR) corresponding to the LP residual $r(n)$ extracted between epoch locations j and (j + 1), N is the number of samples between those epoch locations and $c = 0, 1, 2, ..., N - 1$.

The MPDSS features represent the periodicity nature of the excitation signal. The RMFCC and DCTILPR features represent excitation sub-bands energy contours computed through mel and cosine filters bands. The later has the advantage of using the integrated residual signal that corresponds to pitch synchronized analysis. The effectiveness of these features have been demonstrated with different datasets in different conditions. For example, the DCTILPR features are found to be effective on limited dataset, where as the RMFCC and MPDSS features are useful for both large and limited datasets. We conduct SV experiments in a common platform and select the best representation of excitation source information based residual feature for our investigation.

The speaker verification experiments are conducted with GMM–UBM based system by using NIST-2003 database (The 2003 Nist speaker recognition evaluation plan 2003). The details about NIST-2003 database are given in Sect. 5. The speech signals are processed at 8000 samples/s and voice/unvoiced detections are made by using energy based thresholding. The features are computed from 20 ms overlapping voiced speech frames at the rate of 100 frames/s. There are 20 mel filters are used for computation of MPDSS feature. The first 13 mel-cepstral coefficients (excluding c_0) concatenated with corresponding thirteen delta and delta-delta coefficients are used for representation of MFCC and RMFCC features. The zero mean unit variance normalization is employed to avoid the effect of channel variability (Reynolds et al. 2000). In order to achieve the best possible results the experiments are conducted with different Gaussian mixtures size. In all cases the optimum size is found to be 1024, and the corresponding results are given in Table 1. RMFCC features provide the best recognition accuracy of 18.89%, as compared to 20.93% and 21.38% by DCTILPR and MPDSS features. The MFCC features based baseline system provides 7.54%. By using score level fusion scheme, the RMFCC feature combined with MFCC provides the best performance of 7.30%. Thus, we consider RMFCC as the excitation source information based representative feature for our further studies.

4 Proposed combination scheme

In this scheme the multiple features sets are processed separately and respective models are built independently by using a particular modelling technique. The feature specific models are represented by their respective modelling parameters. The proposed combination scheme generates composite models by padding the parameters of respective feature specific models. This padding process does not increase the dimension of the composite model parameters and thereby reduces the computational complexity. In addition the difficulty of mapping different modelling parameters (for combination) is avoided by use of common modelling technique. While testing, the individual test feature vectors set in similar fashion and placed before the system for evaluation. With reference to the existing score level fusion the proposed combination scheme is differed in the sense that, the common decisions are made in together rather than a consensus from individual opinions. Unlike score level fusion, the scores generated by the proposed scheme are directly used for comparison without assigning any weights. Thus, the proposed scheme does not require the ground truth information, that makes the system more suitable for real time applications.

Table 1 Performance of MPDSS, RMFCC, DCTILPR, MFCC features and their fused representation with GMM–UBM based SV system for NIST-2003 database

Feature	EER (%)
MPDSS	21.38
RMFCC	**18.89**
DCTILPR	20.93
MFCC	7.54
MFCC + MPDSS	7.54
MFCC + RMFCC	**7.30**
MFCC + DCTILPR	7.50

Bold values indicate the best performance

In the current scenario probabilistic approaches are widely used for speaker recognition tasks (Dehak et al. 2011; Reynolds et al. 2000). For example, the GMM is used for speaker identification task, whereas GMM–UBM and i-vector based systems for speaker verification task. The i-vector based SV system is presently considered as the state-of-the-art modelling technique. The proposed combination scheme can be applied with GMM, GMM–UBM and i-vector based systems. The detail procedures of the proposed combination scheme for GMM and i-vector based SR systems are described below. For clear explanation, we consider two feature sets, X = MFCC (vocal-tract) and Y = RMFCC (excitation source). However, the proposed combination scheme can be applied for combined use of any number of features sets.

4.1 Proposed combination scheme for Gaussian modelling based systems

The GMM technique represents the features set in the form of mean, variance and weights. Let us consider by using Gaussian modelling process the train features set X_{tr} from source X generates set of mean vectors μ_x, set of variance vectors Σ_x and set of weights ω_x. The corresponding parameters by train features set Y_{tr} from source Y are μ_y, Σ_y and ω_y, respectively. The mean vectors set μ_z, variance vectors set Σ_z and weights ω_z of the combined GMM system are made by using Eq. 4. The ω_z is scaled by a factor of 1/2 to satisfy the Gaussian modelling constraint $\sum \omega = 1$ (Reynolds et al. 2000). The similar approach can be extended to GMM–UBM process. In that case first the adapted UBM models are built by using individual features set independently, and later padded as explained earlier for building composite adaptive models.

$$\mu_z = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \Sigma_z = \begin{bmatrix} \Sigma_x \\ \Sigma_y \end{bmatrix}, \omega_z = \begin{bmatrix} \omega_x \\ \omega_y \end{bmatrix} * \frac{1}{2} \tag{4}$$

At the time of comparison the resultant test features set Z_{tx} is made by padding test features sets X_{tx} and Y_{tx} from sources X and Y, respectively. Mathematically,

$$Z_{tx} = \begin{bmatrix} X_{tx} \\ Y_{tx} \end{bmatrix} \tag{5}$$

The comparison process is very similar to standard Gaussian modelling based system. The block diagram representation of the proposed composite Gaussian based speaker recognition system is shown in Fig. 1. First, individual GMMs are made by using different features separately. These independent models are then padded (by using Eq. 4) resulting the composite model. In the similar fashion the composite test features set is formed for comparison purpose. It can be observed from the Fig. 1 that feature and model levels computations are independent and thereby retain their original discriminating ability. However, they contributed collectively in giving the decisions. This reduces the confusion from individual decisions resulting an improvement in recognition accuracy.

The step-by-step procedure of the proposed combination scheme for GMM based speaker recognition system:

- The MFCC (X_{tr}) and RMFCC (Y_{tr}) training features are first computed for individual enrolled speakers by using their respective available speech samples.

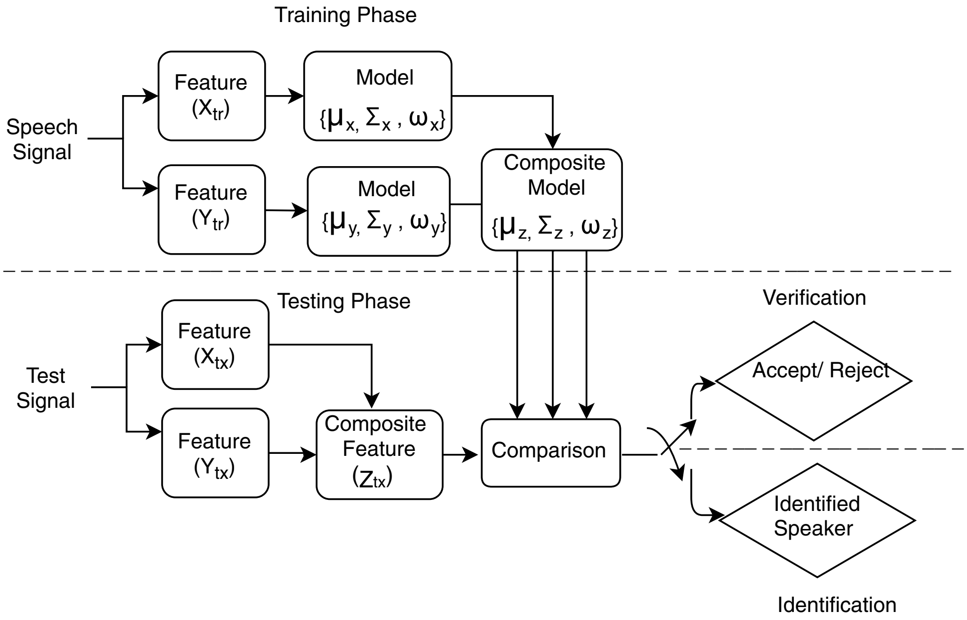

Fig. 1 Block diagram of the proposed GMM–UBM based SV system

- The individual speaker-specific GMM model parameters ($\{\mu_x, \Sigma_x, \omega_x\}$ and $\{\mu_y, \Sigma_y, \omega_y\}$) are computed by using respective X_{tr} and Y_{tr} features for all enrolled speakers.
- The individual speaker-specific composite model parameters ($\{\mu_z, \Sigma_z, \omega_z\}$) are made by using Eq. 4.
- In case of GMM–UBM based system, first the adaptive UBM models are built individually by using respective train features set. The composite adaptive model parameters are made by padding the individual adaptive model parameters in the similar fashion as mentioned in Step-II and Step-III.
- During testing, first the respective test features (X_{tx} and Y_{tx}) are computed from the test speech samples and padded by using the Eq. 5 for building composite test features set (Z_{tx}).
- Finally, the composite features set are used directly for comparison with composite models for different speaker recognition tasks. The comparison process is very similar to standard GMM process (Reynolds 1995).

4.2 Proposed combination scheme for I-vector based system

The classical GMM–UBM based SV systems effectively capture the speaker variabilities but suffers from channel/session variabilities (Kenny et al. 2007a). The later discoveries like, the use of GMM supervector with support vector machine (SVM) and more recently with joint factor analysis (JFA) are used to compensate the channel variability (Kenny et al. 2007b). The further experiments on NIST evaluation dataset proved that channel factors estimated using JFA, which are supposed to model only channel effects, also contain information about speakers. Based on that observation the i-vector based system was developed for speaker verification tasks (Dehak et al. 2011). In this approach a single space referred as the total variability space is defined. The total variability space contains speaker and channel variabilities simultaneously and transforms the GMM mean supervector of an utterance to a low dimensional vector, called as identity vectors or i-vectors for short. The details of the conventional i-vector based SV system can be found in (Dehak et al. 2011).

The proposed combination scheme can also be applied to state-of-the-art i-vector based SV system. The block diagram of the proposed i-vector based combined SV system is shown in Fig. 2. The feature extraction and UBM building processes are followed by usual methods. The respective 0th and centralized 1st order Baum–Welch statistics (sufficient statistics in Fig. 2) are computed separately by using X_{tr} and Y_{tr} features. The composite Baum–Welch statistics are made by using the Eq. 6, where $\{N_x, F_x\}$ and $\{N_y, F_y\}$ represent the Baum–Welch statistics of X_{tr} and Y_{tr} features sets. $\{N_z, F_z\}$ represents the composite Baum–Welch statistics (combined sufficient statistics in Fig. 2).

$$N_z = \left[N_x \; N_y\right], F_z = \left[F_x \; F_y\right], \tag{6}$$

The training of the composite T-matrix is performed by using the composite sufficient statistics and following the similar approach as discussed in (Dehak et al. 2011). Then, the i-vector of the input speech signal is estimated by using the composite T-matrix. Since, the composite T-matrix is performed by joint use of X_{tr} and Y_{tr} features, the corresponding i-vector reflects fused information in together. The rest of the process is similar to conventional i-vectors based system described in (Dehak et al. 2011).

The step-by-step procedure of the proposed combination scheme for i-vector based speaker recognition system:

- The MFCC (X_{tr}) and RMFCC (Y_{tr}) features are first computed from the speech samples of individual enrolled speakers.
- The respective features specific UBMs are built by using the speech samples of the background speakers.
- The individual Baum–Welch statistics ($\{N_x, F_x\}$,$\{N_y, F_y\}$) are computed separately by using X_{tr} and Y_{tr} features and the respective UBMs.
- The composite Baum–Welch statistics ($\{N_z, F_z\}$) is formed by padding the individual parameters as shown in Eq. 6.

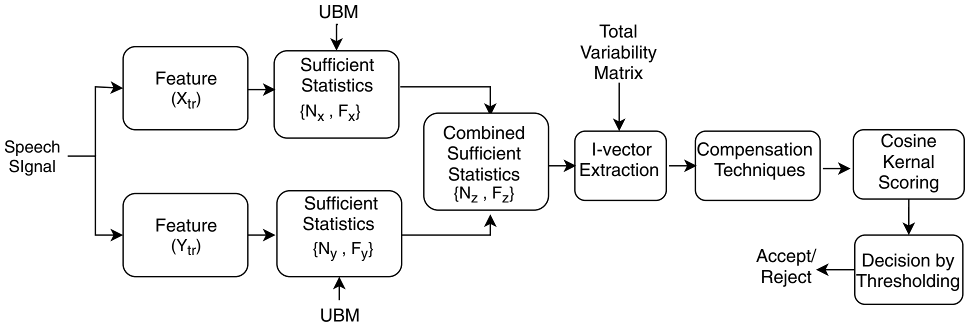

Fig. 2 Block diagram of the proposed i-vector based SV system

- The individual T-matrices of MFCC (X_{tr}) and RMFCC (Y_{tr}) features are computed by using the process explained in (Dehak et al. 2011).
- The composite T-matrix is made by padding the individual T-matrices in similar fashion.
- The i-vectors of the enrolled models are estimated by using the composite Baum–Welch statistics and composite T-matrix.
- The compensation techniques such as LDA and WCCN are applied to reduced the session and channel effects.
- During the testing phase, the similar procedure (Step-VII) is applied to get the i-vectors of the test speech samples.
- Finally the cosine kernel scoring are calculated and decisions are made by thresholding as similar to the conventional i-vector based system.

5 Database description

The potential of the proposed scheme is demonstrated by speaker recognition experiments in clean, noisy and other distorted conditions. The experiments are made with three databases: TIMIT (Garofolo et al. 1993), IITG-MV (Haris et al. 2012) and NIST-2003 (The 2003 Nist speaker recognition evaluation plan 2003). The TIMIT database is used for analysis on clean case and IITG-MV database for noisy case. The TIMIT and IITG-MV databases are segregated into male and female speakers sets for gender independent analysis. The experiments are also made following the standard speaker recognition evaluation (SRE) plan with NIST-2003 database. The brief statistics about all these databases is given below.

5.1 TIMIT database

The TIMIT database consists of 630 (438 males and 192 females) American English speakers phonetically balanced speech data collected in clean environment and thus considered for clean case analysis. The speakers' speech samples consist of ten utterances each roughly of 3 s duration recorded at 16000 samples/s. In case of speaker identification studies the first eight sentences are used for building models and the reaming two for testing, resulting 1260 (876 males and 384 females) identification trials.

In case of speaker verification experiments the first 150 (75 males and 75 females) speakers speech data are used for building UBM models and the remaining 480 (363 males and 117 females) speakers speech data are used for evaluation. In this case, the first eight utterances of the enrolled speakers are used for building reference models and the remaining two utterances are used for testing. The detail statistics is given in Table 2. The trials with respective models are considered as genuine and with others as impostors, resulting 960 (726 males and 234 females) genuine trials and 289956 (262812 males and 27144 females) impostor trials.

Table 2 The statistics of TIMIT database used for conducting speaker identification and verification experiments

Tasks	Speakers set	No. of speakers	Trials	
Identification	Male	438	1260	
	Female	192		
	Total	630		
Verification	Male	363	ZT	IT
			726	262812
	Female	117	234	27144
	Whole-set	480	960	289956

ZT represents Genuine trials and IT represents impostor trials

5.2 IITG-MV database

The IITG-MV speech database contains two sets: Phase-I (81 male and 19 female) and Phase-II (70 male and 30 female). The speech samples are recorded in multi-environments like in laboratories and hostel rooms. While taking recordings, the placement of the sensors was such that speech data can be collected in degraded conditions like background noise and reverberation. So we consider IITG-MV database for noise case analysis. The speaker characteristics including behavioral traits are better reflected in conversational speech. In addition, with respect to read speech the availability of conversation data is comparatively more. It helps in generating large number of test samples. Thus, we choose the conversational speech data for our experimental studies.

Altogether the Phase-I and Phase-II datasets contain 148 (112 male and 36 female) distinct speakers. The complete 148 speakers data are used for speaker identification experiments. The detail statistics is given in Table 3. The first two minutes of speech data of each speaker recordings are used for building respective speaker models (Reynolds et al. 2000). The remaining speech data are divided into several segments of 30 s duration and used as test samples, resulting 2873 identification trials.

The speaker verification process is one-to-one comparison. Therefore we follow the following approach for database development. The female speakers strength of IITG-MV dataset is very less around 36 (19 and 17 speakers in Phase-I and Phase-II respectively). In order to avoid any gender biasing we consider only 45 male and 35 female speakers speech data for speaker verification experiments. The first five speakers speech data ($\simeq$ 1 h/gender) are used for building respective gender independent UBM models. The remaining 70 speakers (40 male and 30 female) speech data are used for speaker verification performance evaluation. In

Table 3 The statistics of IITG-MV database used for conducting speaker identification and verification experiments

Tasks	Speakers set	No. of speakers			Trials	
		Phase-I	Phase-II	Total		
Identification	Male	81	31	112	2873	
	Female	19	17	36		
	Total	100	48	148		
Verification	Male	21	19	40	ZT	IT
					706	27534
	Female	17	13	30	276	8004
	Whole-set	38	22	70	982	35538

ZT represents genuine trials and IT represents impostor trials

this case also the first two minutes of speech data are used for enrollment. The remaining data are converted into several segments of 30 s duration and used for test trials. Each test segment of each speaker is used as a genuine trial for the same target model and an impostor trial against other speakers model of the same gender. The detail statistics about number of verification trials is given in Table 3. Altogether we have 36520 test cases that include 982 genuine and 35538 impostors trials. The 982 genuine tests includes 706 male and 276 female trials.

5.3 NIST-2003 database

The NIST-2003 SRE database contains 356 train speakers speech samples for building models and 2559 test samples for verification trials. The experiments on NIST-2003 database are conducted using the standard SRE plan (The 2003 Nist speaker recognition evaluation plan 2003). The gender independent background models are built by using approximately 40 h speech data from 100 males and 100 females members of Switchboard Corpus II cellular database Linguistic data consortium (2004). These speakers are not included in the NIST-03 evaluation set. As per the SRE plan the NIST-2003 database includes 2212 genuine trials and 25915 impostor trials (The 2003 Nist speaker recognition evaluation plan 2003).

6 Experimental results and discussion

The experimental studies are made to demonstrate the potential of the proposed scheme for combined use of vocal-tract and excitation source information in automatic speaker recognition tasks with GMM, GMM–UBM and I-vector based systems. The vocal-tract information is represented by standard MFCC features. These features are computed from 20 ms overlapping voiced speech frames at the rate of 100 frames/s. The speech signals are processed at 8000 samples/s and voice/unvoiced detections are made by using energy based thresholding. The first 13 coefficients (excluding c_0) concatenated with corresponding 13 delta and delta-delta coefficients are used as the representation of the vocal-tract information. The excitation source information is represented by RMFCC features. These features are computed exactly in similar manner as in case of MFCC except the use of LP residual signal (Eq. 2). In case of experiments with NIST-2003 database the zero mean unit variance normalization is employed to avoid the effect of channel variability (Reynolds et al. 2000). The speaker identification experiments are conducted with GMM based system and results are expressed in terms of identification accuracy (%). The speaker verification experiments are conducted with GMM–UBM based system and the results are expressed in EER (%) and detection-error-trade-off (DET) curves (Martin et al. 1997). The widely used score level scheme is considered as the reference and comparative studies are made to demonstrate the usefulness of the proposed combination scheme. Finally, the potential of the proposed scheme is also demonstrated with state-of-the-art i-vector based SV system.

6.1 Speaker identification experiments

In order to achieve the best possible results, speaker identification experiments are conducted with different Gaussian mixtures size. The optimum size for TIMIT database is found to be 32 and 256 for IITG-MV database. The smaller Gaussian mixtures size in case of TIMIT database is due to the use of comparatively small amount of enrollment data. The experimental results for TIMIT and IITG-MV databases are reported in Table 4. As obvious, in both cases the baseline MFCC feature shows comparatively higher identification accuracy. The performance is relatively worsen in case of IITG-MV database mainly due to the nature of noisy data.

The advantage of using combined evidences can be observed from the fusion of MFCC and RMFCC features. The baseline performances with TIMIT database is improved from 95.39 to 96.19% by score level combination scheme achieving a gain of 0.84%. In implementing the proposed combination scheme the baseline performance is

Table 4 Performance of GMM based SI system with MFCC and RMFCC features, their joint use by score level and proposed combination schemes on TIMIT and IITG-MV databases

System	Identification accuracy (%)	
	TIMIT	IITG-MV
MFCC (baseline)	95.39	90.11
RMFCC	82.14	80.47
Combination		
Score level	96.19	91.19
Proposed scheme	**97.14**	**92.23**

Bold values indicate the best performance

improved to relatively higher value of 97.14% amounting to 1.80% gain. The similar trend is also observed with IITG-MV database, where the baseline performance is improved from 90.11 to 91.19 and 92.23% by score level and proposed combination schemes, respectively. The gain by score level fusion is 1.80%, as compared to 2.35% by the proposed scheme. The results show that the proposed combination scheme can be effectively used with GMM based speaker identification system for application both in clean and noisy conditions.

6.2 Speaker verification experiments

In following studies, we have experimentally shown that the proposed combination scheme effectively helps in reducing the false acceptance and false rejection errors (equivalently the EER) made by classic GMM–UBM and i-vector based SV systems. As mentioned earlier, for gender independent analysis the speaker verification experiments are conducted independently for male and female speakers sets of TIMIT and IITG-MV databases. The gender independent true and false scores are first normalized (with reference to their respective maximum score), and then padded. We refer it as resultant scores set. The overall EER is computed by using resultant scores set.

6.2.1 Speaker verification results with GMM–UBM based system

The experimental results of SV system with TIMIT and IITG-MV databases are reported in Table 5 and the corresponding DET curves are shown in Fig. 3. In this case also we observed that 32 GMM size provides the optimized EER with TIMIT database and 256 GMM for IITG-MV database. As expected, similar to identification case the MFCC feature based baseline system performs comparatively well with both TIMIT and IITG-MV databases. The respective baseline performances are improved while combining with RMFCC features based system, either by score level or by proposed combination scheme. In later case the relative improvements are comparatively more. For example, in case of overall performance, the relative improvements (Ref. last row of the Table 5) by score level scheme are 17.54% and 18.77% for TIMIT and IITG-MV databases. Similarly, the proposed scheme provide 26.10% and 21.85% relative improvement, approximately 8.5% and 3% higher improvements than the score level combination scheme. The similar trends are also observed with individual male and female trials.

Notice that the baseline performances are not consistent but largely varying with male and female speakers trials. The performance is comparatively poor for female trials. In case of TIMIT database (Table 2) the ratio of impostor-to-genuine female trials is 116, as compared to 362 for male trials. Similarly, in case of IITG-MV database (Table 3) the impostor-to-genuine female trials is 29, as compared to 39 for male trials. The EER is a statistical measure. The difference in proportion of impostor-to-genuine trials may be the

Table 5 Performance of GMM–UBM based SV system with MFCC and RMFCC features, and their joint use by score level and proposed combination schemes on TIMIT and IITG-MV databases

System	EER (%)					
	TIMIT			IITG-MV		
	Male	Female	Overall	Male	Female	Overall
MFCC (baseline)	2.61	4.70	4.79	2.97	3.62	3.25
RMFCC	4.40	10.63	5.83	5.43	4.71	5.09
Score level combination	2.20	4.27	3.95	2.87	2.99	2.64
Proposed scheme	**2.06**	**2.56**	**3.54**	**2.49**	**2.53**	**2.54**
Rltv (%)						
Score level	15.71	9.15	17.54	3.37	17.40	18.77
Proposed scheme	**21.07**	**45.53**	**26.10**	**16.16**	**31.11**	**21.85**

Bold values indicate the best performance

Rltv represents the relative improvement by a combination scheme w. r. t. the baseline system

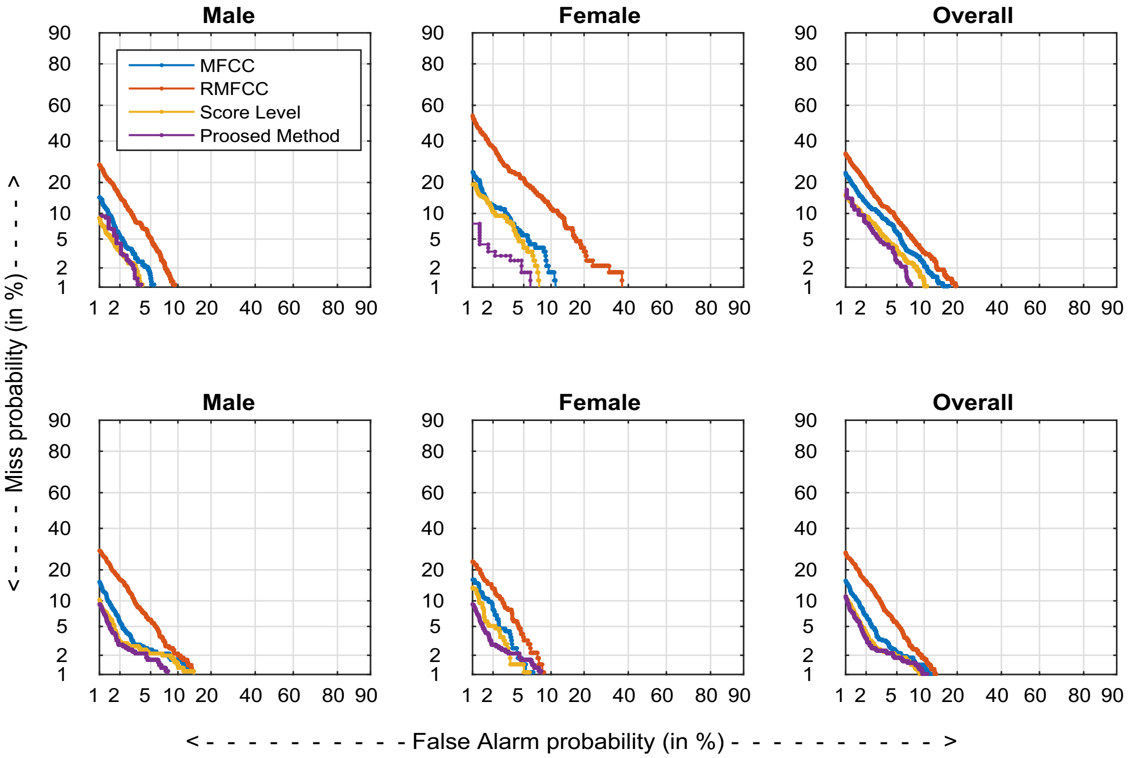

Fig. 3 DET plot showing the performance of the gender independent analysis of GMM–UBM based SV system with MFCC and RMFCC features, and their joint use by score level and proposed combination schemes on TIMIT and IITG-MV database. First row shows the results on TIMIT database, whereas second row shows the results on IITG-MV database

reason for that inconsistency in performance. This may also be the reason that affects the overall performance. This is where the use of combined evidences plays a helping role towards maintaining the consistency with performance. It can be observed from the last two rows of the Table 5 that the combined use of evidences reduces the differences of performance with female and male trials. In that context the proposed scheme maintains comparatively stronger consistency. In case of score level scheme the performance differences with female and male trials are 2.07 with TIMIT database and 0.12 with IITG-MV database. The corresponding parameters with respect to proposed combination scheme are 0.5 and 0.04, respectively. The relative improvements are comparatively higher for female trials. This may be due to relatively poor baseline performance with female trials.

The potential of the proposed scheme is also demonstrated with standard NIST-2003 database for further clarity. The experimental results of MFCC and RMFCC features are refereed from Table 1 and reported in Table 6. The corresponding DET curves are shown in Fig. 4. The results show the similar trend. The MFCC features based baseline performance of 7.54% is improved to 7.30% and 3.18% by score level and proposed combination schemes, respectively. In former case the relative improvement is 3.18% as compared to 8.35% by the proposed scheme.

6.2.2 Speaker verification results with I-vector based system

The proposed combination scheme can also be effectively used with state-of-the-art i-vector based system. To demonstrate that we employed i-vector based approach and repeat speaker verification experiments with NIST-2003 database. We follow exactly similar procedure for extraction of MFCC and RMFCC features. Switchboard Corpus II cellular database is used as development data to build the universal background model (UBM) and the T-matrix. A gender-independent UBM of 1024 mixtures is trained using approximately 10 h data taken from the development set. To eliminate the gender biasing, we have made two separate male and female UBM model of 512 size, and finally pooled together to built the gender independent UBM. The Baum–Welch statistics of training and testing data are estimated separately and finally pooled together as described in Sect. 4.2. Similarly, the Baum–Welch statistics of the development set are computed separately and pooled together to get the composite statistics. The low dimensional i-vector representation is derived from the 400-dimensional T-matrix. The i-vector based speaker modeling has both speaker and channel information. The 150-dimensional

Table 6 Performance of GMM–UBM based SV system with MFCC and RMFCC features, and their joint use by score level and proposed combination schemes on NIST-2003 database

System	EER (%)
MFCC	7.54
RMFCC	18.89
Score level combination	7.30
Proposed scheme	**6.91**
Rltv (%)	
Score level	3.18
Proposed scheme	**8.35**

Bold values indicate the best performance

Rltv represents the relative improvement by a combination scheme w. r. t. the baseline system

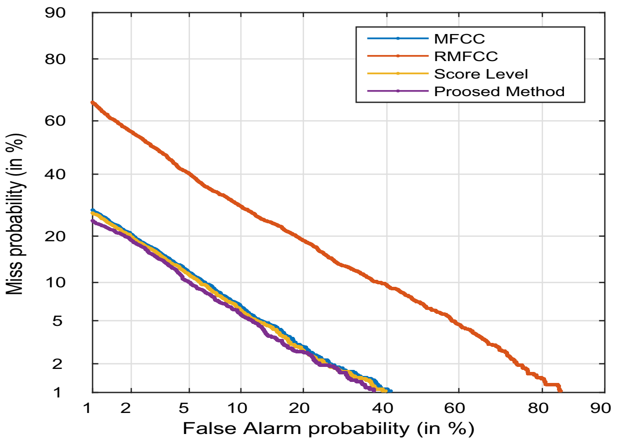

Fig. 4 DET plot showing the performance of GMM–UBM based SV system with MFCC and RMFCC features, and their joint use by score level and proposed combination schemes on NIST-2003 database

Table 7 Performance of i-vector based SV system with MFCC and RMFCC features, their joint use by score level and proposed combination schemes on NIST-2003 database

System	EER (%)
MFCC	2.36
RMFCC	10.16
Score level combination	2.16
Proposed scheme	**2.05**
Rltv (%)	
Score level	8.4
Proposed scheme	**13.13**

Bold values indicate the best performance

Rltv represents the relative improvement by a combination scheme w. r. t. the baseline system

LDA and 400-dimensional WCCN are used as channel/session compensation techniques. The EER is measured by using the cosine kernel scores.

The experimental results are reported in Table 7 and the corresponding DET curves are shown in Fig. 5. The MFCC feature based baseline system provides the best performance of 2.36%, as compared to 10.16% by RMFCC feature. The score level combination scheme improves the baseline performance from 2.36 to 2.16%, reflecting a relative improvement of 8.4%. The proposed combination scheme improves the baseline performance to 2.05%, showing a relative improvement of 13.14%. This demonstrate the usefulness of the proposed scheme with i-vector based system as well.

To summarize, the experimental results demonstrate that the proposed combination scheme well exploit the benefit of using joint evidences from MFCC and RMFCC features for speaker recognition tasks. The proposed scheme can be effectively used for GMM, GMM–UBM and i-vector based speaker recognition systems. It performs well with clean, noisy and in other environmental variations. As compared to the fusion of evidences by score level combination, the proposed scheme provides comparatively higher improvements in terms of speaker recognition performance without the need of the ground truth information. The experimental results show that the proposed scheme performs comparatively well with speaker verification tasks. The reason may be due to the nature of the recognition process.

7 Statistical significance of the proposed scheme

The statistical significance of the proposed combination scheme is investigated by measuring the confidence interval (CI) and p-value measurements from hypothesis test with genuine trials. In speaker verification task a genuine claim is accepted when the claimant's average log-likelihood score is above the requisite threshold (*Th*) value. As such, the *null hypothesis* H_o and *alternative hypothesis* H_1 (Neyman 1937) can be defined as:

$$
\begin{aligned}
H_o &: \mu < Th \\
H_a &: \mu >= Th
\end{aligned}
\tag{7}
$$

where μ is the average log-likelihood score of the test trails. The CI represents the number of times the decisions are taken confidently. The p-value represents the number of times the decisions are made randomly. Ideally, a good SV system should reflect higher CI and lower p-value. We consider 500 randomly chosen genuine trials from NIST-2003 database and evaluate CI and p-value of score-level fusion and proposed combination schemes. The values are

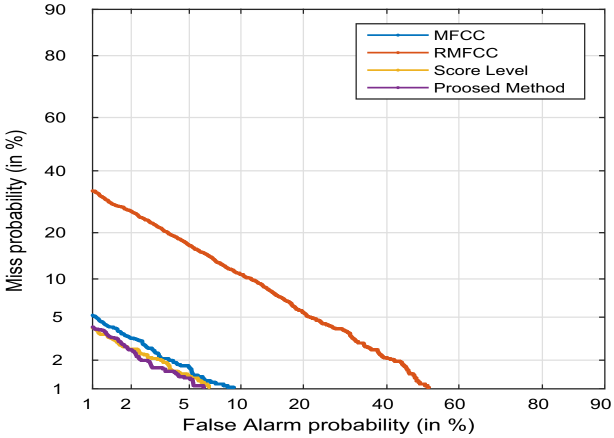

Fig. 5 DET plot showing the performance of I-vector based SV system with MFCC and RMFCC features, and their joint use by score level and proposed combination schemes on NIST-2003 database

Table 8 Confidence interval (CI) and p-values of score level and proposed combination schemes

Type of scheme	Confidence interval (%)	P value
Score level combination scheme	96.6	0.034
Proposed combination scheme	**97.4**	**0.025**

Bold values indicate the best performance

The CI and p-values are computed from 500 randomly chosen genuine trials of NIST-2003 database

reported in Table 8. The proposed scheme shows comparatively higher confidence interval of 0.8% than the score level combination scheme. The p-values show that the number of times in taking random decisions by the proposed scheme is also comparatively less (around 0.9%). These observations reflect the statistical significance of the proposed combination scheme.

8 Conclusion

In this work, we have proposed a combination scheme that provides improved speaker recognition performance. The idea is that, while taking decisions the collective contribution of multiple evidences may be more effective. This is achieved by padding feature-specific models and test feature vectors in similar fashion. As compared to the widely used score level fusion, the proposed scheme provides improved performance and robustness with GMM based speaker identification system, GMM–UBM and i-vector based speaker verification systems for both in clean and noisy environments. The proposed scheme is statistically more significant than score level fusion, and does not require any ground truth information. Further, the proposed combination scheme is independent of feature extraction processes and modeling techniques. As such, it can be used for other similar kind speech processing applications like in speech recognition, language identification and dialect identification tasks.

Acknowledgements This research work is funded by Department of Electronics and Information Technology (DeitY), Govt. of India through the project "Development of Excitation Source Features Based Spoof Resistant and Robust Audio-Visual Person Identification System". The research work is carried out in Speech Processing and Pattern Recognition (SPARC) laboratory at National Institute of Technology Nagaland, Dimapur, India.

References

Altnccay, H., & Demirekler, M. (2000). An information theoretic framework for weight estimation in the combination of probabilistic classifiers for speaker identification. *Speech Communication*, *30*(4), 255–272.

Altnccay, H., & Demirekler, M. (2003). Speaker identification by combining multiple classifiers using dempster-shafer theory of evidence. *Speech Communication*, *41*(4), 531–547.

Atal, B. S. (1976). Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, *64*(4), 460–475.

Beigi, H. (2011). *Fundamentals of speaker recognition*. Berlin: Springer.

Campbell, J. P. (1997). Speaker recognition: A tutorial. *Proceedings of IEEE*, *85*(9), 1437–1462.

Campbell, W. M., Sturim, D. E., & Reynolds, D. A. (2006). Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, *13*(5), 308–311.

Das, R. K., & Prasanna, S. R. M. (2016). Exploring different attributes of source information for speaker verification with limited test data. *The Journal of the Acoustical Society of America*, *140*(1), 184–190.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798.

Djemili, R., Bedda, M., & Bourouba, H. (2007). A hybrid gmm/svm system for text independent speaker identification. *International Journal of Computer and Information Science & Engineering*, *1*(1).

Duda, R. O., Hart, P. E., & Stork, D. G. (1973). *Pattern classification and scene analysis*. New York: Wiley.

Farrell, K., Kosonocky, S., & Mammone, R. (1994). Neural tree network/vector quantization probability estimators for speaker recognition. In *Proceedings of the 1994 IEEE workshop on neural networks for signal processing*, pp. 279–288.

Feustel, T. C., Logan, R. J., & Velius, G. A. (1988). Human and machine performance on speaker identity verification. *The Journal of the Acoustical Society of America*, *83*(S1), S55–S55.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustic, Speech, and Signal Processing*, *29*, 254–272.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., et al. (1993). *Timit acoustic-phonetic continuous speech corpus*. Philadelphia: Linguistic data consortium.

Haris, B. C., Pradhan, G., Misra, A., Prasanna, S. R. M., Das, R. K., & Sinha, R. (2012). Multi-variability speaker recognition database in Indian scenario. *International Journal of Speech Technology*, *15*(4), 441–453.

Hermansky, H., & Morgan, N. (1994). Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, *2*(4), 578–589.

Hosseinzadeh, D., & Krishnan, S. (2007). Combining vocal source and mfcc features for enhanced speaker recognition performance using gmms. In *IEEE* 9th *Workshop on Multimedia Signal Processing*, pp. 365–368.

Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007). Speaker and session variability in gmm-based speaker verification. *IEEE Transactions on Audio Speech and Language Processing*, *15*(4), 1448.

Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007). Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(4), 1435–1447.

Linguistic data consortium, switchboard cellular part 2 audio. (2004). Retrieved from, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jspcatalogId=LDC2004S07.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The det curve in assessment of detection task performance. In *Technical Report*, National Institute of Standards and Technology Gaithersburg MD.

Mashao, D. J., & Skosan, M. (2006). Combining classifier decisions for robust speaker identification. *Pattern Recognition*, *39*(1), 147–155.

Murty, K. S. R., & Yegnanarayana, B. (2006). Combining evidence from residual phase and mfcc features for speaker recognition. *IEEE Signal Processing Letters*, *13*(1), 52–55.

Nakagawa, S., Wang, L., & Ohtsuka, S. (2012). Speaker identification and verification by combining mfcc and phase information. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(4), 1085–1095.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London A*, *236*(767), 333–380.

Pati, D., & Prasanna, S. R. M. (2010). Speaker recognition from excitation source perspective. *IETE Technical Review*, *27*(2), 138–157.

Pati, D., & Prasanna, S. R. M. (2011). Subsegmental, segmental and suprasegmental processing of linear prediction residual for speaker information. *International Journal of Speech Technology*, *14*(1), 49–64.

Pati, D., & Prasanna, S. R. M. (2013). A comparative study of explicit and implicit modelling of subsegmental speaker-specific excitation source information. *Sadhana*, *38*(4), 591–620.

Poh, N., & Kittler, J. (2008). Incorporating model-specific score distribution in speaker verification systems. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(3), 594–606.

Prasanna, S. R. M., Gupta, C. S., & Yegnanarayana, B. (2006). Extraction of speaker-specific excitation information from linear prediction residual of speech. *Speech Communication*, *48*(10), 1243–1261.

Ramachandran, R. P., Farrell, K. R., Ramachandran, R., & Mammone, R. J. (2002). Speaker recognition-general classifier approaches and data fusion methods. *Pattern Recognition*, *35*(12), 2801–2821.

Ramakrishnan, A., Abhiram, B., & Prasanna, S. R. M. (2015). Voice source characterization using pitch synchronous discrete cosine transform for speaker identification. *The Journal of the Acoustical Society of America*, *137*(6), EL469–EL475.

Reynolds, D. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, *17*, 91–108.

Reynolds, D. A. (1995). Large population speaker identification using clean and telephone speech. *IEEE Signal Processing Letters*, *2*(3), 46–48.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, *10*(1–3), 19–41.

The 2003 Nist speaker recognition evaluation plan (2003). In *Proceedings of NIST Speaker Recognition Workshop*, College Park, MD.

Venturini, A., Zao, L., & Coelho, R. (2014). On speech features fusion, α-integration Gaussian modeling and multi-style training for noise robust speaker classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(12), 1951–1964.

Wong, L. P., & Russell, M. (2001). Text-dependent speaker verification under noisy conditions using parallel model combination. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP01)*, 1, 457-460.

Yegnanarayana, B., Prasanna, S. R. M., Zachariah, J. M., & Gupta, C. S. (2005). Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system. *IEEE Transactions on Speech and Audio Processing*, *13*(4), 575–582.

Yegnanarayana, B., Reddy, K. S., & Kishore, S. P. (2001). Source and system features for speaker recognition using ANNN models. In *Proceedings of Acoustics, Speech, and Signal Processing (ICASSP-01)* (Vol. 1, pp. 409–412).