IEICE TRANS. INF. & SYST., VOL.E102–D, NO.2 FEBRUARY 2019

PAPER

Speaker-Phonetic I-Vector Modeling for Text-Dependent Speaker Verification with Random Digit Strings

Shengyu YAO[†,††a)], **Ruohua ZHOU**[†,††b)], *Nonmembers*, *and* **Pengyuan ZHANG**[†,††], *Member*

SUMMARY This paper proposes a speaker-phonetic i-vector modeling method for text-dependent speaker verification with random digit strings, in which enrollment and test utterances are not of the same phrase. The core of the proposed method is making use of digit alignment information in i-vector framework. By utilizing force alignment information, verification scores of the testing trials can be computed in the fixed-phrase situation, in which the compared speech segments between the enrollment and test utterances are of the same phonetic content. Specifically, utterances are segmented into digits, then a unique phonetically-constrained i-vector extractor is applied to obtain speaker and channel variability representation for every digit segment. Probabilistic linear discriminant analysis (PLDA) and s-norm are subsequently used for channel compensation and score normalization respectively. The final score is obtained by combing the digit scores, which are computed by scoring individual digit segments of the test utterance against the corresponding ones of the enrollment. Experimental results on the Part 3 of Robust Speaker Recognition (RSR2015) database demonstrate that the proposed approach significantly outperforms GMM-UBM by 52.3% and 53.5% relative in equal error rate (EER) for male and female respectively.

key words: *speaker verification, text-dependent, speaker-phonetic, random digit strings, i-vector, phonetically-constrained*

1. Introduction

Speaker verification refers to verifying the claimed individual based on his/her voice. Rapid development has been experienced in this field during the last decade. And currently, the newly introduced channel compensation techniques such as joint factor analysis (JFA), i-vector [1], [2] and PLDA have become popular and made a great contribution to large improvement of the verification performance. Guided by the NIST speaker recognition evaluations [3], these methods have been applied successfully to the text-independent speaker verification task.

In many applications, speaker verification systems are required to be capable of obtaining satisfactory performance for short utterances. However, the i-vector based system suffers from severe performance degradation when the enrollment and test utterances are short, even though it propagates the uncertainty of the i-vector estimation to the PLDA [4]–[6]. This is mainly because of the mismatch of the phrase contents between enrollment and test utterance [7]. Text-dependent speaker verification methods are thus used to avoid the mismatch problem and achieve much better performance than text-independent speaker verification.

Reports in [8] have defined four challenges for text-dependent task. They are the common pass-phrase with abundant background data, the common pass-phrase with scarce background data, the randomized pass-phrase with constrained vocabulary and the unique pass-phrase with unconstrained vocabulary. Recently, increasing demand for voice-based access control applications has attracted our attention on the third situation which is so called random digit strings task. Part 3 of RSR2015 database [9] has been designed for speaker verification with random digit strings, where the user is prompted by the system to utter random sequences of digits.

Attempt has been made to adopt the text-independent techniques for the speaker verification task with random digit strings. However, unfortunately, the state-of-the-art i-vector/PLDA method is found ineffective in this situation, and better performance can be obtained by using simpler GMM-UBM techniques [9], [10]. This incapacity of the conventional i-vector method can be explained from three aspects, the co-articulation effects between digits [8], the lack of in-domain training data and the inadequacy of out-of-domain datasets [13]. These three difficulties make it inappropriate to directly apply text-independent i-vector approach to random digit strings task.

Some studies on text-dependent speaker verification task are deserve to be reported first. In [10], a phrase-dependent PLDA model is proposed to make use of the phonetic content. It shows better performance by training i-vector extractor with text-dependent database. In [11], phone-centric local variability vector is proposed, each loading matrix is corresponding to a monophone, and the i-vector is the concatenation of monophone local vectors. The work in [12] uses a phrase-specific HMM for every utterance and extracts Baum-Welch statistics in a phonetic-dependent way. Frames aligned to the same monophone are used to collect sufficient statistics with the corresponding state. For i-vector estimating, it concatenates sufficient statistics of all monophones, then the i-vector extractor is trained and used in the usual way. These works show that phonetic content plays a major role in text-dependent speaker verification performance.

To utilize the phonetic content, many subsystem based methods are widely used in recent years. The authors in

Manuscript received September 11, 2018.
Manuscript publicized November 19, 2018.

[†]The authors are with Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, China.
[††]The authors are with University of Chinese Academy of Sciences, China.
a) E-mail: yaoshengyu@hccl.ioa.ac.cn
b) E-mail: zhouruohua@hccl.ioa.ac.cn (Corresponding author)

DOI: 10.1587/transinf.2018EDP7310

[9] propose HiLAM approach, which makes each state of a HMM to be a GMM that models a speaker's one of the 10 digits and deploys the standard Viterbi algorithm for evaluating. In [15], a state-based JFA approach is proposed where one UBM and one corresponding JFA are trained for each digit. The work in [16] builds a DNN/i-vector based subsystem for each single digit individually, and LDA is used to reduce i-vector dimension. In [17], the authors report similar approach, the difference is that they use 10 simple GMM/i-vector subsystems with digit-dependent PLDA modeling. Earlier studies [18], [19] also have tried the same idea and train a i-vector based subsystem for each word or state. The commonality of the above approaches is that they segment the utterances into phonetic units (digits, words or states) and makes each pair of enrollment and test segments to be of the same phonetic content, moreover they utilize the alignment information of a utterance by building multiple subsystems, each subsystem is only used for a corresponding unit and mutually independent with others. However, application of these subsystem based methods to random digit strings task will suffers from two great difficulties. One is related to the sparsity of training data, since training data has to be divided into different digit segments, only the segments with same digit will be used to build an individual system and the subsystem seems to be unreliable trained with limited data. Another is the variability of individual subsystems, as it would challenge the score fusion.

An alternative approach is reported in [13], it has similarity with the subsystem based methods, but uses a single JFA-JDB system for modeling segments corresponding to all 10 digits. This work explores three different ways to collect Baum-Welch statistics for obtaining local and global vectors. The authors investigate different concatenations of using local and global vectors, then attain encouraging performance by fusing their scores. Despite its good results, using the "multi-tier" JFA and the fusion of multiple systems is complex. We learn from this work and think that using a common system instead of multiple subsystems might be a good choice in practical applications.

In this paper, we propose a scheme of using speaker-phonetic i-vector modeling and show great promotion for text-dependent speaker verification with random digit strings. We firstly segment utterances into individual digits by using HMM based speech recognizer and make each pair of enrollment and test segments to be of the same digit. Then, we proposed the speaker-phonetic modeling scheme that uses a common i-vector based system for low-dimensional representation of segments belonging to all 10 digits. Which differs in using one common system, versus 10 individual systems of the subsystem based methods. For the purpose of utilizing phonetic content to better modeling each digit segment, we change the hypothesis of conventional i-vector method and train a phonetically-constrained i-vector extractor for all digits. And a text-dependent backend is used to compensate for the channel variability. We finally make decision by combining the digit scores. When compared with previous works, the proposed method is more reliable and convenient in practical applications. Since that we only need to optimize one system instead of ten as subsystem based method with limited training data, and spend low computational complexity without fusing multiple systems as [13]. Moreover, experiments show that our proposed approach performs the best.

The remainder of this paper is organized as follows: Sect. 2 briefly introduces conventional i-vector based techniques. In Sect. 3, the proposed methods are explained in detail. Experiments are presented in Sect. 4 and we finally conclude in Sect. 5.

2. I-Vector Based System

In this section, we review the basics of i-vector system in order to facilitate the description with the proposed techniques in the following section.

2.1 I-Vector Extraction

An i-vector extractor projects the sequence of feature vectors onto a lower dimensional total variability space. For the purpose of performing the projection, Baum-Welch statistics are collected for each mixture component of UBM. Given a speech segment, the speaker and channel dependent super-vector M of concatenated GMM means is modeled as

$$M = m + Tw \tag{1}$$

where m is the UBM mean super-vector, T refers to total variability matrix of low-rank and w is a latent variable with standard normal distribution. For each speech segment, its i-vector is the maximum-a-posterior point estimate of latent variable w.

2.2 Scoring and Channel Compensation

Channel compensation is crucial for the good performance of i-vector speaker verification system. The speaker verification score can be obtained by directly computing the cosine distance between the enrollment i-vector ω_e and the test i-vector ω_t

$$CD_{e,t} = \frac{\omega_e^T \omega_t}{\|\omega_e\| \|\omega_t\|} \tag{2}$$

In this modeling however, the system is unable to model and reduce channel effects, consequently makes that the i-vector extraction simply plays the role of feature extractor rather than modeling speaker and channel effects. In order to compensate for the channel variability, linear discriminant analysis (LDA) is applied to precondition i-vectors. LDA transformation attempts to minimize the intra-class variance caused by channel effects while maximizing the variance between speakers. This is achieved by optimizing the following ratio

$$J(v) = \frac{v^T S_b v}{v^T S_w v} \tag{3}$$

In Eq. (3), S_b is the between-class variance, S_w is the within-class variance and v is the space direction. The optimization of this ratio is used to define a transformation matrix A, which is the generalised eigenvectors with highest eigenvalues of the equation

$$S_b v = \lambda S_w v \tag{4}$$

The i-vectors are then multiplied by using transformation matrix A^T.

In addition to the cosine similarity scoring, PLDA is more often used to compute the log-likelihood ratio (LLR) of the given pair of i-vectors

$$LLR_{e,t} = log \frac{p(\omega_e, \omega_t|H_s)}{p(\omega_e, \omega_t|H_d)} \tag{5}$$

where H_s is the hypothesis that the two i-vectors belong to the same speaker and H_d is the contrary. In PLDA modeling, it assumes that i-vector ω can be the combination of three term

$$\omega = \mu + Uy + \varepsilon \tag{6}$$

where μ is the i-vector mean, y is a speaker factor having a normal prior distribution, U is the speaker subspace matrix, and ε is the residual term with a zero mean and full covariance matrix Gaussian distribution. Before PLDA scoring, the i-vectors are length normalised [14] to make the distribution of the i-vectors more Gaussian-like.

In the case of multiple segments for each speaker enrollment, we simply average i-vectors from the given speaker.

2.3 Score Normalization

Score normalization is widely used in speaker verification. In our work, symmetric normalization (s-norm) is applied since it has been found to perform the best in text-dependent speaker verification task [20]. For the obtained evaluation score $s_{e,t}$ of a pair of enrollment and test utterances, its normalized score $s_{e,t}{}'$ is computed as

$$s_{e,t}{}' = \frac{s_{e,t} - \mu_1}{\sigma_1} + \frac{s_{e,t} - \mu_2}{\sigma_2} \tag{7}$$

where μ_1 and σ_1 are the mean and standard deviation of scores between enrollment target in evaluation data and test segment of imposter cohort in training data, while μ_2 and σ_2 are obtained by scoring the trial of test segment and the imposter target, from evaluation and training set respectively.

3. Proposed Speaker-Phonetic I-Vector Modeling

In this section, we give a detailed description of the proposed speaker-phonetic i-vector modeling method. The proposed method starts with obtaining digit segments from the database, and then applies i-vector based system to model speaker and phonetic information over specific digit segments. The training methods for phonetically-constrained i-vector extractor and channel compensation used in the random digit string task are presented. In addition, we describe a score compensation method in order to compensate for the unreliability of segmentation.

3.1 Segmentation

To segment the utterance into digits, previous methods such as DTW and semi-continuous HMM with states corresponding to digits seem to be suitable. However, the main difficulty here is the co-articulation effects caused by the randomness of digit strings, which remains a problem for these alignment methods. In order to obtain accurate digit segments, we use the traditional continuous density HMMs method since it better models the context of speech to obtain more reliable alignment information.

After HMMs decoding, we then get alignment labels of the dataset and segment utterances into digits. Considering the prerequisite for doing cepstral mean subtraction is phonetically balanced in the segment [21], situation of one digit per segment in the proposed approach is unsuitable to apply mean normalization. So, our works do not directly segment the waveform of a given utterance into pieces in the first place. Actually, we do mean normalization on the whole utterance and finally use its forced alignment labels to select feature frames of the corresponding digit group to complete the segmentation process.

3.2 Speaker-Phonetic I-Vector Modeling

The basic idea of the approach is to turn the randomized pass-phrase task into a fixed pass-phrase one. Noting that digit is constrained here, it is appropriate to make each pair of enrollment and test segments to be of the same digit. In this term, the modeling phonetic content is constrained to a specific digit.

To be specific, the proposed scheme first estimates a UBM by using the concatenate features of valid voice pieces. Then a single i-vector extractor is obtained by applying our phonetically-constrained training way. After that, given a target speaker's enrollment utterances that each consists a sequence of random number covering all 10 digits from 0 to 9, we can easily get 10 groups of digit segments according to the forced alignment label. Using the i-vector extractor, an i-vector is extracted for each digit of the speaker. During evaluation, after segmenting the test utterance, the i-vector for each test digit is extracted and followed by cosine similarity scoring (directly or transformed by LDA) or PLDA evaluating with the corresponding enrollment digit i-vector. The score of the test utterance is finally computed by averaging all digit scores.

$$s_{e,t} = \frac{1}{N} \sum_{n=0}^{M} \delta_{t,n} score(\omega_{e,n}, \omega_{t,n}) \tag{8}$$

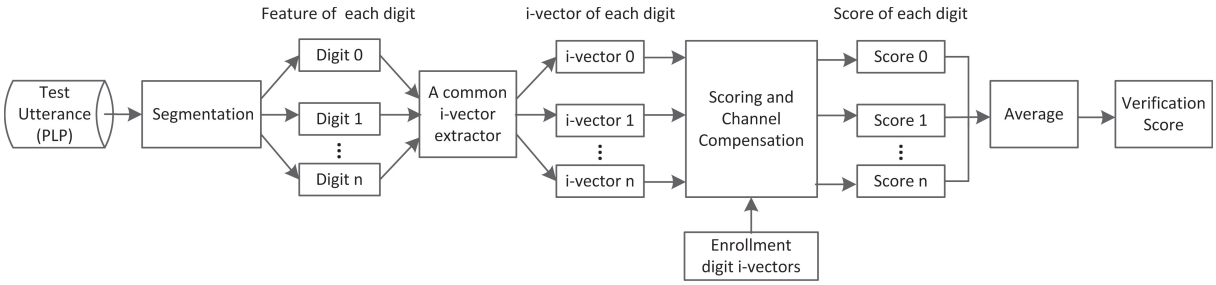

Fig. 1 Block diagram of speaker-phonetic i-vector modeling framework.

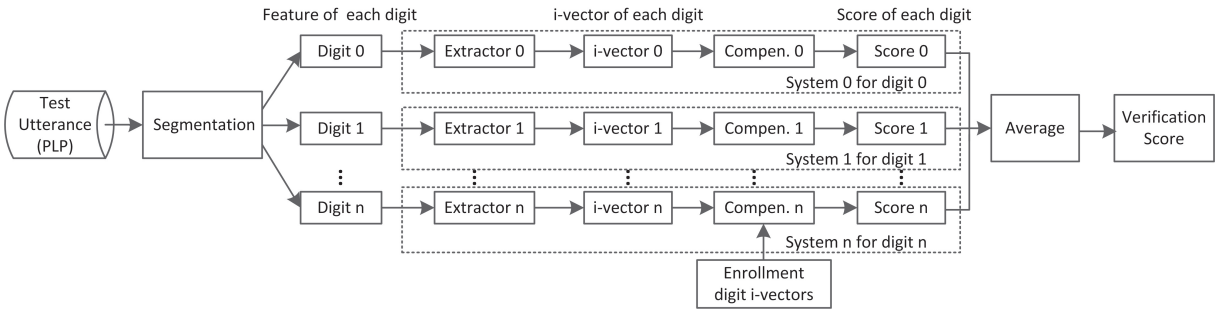

Fig. 2 Block diagram of i-vector based subsystems modeling framework. In the framework, the Extractor n and Compen. n are the i-vector extractor and the channel compensation (including the LDA or PLDA) built specifically for digit n.

where, $\omega_{e,n}$ and $\omega_{t,n}$ are the i-vectors for digit n of the enrollment and test utterance, respectively. And $M = 9$, $N = \sum \delta_{t,n}$, $\delta_{t,n} = 1$ when digit n is in the test utterance, $\delta_{t,n} = 0$ otherwise. Figure 1 shows the block diagram of the proposed verification framework. Instead of building subsystem for every digit as [16], [17], the speaker-phonetic i-vector modeling method uses a single system with a common i-vector extractor and channel compensation model for all digits. Figure 2 gives the block diagram of i-vector based subsystem modeling framework as described in [16], [17], each dotted box is a subsystem trained specifically for digit n, n=0,...,9. By comparing between Fig. 1 and Fig. 2, the proposed speaker-phonetic i-vector modeling framework can show its specific characteristics, that is, the same i-vector extractor and the same channel compensation model are used for processing of all ten digits. The intentions of the speaker-phonetic i-vector modeling method are to avoid sparsity of training data and variability of subsystems (97 speakers of each digit are unreliable for training a subsystem). From this perspective, the proposed speaker-phonetic i-vector modeling method is more suitable to use in practical applications. Next section will present the training methods for i-vector, LDA and PLDA in detail.

3.3 Phonetically-Constrained I-Vector Extractor and Channel Compensation

The i-vector extraction here differs from the conventional one on the hypothesis of Eq. (1). In normal, the total variability space T estimation treats entire set of utterances as having been produced by different speakers. In other words, it models both speaker and channel factors, supposing that words in an utterance share a common latent variable. The T space mixes up content information by collecting Baum-Welch statistics and is effective in the phrase level of granularity. However, it is incompetent to our task. As the speaker-phonetic i-vector framework enables us to evaluate segments corresponding to the same digit from the enrollment and test utterances, it demands that the i-vector approach can well model both speaker and phonetic variabilities in term of one digit per segment. So, we make a change to the hypothesis of Eq. (1) and model the speaker and channel super-vector M_j of the j'th digit segment in a given utterance as

$$M_j = m + T\omega_j \tag{9}$$

where ω_j is a latent variable for the j'th digit segment. In Eq. (9), we assume that different digits have different latent variables. The new hypothesis changes the modeling level of granularity from the phrase to the digit. The other words, it constrains the phonetic content to be a specific digit number instead of a phase of mixing content, and is supposed to be more targeted for prominently modeling the speaker variability with the phonetic information of one digit per segment. In this work, we segment the utterance into digits, and the training data with one digit per segment is used to estimate a common phonetically-constrained T space for i-vector extraction.

With the phonetic content modeling in term of separate digits for an utterance, text-dependent training setting is appropriate for LDA and PLDA. We also use the phonetically-constrained modeling method proposed in [22] in our exper-

iments, and define each class for estimation as the combination of both speaker and phonetic content of digit. By this setting, the true number of classes is 970 (97 speakers and 10 digits), which therefore is sufficient for well training the LDA and PLDA.

3.4 Compensation for Scores

It is worth mentioning that the force alignment label from continuous density HMMs cannot be completely correct. The wrong identification result and the unperfect frame boundary for digit can bring unreliability to the evaluation of our system. To take into account the uncertainties in segmentation with the force alignment label, confidence measure (CM) [23] is used to weight the digit scores before combination. The decision score is computed as

$$\tilde{s}_{e,t} = \frac{1}{\tilde{N}} \sum_{n=0}^{M} \alpha_{t,n} score(\omega_{e,n}, \omega_{t,n}) \tag{10}$$

where confidence for digit n of the test utterance is represented as $\alpha_{t,n}$, and $\tilde{N} = \sum \alpha_{t,n}$, $\alpha_{t,n} = 0$ if digit n is not appear in the test utterance.

4. Experiments and Results

This section describes the database used for evaluation, the building-up of our experiments and the comparison results. In all experiments, we report results in terms of EER and minimum Normalized Detection Cost Function as defined for NIST SRE08 (minDCF_{08}) and NIST SRE10 (minDCF_{10}). In detection error tradeoff (DET) curves, the square and star markers correspond to minDCF_{08} and minDCF_{10} points, respectively. Results statistic and DET curves are obtained by using the BOSARIS toolkit [25].

4.1 Database

Experiments are conducted on the RSR2015 (Part 3) database, a publicly available speech corpus recorded by Institute for Infocomm Research in Singapore [9]. Part 3 of RSR2015 database is devoted to speaker verification using randomly prompted English digit strings, and is divided into background (*bkg*), development (*dev*) and evaluation (*eval*) subsets. Table 1 shows the number of speakers for each subset. Six mobile devices were used for recording the database. Three of them were assigned to each speaker, then the speaker used each device to record three sessions of the prompted sequences. Thus, we have nine sessions in total, they each contains 3 10-digit and 10 5-digit utterances. Each speaker model is enrolled with 3 10-digit utterances in the same session, and three sessions (session 1,4,7 according to the protocol in [9]) recorded with the same handset are chose for enrollment, while 5-digit utterances from the remaining six sessions are used to build the test set.

Table 1 Number of speakers in RSR2015.

Subset	Male	Female
bkg	50	47
dev	50	47
eval	57	49

Table 2 Number of trials in RSR2015 Digits.

Subset	Gender	Target	Nontarget
dev	Male	5154	251310
dev	Female	5061	232806
eval	Male	6120	342720
eval	Female	5283	253584

4.2 Experimental Setups

We use standard 20-dimensional PLPs with its first and second derivatives to form the feature of 60-dimension followed by mean and variance normalization. Features in our experiments are extracted from 25 ms Hamming windowed signals with 15 ms overlaps by using Kaldi [24].

We also adopt Kaldi toolkit to train the continue HMMs. Contrary to text-independent tasks, we do not apply voice activity detection (VAD) directly to the utterances, as VAD errors would do harm to the Viterbi alignment. Therefore, we add silence labels to the beginning and the end of each utterance and use a silence HMM model for silences. Then, the process of VAD is done by dropping the frames aligned to silence model. Continuous HMMs with feature vector doing maximum likelihood linear transformation after applying LDA are trained to classify 1164 triphone tied states (senones). We use the *bkg* subset and digit context of each utterance to train the continuous HMMs. The decoder of HMMs gives the alignment results in the way of labeling the beginning and the end times for each digit.

We take GMM-UBM approaches as our benchmark since it performs well due to the exceptionally low channel effect in the RSR2015. It needs to be emphasized that all the processes are on gender-independent setting unless stated otherwise. The UBM, i-vector extractor, LDA and PLDA training are performed by using the *bkg* subset. The number of mixture components for UBM is 256. We use a total variability space of 400 factors, LDA transformation projecting the i-vectors to 200-dimension and 400-dimension PLDA.

The *dev* and *eval* subset are used for verification, the target-correct (target speaker with correct pass-phrase) and the imposter-correct (imposter with correct pass-phrase) trials are used as target and nontarget since our concerns with verifying speaker's identities instead of the lexical context. The total number of trials is given for each gender in Table 2. Compared with the trials in [13], we evaluate without any wave file selection, which is more challenge for achieving better performance. However, there is still a little difference in the male trial for *eval* compared with [9], as we experimentally find that one utterance has been recorded with all silence and thus is automatically deleted. The set of enrollment and test utterances on *bkg* is used for s-norm.

4.3 Results

4.3.1 Comparison with the Benchmark

In this section, we compare the speaker-phonetic i-vector modeling method with the benchmark. In order to highlight the effectiveness of applying the speaker-phonetic modeling scheme, we also include text-independent i-vector system for comparison. Experiments in this section use an i-vector extractor trained in the conventional way and only the cosine similarity scoring for the i-vector based systems. Table 3 and Table 4 present the results on *dev* and *eval* sets. The notation within matrix entries means male/female.

From these result tables, it is clear that GMM-UBM outperforms the text-independent i-vector system (Text-in./con., the notation /con. means the use of conventional i-vector extractor) and thus confirms the discussion in Sect. 1. However, better performance can be achieved by combining the speaker-phonetic scheme with i-vector extraction (Spk-phon./con.). To make it obvious, we see that when compared to GMM-UBM, the Spk-phon./con. system can get significant reduction in EER, minDCF_{08} and minDCF_{10} by a factor of 24.03%, 35.92% and 33.10% on male trial of *eval* set, respectively. This result can be attributed to evaluating in a text-dependent situation as designed by speaker-phonetic framework.

4.3.2 Phonetically-Constrained I-Vector Extractor vs. Conventional I-Vector Extractor

In Sect. 3.3, we presented that the proposed phonetically-constrained i-vector extractor which takes the phonetic content into consideration is more suitable for speaker-phonetic i-vector modeling. In order to prove it experimentally, we compare the performances of speaker-phonetic i-vector system with i-vector extractors in the conventional and the phonetically-constrained (Spk-phon./phon., the notation /phon. means the use of phonetically-constrained i-vector extractor) training ways in Table 5. The results indicate that using the phonetically-constrained training way leads to a big improvement in performance in all criteria (EER, minDCF_{08} and minDCF_{10}).

Table 3 Results on RSR2015 Digits, *dev* trials

System	EER(%)	minDCF_{08}	minDCF_{10}
Text-in./con.	6.02/10.45	0.274/0.459	0.668/0.867
GMM-UBM	5.07/8.77	0.214/**0.366**	0.534/0.792
Spk-phon./con.	**4.63/7.84**	**0.196**/0.376	**0.510/0.780**

Table 4 Results on RSR2015 Digits, *eval* trials

System	EER(%)	minDCF_{08}	minDCF_{10}
Text-in./con.	5.96/10.36	0.307/0.476	0.854/0.901
GMM-UBM	5.20/7.97	0.284/0.380	0.855/**0.788**
Spk-phon./con.	**3.95/7.48**	**0.182/0.375**	**0.572**/0.810

4.3.3 The Effects of Compensations for Channel and Score

Table 6 shows the results of applying LDA preconditioning and PLDA scoring for channel compensation. The great degradation in all criteria indicates that the speaker-phonetic i-vector modeling framework can be followed by trainable backends to gain more promotion in performance. Spk-phon./phon. with PLDA scoring as the backend performs the best in most criteria.

To compensate for the uncertainty in segmentation, we weight the digit scores with CM. Results on *eval* subset are shown in Table 7, and the corresponding DET curves are given in Fig. 3 and Fig. 4. We observe limited performance improvement after applying CM scores compensation, however the EER, minDCF_{08} and minDCF_{10} of the conventional i-vector extractor system even get worse. These results are due to the accurate segmentation with force alignment label obtained by continuous HMMs, which can achieve 4.12% and 3.60% in word error rate on male and female, respectively. The deterioration in Spk-phon./con. after compensating with CM shows the weakness of conventional subspace in modeling digit segment. The proposed Spk-phon./phon. system with PLDA and CM scores compensation drops the EER to 2.48% and 3.71% on the *eval* set on male and female, respectively. And when compared with the GMM-UBM, the relative reductions are 52.3% and 53.5%.

We want to make a comparison with the JFA followed by JDB in [13], which fuses 6 systems to attain the EER

Table 5 Comparison of results by using different i-vector extractor training methods.

Subet	System	EER(%)	minDCF_{08}	minDCF_{10}
dev	Spk-phon./con.	4.63/7.84	0.196/0.376	0.510/0.780
dev	Spk-phon./phon.	**4.17/6.56**	**0.175/0.328**	**0.476/0.758**
eval	Spk-phon./con.	3.95/7.48	0.182/0.375	0.572/0.810
eval	Spk-phon./phon.	**3.36/6.25**	**0.163/0.320**	**0.533/0.781**

Table 6 Results with channel compensation.

Subset	System	EER(%)	minDCF_{08}	minDCF_{10}
dev	Spk-phon./con.	4.63/7.84	0.196/0.376	0.510/0.780
dev	+ LDA	4.03/4.72	0.174/0.247	0.503/0.672
dev	+ PLDA	3.71/4.47	0.169/0.243	0.495/0.676
dev	Spk-phon./phon.	4.17/6.56	0.175/0.328	0.476/0.758
dev	+ LDA	3.55/4.24	0.155/0.218	0.483/0.619
dev	+ PLDA	**3.29/3.96**	**0.148/0.209**	**0.468/0.599**
eval	Spk-phon./con.	3.95/7.48	0.182/0.375	0.572/0.810
eval	+ LDA	3.25/5.09	0.158/0.266	0.523/0.737
eval	+ PLDA	2.91/4.84	0.153/0.248	0.522/0.710
eval	Spk-phon./phon.	3.36/6.25	0.163/0.320	0.533/0.781
eval	+ LDA	2.84/4.14	0.138/0.223	**0.472**/0.649
eval	+ PLDA	**2.51/3.88**	**0.133/0.201**	0.476/**0.583**

Table 7 Results on RSR2015 Digits, *eval* trials by compensation with confidence.

System	EER(%)	minDCF_{08}	minDCF_{10}
Spk-phon./con.+cm	3.97/7.49	0.183/0.376	0.573/0.819
Spk-phon./phon.+cm	3.31/6.18	0.162/0.318	0.532/0.785
+ LDA,cm	2.81/4.06	0.137/0.220	**0.469**/0.650
+ PLDA,cm	**2.48/3.71**	**0.131/0.198**	0.472/**0.586**

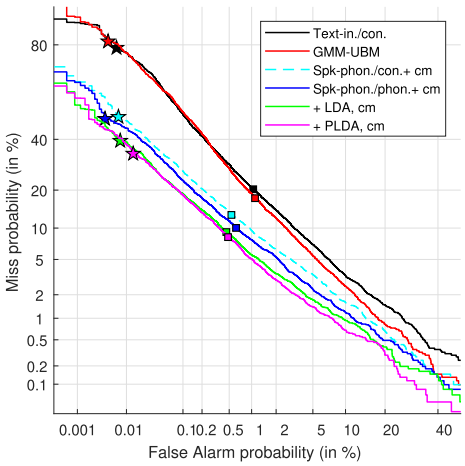

Fig. 3 Results on RSR2015 Digits, male - *eval* set.

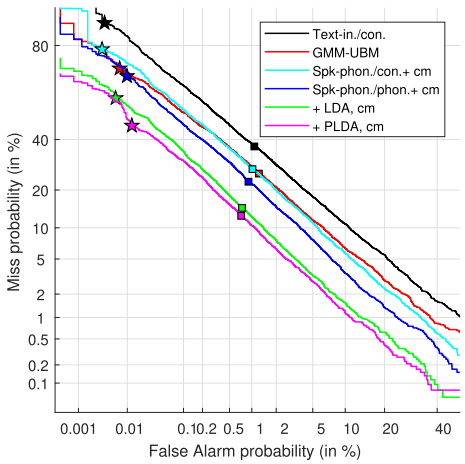

Fig. 4 Results on RSR2015 Digits, female - *eval* set.

equal to 2.61% and 3.76% on male and female (we concern only the random digit situation and thus compare with the results using regular s-norm), respectively. However, difference between evaluation trials with and without wave file selection makes it inappropriate to tell which result is better. The work [13] rejects utterances that are too short or too noisy, and leaves out speaker models with less than 3 enrollment utterances. Thus, when compared with [13], our experiments are carried out in a much worse data condition, but better results are achieved. So, the proposed Spk-phon./phon. with PLDA and CM scores compensation is competitive with the fusing systems in [13].

4.3.4 The Effects of Increasing the Number of Training Speakers

We also include experiments using both *bkg* and *dev* subsets as training data, intend to show the effects of increasing the number of training speakers. Gender-dependent setting is applied to take full use of the increasing amount of training data, since the gender-dependent systems perform better in text-dependent task, especially for female [26]. Noting that we use the gender-independent setting when restricted the use of *bkg* data only for training, because it is insufficient to fine train the gender-dependent models (total variability space, LDA and PLDA) with an extremely small amount of speakers (47 females and 50 males) and limited duration of utterances.

Specifically, the utterances of 100 female and 94 male speakers are used to train the gender-dependent UBMs, i-vector extractors, LDAs and PLDAs. The continuous HMMs system remains gender-independent setting and is trained with all the training data. Results are shown in Table 8, we can see significant improvements (especially for females) attained for all the results, as the adding of the training data and the gender information make the transformations in systems to be more robustly estimated.

Table 8 Results on RSR2015 Digits, *eval* trials by adding *dev* subset to the training data.

System	EER(%)	minDCF_{08}	minDCF_{10}
GMM-UBM	4.79/7.33	0.261/0.350	0.812/0.811
Spk-phon./con.+cm	3.63/6.30	0.174/0.327	0.572/0.774
+ LDA,cm	2.60/3.93	0.127/0.220	0.431/0.688
+ PLDA,cm	2.43/3.62	0.122/0.208	0.440/0.643
Spk-phon./phon.+cm	3.00/4.71	0.152/0.249	0.505/0.704
+ LDA,cm	2.34/3.14	0.115/0.177	**0.416**/0.603
+ PLDA,cm	**2.17/2.99**	**0.110/0.161**	0.431/**0.527**

Table 9 Comparison with previous results on RSR2015 digits.

System	EER(%)	minDCF_{08}	minDCF_{10}
HiLAM [9]	5.32/10.87	0.326/0.469	-/-
JFA-JDB [13]	2.61/3.76	0.139/0.195	0.523/0.623
Seg.DNN [16]	2.47/3.44	0.131/0.164	-/-
Spk-phon./phon. + PLDA,cm	**2.17/2.99**	**0.110/0.161**	**0.431/0.527**

4.3.5 Comparison with Previous Methods

In this section, our best results are compared with previous stat-of-the-art methods on RSR2015 digits. For JFA-JDB [13], we compare with its best results using regular s-norm as we concern only the random digit situation. For HiLAM [9] and Seg.DNN [16], the minDCF_{10} results are not available. However, they can still be evaluated in terms of EER and minDCF_{08}. The comparison results are in Table 9. Obviously, we achieve the best results and improve results of HiLAM, JFA-JDB and Seg.DNN with a large margin.

5. Conclusion and Future Work

In this paper, we proposed a speaker-phonetic framework with phonetically-constrained i-vector modeling method for text-dependent speaker verification with random digit strings, enabling us to use the digit alignment information in an i-vector framework and get decision scores in the fixed-phrase situation. We first trained a force alignment system, and then used its alignment labels for the dataset to segment utterances into digits. With the digit segments of training

data, a phonetically-constrained i-vector extractor was built to better model both speaker variance and phonetic content for digit segments. Finally, simple channel compensation methods were applied by using the text-dependent training setting, and confidence information of force alignment was used to compensate for the digit scores. Experimental results showed that the proposed Spk-phon./phon. with PLDA and CM scores compensation significantly improved the verification performance over the GMM-UBM approach and attained EER equal to 2.48% and 3.71%, minDCF_{08} equal to 0.131 and 0.198 and minDCF_{10} equal to 0.472 and 0.586 on male and female, respectively. By increasing the number of training speakers, we achieved an extremely effective system with EER equal to 2.17% and 2.99% on RSR2015 digits for male and female respectively.

We also showed that due to the limited number of training speakers, the gender-independent modeling was less robustly estimated. In the future, we will use larger databases to build gender-dependent setting for testing the proposed method.

Acknowledgments

This work is partially supported by the National Key Research and Development Program (Nos. 2016YFC0800503).

References

[1] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," IEEE Trans. Audio, Speech, Language Process., vol.15, no.4, pp.1435–1447, May 2007.

[2] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Trans. Audio, Speech, Language Process., vol.19, no.4, pp.788–798, May 2011.

[3] National Institute of Standards and Technology, "Speaker recognition evaluation," available at http://www.nist.gov/speech/tests/spk

[4] P. Kenny, T. Stafylakis, P. Ouellet, M.J. Alam, and P. Dumouchel, "PLDA for Speaker Verification with Utterances of Arbitrary Duration," Proc. IEEE Int. Conf. Acoustics, Speech Signal Process., pp.7649–7653, Vancouver, Canada, 2013.

[5] S. Cumani, O. Plchot, and P. Laface, "Probabilistic linear discriminant analysis of i-vector posterior distributions," Proc. IEEE Int. Conf. Acoustics, Speech Signal Process., pp.7644–7648, Vancouver, Canada, 2013.

[6] S. Cumani, O. Plchot, and P. Laface, "On the use of i-vector posterior distributions in Probabilistic Linear Discriminant Analysis," IEEE Trans. Audio, Speech, Language Process., vol.22, no.4, pp.846–857, April 2014.

[7] L. Li, D. Wang, C. Zhang, and T.Z. Zheng, "Improving short utterance speaker recognition by modeling speech unit classes," IEEE Trans. Audio, Speech, Language Process., vol.24, no.6, pp.1129–1139, June 2016.

[8] T. Stafylakis, P. Kenny, M.J. Alam, and M. Kockmann, "Speaker and channel factors in text-dependent speaker recognition," IEEE Trans. Audio, Speech and Language Processing, vol.24, no.1, pp.65–78, Jan. 2016.

[9] A. Larcher, K.A. Lee, B. Ma, and H. Li, "Text-dependent speaker verication: Classifiers, databases and RSR2015," Speech Commun., vol.60, no.3, pp.56–77, May 2014.

[10] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel, "I-Vector/PLDA variants for text-dependent speaker recognition," Aug. 2013, available: http://www.crim.ca/perso/patrick.kenny

[11] L. Chen, KA. Lee, B. Ma, W. Guo, H. Li, and L.R. Dai, "Phone-centric local variability vector for text-constrained speaker verification," Proc. Interspeech, 2015.

[12] H. Zeinali, L. Burget, H. Sameti, and J. Cernockỳ, "Spoken Pass-Phrase Verification in the i-vector Space," Speaker&Language Recognition Workshop, 2018.

[13] T. Stafylakis, M.J. Alam, and P. Kenny, "Text-dependent speaker recognition with random digit strings," IEEE Trans. Audio, Speech, Language Process., vol.24, no.7, pp.1194–1203, July 2016.

[14] D. Garcia-Romero, and CY. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," Proc. Annual Conference of the International Speech Commun. Association, Florence, Italy, pp.249–252, 2011.

[15] S. Novoselov, T. Pekhovsky, A. Shulipa, and A. Sholokhov, "Text-dependent GMM-JFA system for password based speaker verivication," Proc. IEEE Int. Conf. Acoustics, Speech Signal Process., Florence, Italy, 2014.

[16] J. Yan, X. Lei, G. Wang, and Z.H. Fu, "A Segmental DNNi-vector Approach for DigitPrompted Speaker Verification," Proc. Asia-pacific Signal&Information Processing Association Summit&Conference, pp.1–5, 2017.

[17] P. Chen, Q. Wu, and G. Hu, "Digit-dependent local i-vector for text-prompted speaker verification with random digit sequences," Proc. International Symposium on Chinese Spoken Language Processing, pp.1–5, 2017.

[18] H. Zeinali, E. Kalantari, H. Sameti, and H. Hadian, "Telephony text-prompted speaker verification using i-vector representation," Proc. IEEE Int. Conf. Acoustics, Speech Signal Process., pp.4839–4843, 2015.

[19] O. Büyük, "Sentence-HMM state-based i-vector/PLDA modelling for improved performance in text dependent single utterance speaker verification," let Signal Processing, vol.10, no.8, pp.918–923, Oct. 2016.

[20] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," Proc. Odyssey, p.14, 2010.

[21] R. Mammone, X. Zhang, and R. Ramachandran, "Robust speaker recognition: A feature-based approach," IEEE Signal Process. Mag., vol.13, no.5, pp.58–71, Sept. 1996.

[22] A. Larcher, KA. Lee, B. Ma, and H. Li, "Phonetically-constrained PLDA modeling for text-dependent speaker verification with multiple short utterances," Proc. IEEE Int. Conf. Acoustics, Speech Signal Process., pp.4052–4056, Vancouver, Canada, 2013.

[23] H. Jiang, "Confidence measures for speech recognition: A survey," Speech Commun., vol.45, no.4, pp.455–470, April 2005.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," IEEE Workshop on Automatic Speech Recognition and Understanding, 2011.

[25] N. Brümmer, and ED. Villiers, "The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF," Proc. NIST SRE Analysis Workshop, Atlanta, GA, 2011.

[26] A. Kanervisto, V. Vestman, M. Sahidullah, V. Hautamäki, and T. Kinnunen, "Effects of gender information in text-independent and text-dependent speaker verification," Proc. IEEE Int. Conf. Acoustics, Speech Signal Process., New Orleans, USA, 2017.

Shengyu Yao received the B.E. degree from Tsinghua University. He is now a M.S. & Ph.D. candidate of Key Laboratory of Speech Acoustics and Content Understanding at Institute of Acoustics, Chinese Academy of Sciences (IACAS), University of Chinese Academy of Sciences. His research interests include machine learning, speech signal processing and robust speaker recognition.

Ruohua Zhou received the B.S. degree from the Electronics Engineering Department, Beijing Institute of Technology, Beijing, China, in 1994, the M.S. degree of engineering in microelectronics and semiconductor devices from Microelectronics R&D Center, Chinese Academy of Sciences, Beijing, in 1997, and the Ph.D. degree from the Signal Processing Laboratory (LTS), Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. Currently he is a Professor at Key Laboratory of Speech Acoustics and Content Understanding at IACAS.

Pengyuan Zhang received the Ph.D. in the Institute of Acoustic, Chinese Academy of Sciences. He is a now a researcher in the Key Laboratory of Speech Acoustics and Content Understanding at IACAS. His research interests include large vocabulary continuous speech recognition. He is currently a member of IEICE.