DOI 10.1007/s11265-015-1019-z

Generalized I-vector Representation with Phonetic Tokenizations and Tandem Features for both Text Independent and Text Dependent Speaker Verification

Ming Li[1,2] · **Lun Liu**[3] · **Weicheng Cai**[4] · **Wenbo Liu**[1,5]

Received: 14 January 2015 / Revised: 26 May 2015 / Accepted: 17 June 2015

Abstract This paper presents a generalized i-vector representation framework with phonetic tokenization and tandem features for text independent as well as text dependent speaker verification. In the conventional i-vector framework, the tokens for calculating the zero-order and first-order Baum-Welch statistics are Gaussian Mixture Model (GMM) components trained from acoustic level MFCC features. Yet besides MFCC, we believe that phonetic information makes another direction that can benefit the system performance. Our contribution in this paper lies in integrating phonetic information into the i-vector representation by several extensions, forming a more generalized i-vector framework. First, the tokens for calculating the zero-order statistics is extended from the MFCC trained GMM components to phonetic phonemes, trigrams and tandem feature trained GMM components, using phoneme posterior probabilities. Second, given the zero-order statistics (posterior probabilities on tokens), the feature used to calculate the first-order statistics is also extended from MFCC to tandem feature, and is not necessarily the same feature employed by the tokenizer. Third, the zero-order and first-order statistics vectors are then concatenated and represented by the simplified supervised i-vector approach followed by the standard Probabilistic Linear Discriminant Analysis (PLDA) back-end. We study different token and feature combinations, and we show that the feature level fusion of acoustic level MFCC features and phonetic level tandem features with GMM based i-vector representation achieves the best performance for text independent speaker verification. Furthermore, we demonstrate that the phonetic level phoneme constraints introduced by the tandem features help the text dependent speaker verification system to reject wrong password trials and improve the performance dramatically. Experimental results are reported on the NIST SRE 2010 common condition 5 female part task and the RSR 2015 part 1 female part task for text independent and text dependent speaker verification, respectively. For the text independent speaker verification task, the proposed generalized i-vector representation outperforms the i-vector baseline by relatively 53 % in terms of equal error rate (EER) and norm minDCF values. For the text dependent speaker verification task, our proposed approach also reduced the EER significantly from 23 % to 90 % relatively for different types of trials.

✉ Ming Li
liming46@mail.sysu.edu.cn

Lun Liu
liul69@mail2.sysu.edu.cn

Weicheng Cai
wilsoncai@foxmail.com

Wenbo Liu
wenboliu@andrew.cmu.edu

[1] SYSU-CMU Joint Institute of Engineering, Sun Yat-sen University, Guangdong, China

[2] SYSU-CMU Shunde International Joint Research Institute, Guangdong, China

[3] School of Mobile Information Engineering, Sun Yat-sen University, Guangdong, China

[4] School of Information Science and Technology, Sun Yat-sen University, Guangdong, China

[5] Department of ECE, Carnegie Mellon University, Pittsburgh, PA, USA

Published online: 02 July 2015

Keywords Text independent speaker verification · Text dependent speaker verification · I-vector · Tandem feature · Language identification

1 Introduction

Total variability i-vector modeling has gained significant attention in speaker verification due to its excellent performance, compact representation and small model size [3]. In this framework, first, zero-order and first-order Baum-Welch statistics are calculated by projecting the acoustic level Mel-frequency cepstral coefficients (MFCC) features on those Gaussian Mixture Model (GMM) components using the occupancy posterior probability. Second, in order to reduce the high dimension of the concatenated statistics supervectors, a single factor analysis is adopt to generate a low dimensional total variability space which jointly models language, speaker and channel variabilities all together [4]. Third, within this i-vector space, variability compensation methods, such as Within-Class Covariance Normalization (WCCN) [7], Linear Discriminative Analysis (LDA) and Nuisance Attribute Projection (NAP) [1] are performed to reduce the variability for subsequent scoring methods (e.g., cosine similarity [3], Support Vector Machine (SVM) [2], sparse representation [18], Probabilistic Linear Discriminant Analysis (PLDA) [19, 23], deep belief networks [2], etc.).

Compared to the back-end modeling methods (as mentioned above), the frontend concatenated first-order statistics supervectors before the i-vector factor analysis are less studied. Conventionally, in the i-vector framework, the tokens for calculating the zero-order and first-order Baum-Welch statistics are the MFCC features trained GMM components. Such choice of token units may not be the optimal solution and is pending more detailed investigation. [14] recently proposed a generalized i-vector framework where decision tree senones (tied triphone states) in a general Deep Neural Network (DNN) based Automatic Speech Recognition (ASR) system are employed as new type of tokens for calculating statistics, rather than the conventional MFCC trained GMM components. Although the features for calculating the first-order statistics remain the same (MFCC), the phonetically-aware tokens trained by supervised learning can provide better token separation and discrimination. This enables the system to compare different speakers' voices token by token with more accurate token alignment, which leads to significant performance improvement on the text independent speaker verification task. Nevertheless, there are several other phonetic units (e.g. monophone states, phonemes, trigrams, etc.) with larger scale that have the potential to be considered as tokens as well. The frame level posterior probabilities of these phonetic tokens can also be converted into tandem features followed by the standard GMM to fit the conventional GMM framework.

This motivates us to further investigate different alternative configurations of phonetic tokens and features for zero-order and first-order statistics calculation in a generalized framework and apply them to the text independent speaker verification task. First, we explore the commonly used monophone states as the phonetic tokens and extend to even larger units such as trigrams. In this way, the bag of trigrams vector in the vector space modeling [15] is exactly the zero-order statistics on these trigrams. Second, since the number of monophone states is much smaller than the number of tied triphone states, we converted the phoneme posterior probabilities into tandem features [6, 9] and then apply GMM on top of it to generate large components tokens. This is also motivated by the hierarchical phoneme posterior probability estimator in [22]. In this setup, the GMM statistics calculation remains the same as the i-vector baseline except that the GMM is trained on the tandem features.

This phoneme posterior probability (PPP) based tandem feature has been reported to be effective in both ASR [6, 9, 29] and language identification (LID) tasks [5, 27] as front end features. GMM mean supervector modeling and conventional i-vector modeling are used to model this tandem feature in [27] and [5] for LID. In both methods, the tandem feature outperformed the shifted-delta-cepstral (SDC) feature relatively by more than 30 %. We note that the conventional i-vector modeling on tandem features (in [5]) is a special case in our generalized i-vector framework where tandem features and the derived GMM components are considered as features and tokens, respectively.

Since the features for token extraction and the features for first-order statistics calculation are not necessary the same [14], we show that in terms of first-order statistics calculation, MFCC is superior than tandem features for speaker verification. We further explore the hybrid features which concatenate the acoustic MFCC and the phonetic tandem features at the frame level as a feature level fusion approach. Such setup not only achieves better performance but also directly fit the conventional i-vector framework. Moreover, the tandem feature may be less informative in noisy environments where phoneme recognition accuracy is low. Adding MFCC on top of tandem features could complement and benefit the zero-order statistics.

Furthermore, we demonstrate that the phonetic level phoneme constraints introduced by the tandem features also help the text dependent speaker verification system to reject wrong password trials and therefore improve the performance dramatically. In [8], text dependent speaker verification is defined as a speaker verification task in which the lexicon used in the test phase is a subset of the lexicon pronounced by the speaker during the enrollment. By forcing the lexicon contents of enrollment and testing phrases to be the same (verbal password), higher accuracy with shorter duration utterances can be achieved [11, 13, 21, 26]. In [13], the Hierarchical multi-Layer Acoustic Model (HiLAM) outperforms the conventional i-vector approach since the latter one does not explicitly take advantages of the temporal structure of those text dependent speech utterances. However, the HiLAM approach requires specific acoustic model composing for each known password lexicon content which may not work well on accented speech, dialect or out-of-vocabulary words. In this work, we enhance the robustness of i-vector representation against the lexicon contents by adding phoneme level phonetic tokenizations and tandem features into the generalized i-vector framework. The proposed feature level fusion approach (concatenating MFCC and tandem features together as hybrid features) significantly reduces the error rate of trials with wrong lexicon contents. However, it can also make the system vulnerable to those trials where imposter speakers utter the same password as the target speaker. The solution in this work is to fuse the i-vector baseline and the proposed hybrid-GMM-hybrid system together at the score level to achieve performance improvement for all three types of trials.

In summary, we present a generalized i-vector representation framework with phonetic tokenizations and tandem features for both text independent as well as text dependent speaker verification. The contributions are as follows: (1) The tokens for calculating the zero-order statistics is extended from the MFCC trained GMM components to a variety of other phonetic units. (2) Given the zero-order statistics, the feature for calculating the first-order statistics is also extended from MFCC to tandem features and is not necessarily the same feature employed by the tokenizer. (3) We study different system setups with different tokens and features. We show that the feature level fusion of acoustic level MFCC features and phonetic level tandem features with GMM based i-vector representation achieves the best performance for text independent speaker verification while it can keep the current system structure unchanged. (4) We demonstrate that the phonetic level phoneme constraints introduced by the tandem features help the text dependent speaker verification system to reject wrong password trials and improve the performance dramatically when the contents are wrong. But the results on those trials where imposter and target speakers utter the same lexicon content degrades. We therefore propose a score level fusion approach to achieve performance improvements for all types of trials. To our best knowledge, this is the first time that the generalized representation framework with phonetic tokenization and tandem features is applied to the i-vector based text dependent speaker verification system.

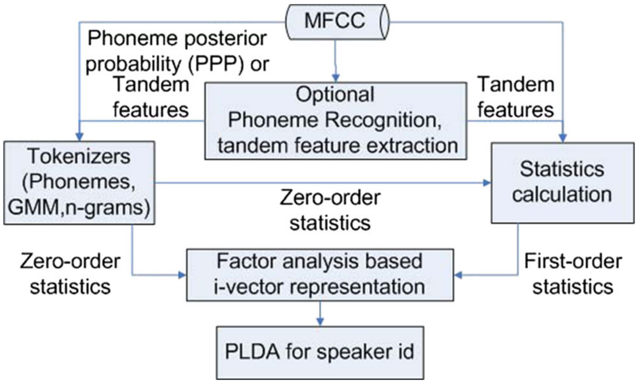
MFCC
Phoneme posterior probability (PPP) or Tandem features
Optional Phoneme Recognition, tandem feature extraction
Tandem features
Tokenizers (Phonemes, GMM,n-grams)
Zero-order statistics
Statistics calculation
Zero-order statistics
Factor analysis based i-vector representation
First-order statistics
PLDA for speaker id

Figure 1 The generalized i-vector framework.

The remainder of the paper is organized as follows. The baseline and the proposed algorithms are explained in Section 2. Experimental results and discussions are presented in Section 3 while conclusions are future works are provided in Section 4.

2 Methods

The overview of the proposed generalized i-vector framework is shown in Fig. 1. Our generalized framework extends the choices of tokens and features for statistics calculation while keeps the factor analysis, variability compensation and subsequent modeling the same way as the conventional i-vector method. Table 1 and Fig. 3 demonstrates the five different tokens that we explored in this work as well as the processes to extract them. The statistics calculation, factor analysis based i-vector baseline and our simplified version simplified supervised i-vector are first described in Section 2.1 as background. The statistics calculation with new types

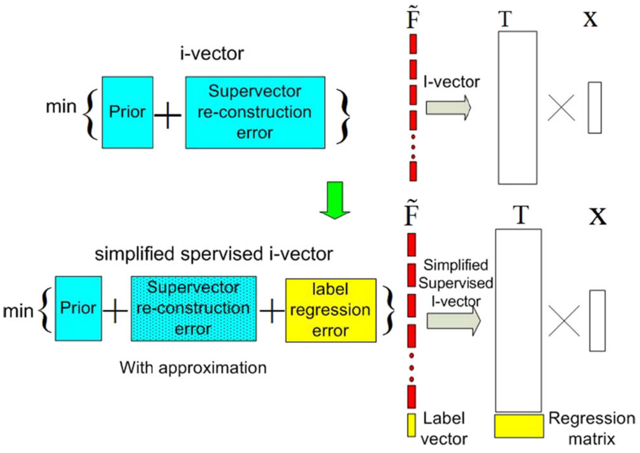

Figure 2 Schematic of the factor analysis based i-vector and simplified supervised i-vector modeling [16, 17].

Table 1 The proposed methods with different combinations of tokens and features for zero-order and first-order statistics calculation.

Methods	Tokens for zero-order statistics	Feature for first order statistics
i-vector Baseline	MFCC feature trained GMM components	MFCC
Phonemes-MFCC	Monophone states	MFCC
Tandem-GMM-MFCC	Tandem feature trained GMM components	MFCC
Trigrams-MFCC	Trigrams	MFCC
Tandem-GMM-Tandem	Tandem feature trained GMM components	Tandem
Hybrid-GMM-Hybrid	Hybrid feature trained GMM components	MFCC+Tandem

of phonetic tokens and tandem features in the generalized i-vector framework is then introduced in Section 2.2. Finally, PLDA modeling and score level fusion is presented in Section 2.3.

2.1 I-vector Baseline and the Simplified Supervised i-vector

Given a C component GMM Universal Background Model (UBM) model λ with $\lambda_c = \{p_c, \mu_\mathbf{c}, \Sigma_\mathbf{c}\}, c = 1, \cdots, C$ and an utterance with an L frame feature $\{\mathbf{y_1}, \cdots, \mathbf{y_L}\}$, the zero-order and centered first-order Baum-Welch statistics on the UBM are calculated as follows:

$$N_c = \sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda) \tag{1}$$

$$\mathbf{F_c} = \sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda)(\mathbf{y_t} - \mu_\mathbf{c}) \tag{2}$$

where $c = 1, \cdots, C$ is the GMM component index and $P(c|\mathbf{y_t}, \lambda)$ is the occupancy posterior probability for $\mathbf{y_t}$ on λ_c. The corresponding centered mean supervector $\tilde{\mathbf{F}}$ is generated by concatenating all the $\tilde{\mathbf{F}}_\mathbf{c}$ together:

$$\tilde{\mathbf{F}}_\mathbf{c} = \frac{\sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda)(\mathbf{y_t} - \mu_\mathbf{c})}{\sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda)}. \tag{3}$$

The centered mean supervector $\tilde{\mathbf{F}}$ can be projected as follows:

$$\tilde{\mathbf{F}} \rightarrow \mathbf{Tx}, \tag{4}$$

where $\mathbf{T}$ is a rectangular total variability matrix of low rank and $\mathbf{x}$ is the so-called i-vector [3]. Considering a C-component GMM and D dimensional acoustic features, the total variability matrix $\mathbf{T}$ is a $CD \times K$ matrix which is estimated the same way as learning the eigenvoice matrix in [10] except that here we consider that every utterance is produced by a new speaker [3].

As shown in Fig. 2, we recently proposed the simplified supervised i-vector method [16, 17] which achieves comparable performance to the conversional i-vector baseline and at the same time reduces the computational cost by a factor of 100. Since this method relies on the same set of statistics and is more efficient, it is employed as the factor analysis based dimensionality reduction method for all the experiments in this work.

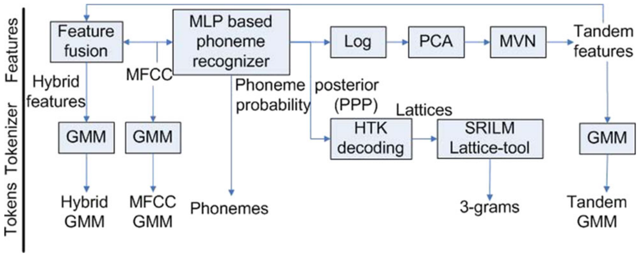

Figure 3 Tokens for zero-order statistics calculation.

2.2 Statistics Calculation in the Generalized Framework

In our generalized i-vector framework, the zero-order and first-order statistics for the j^{th} utterance are calculated as follows:

$$N_c = \sum_{t=1}^{L} P(c|\mathbf{z_t^j}, \hat{\lambda}) \tag{5}$$

$$\mathbf{F_c} = \sum_{t=1}^{L} P(c|\mathbf{z_t^j}, \hat{\lambda})(\mathbf{y_t^j} - \hat{\mu}_\mathbf{c}) \tag{6}$$

$$\hat{\mu}_\mathbf{c} = \frac{\sum_{j=1}^{J}\sum_{t=1}^{L} P(c|\mathbf{z_t^j}, \lambda)\mathbf{y_t}}{\sum_{j=1}^{J}\sum_{t=1}^{L} P(c|\mathbf{z_t^j}, \lambda)}. \tag{7}$$

where $c = 1, \cdots, C$ is the new token index and $P(c|\mathbf{z_t^j}, \hat{\lambda})$ is the posterior probability for the j^{th} utterance's feature vector at time t on the c^{th} token. Note that the feature $(\mathbf{z_t})$ used to calculate the posterior probability $P(c|\mathbf{z_t}, \hat{\lambda})$ and the feature $(\mathbf{y_t})$ for cumulating the first-order statistics $\mathbf{F_c}$ are not necessarily the same. They can be different just as shown in Table 1. Global mean $\hat{\mu}_\mathbf{c}$ is computed using all the training data in the same way as the mean parameter estimation in GMM. Similarly, we also calculated the second-order statistics for the simplified supervised i-vector modeling.

The proposed methods with different combinations of tokens and features for statistics calculation are shown in Table 1. First, in the conventional i-vector baseline, both $\mathbf{z_t}$ and $\mathbf{y_t}$ in Eqs. 5, 6 are MFCC features and the tokens are the MFCC trained GMM components. Second, in the Phonemes-MFCC system, the tokens are the monophone states and the posterior probability $P(c|\mathbf{z_t}, \hat{\lambda})$ is the phoneme posterior probability (PPP). We employed the multilayer perceptron (MLP) based phoneme recognizer [24] with acoustic models from five different languages, namely Czech, Hungarian, Russian, English and Mandarin. The models for the first three languages were trained on SpeechDat-E databases and provided in [24]. Additionally, we adopted the toolkit in [24] and trained the English and Mandarin based models both with 1000 neurons in all nets using the switchboard, fisher databases and the call friend, call home databases, respectively.

Since there are only limited amount of monophone state tokens (around 8 times less than the GMM components for English), the system performance is affected due to the broad coverage of each phoneme token. Here we propose two different methods to generate tokens with comparable size of GMM components. First, the PPP features are converted into tandem features by log transform, principal component analysis (PCA) and mean variance normalization (MVN) [6, 9, 27] as shown in Fig. 3. Then we directly consider this tandem feature as $\mathbf{z_t}$ in Eqs. 5, 6 and train a GMM on top of it to generate the Tandem-GMM tokens. In this setup, the entire GMM statistics calculation remains the same except that the GMM model is trained on the tandem features. Second, we increase the time scale of tokens and adopt the trigrams as the new type of tokens. As shown in Fig. 3, HTK toolkit [28] is used to decode the PPP features and output a lattice file for each utterance which is further processed into n-gram counts and n-gram indexes by the lattice-tool toolkit [25]. The decoded n-gram counts are considered as the posterior probability and the mean of features within this n-gram's range is accounted as $\mathbf{y_t}$ where t indexes the whole n-gram here.

Both tandem features and MFCC features can be used (as $\mathbf{z_t}$) to train a GMM tokenizer and both could be projected on tokens (as $\mathbf{y_t}$) for calculating the first-order statistics. Therefore, we further explore the hybrid features which concatenate the acoustic MFCC feature and the phonetic tandem features at the frame level for both purposes. This feature level fusion setup not only achieves better performance but also directly fit the conventional i-vector framework.

2.3 Back-end Modeling

Once we have the low dimensional i-vectors extracted from the generalized framework, PLDA is applied as the back-end modeling for both text independent and text dependent speaker verification tasks. In the text dependent task, each speaker with each lexicon content password is considered as a class and different phrases from the same speaker are labeled with separate classes in the PLDA model training. We simply employed the weighted summation fusion approach at the score level with parameters tuned by cross validation to further enhance the text dependent speaker verification performance.

3 Experimental Results

3.1 Results on text Independent Speaker Verification

We first conducted experiments on the NIST 2010 speaker recognition evaluation (SRE) corpus [20] for the text independent speaker verification task. Our focus is the female part of the common condition 5 (a subset of tel-tel) in the core task. We used equal error rate (EER) and the 2008 and 2010 normalized minimum decision cost value (norm minDCF) as the metrics for evaluation [20]. For cepstral feature extraction, a 25ms Hamming window with 10ms shifts was adopted. Each utterance was converted into a sequence of 36-dimensional feature vectors, each consisting of 18 MFCC coefficients and their first derivatives. We employed the Czech phoneme recognizer [24] to perform the voice activity detection (VAD) by simply dropping all frames that are decoded as silence or speaker noises. Feature warping is applied to mitigate variabilities.

The training data for NIST 2010 task include Switchboard II part1 to part3, NIST SRE 2004, 2005, 2006 and 2008 corpora on the telephone channel. The gender-dependent GMM UBM consists of 1024 mixture components. Token numbers are shown in Table 2 and the tandem feature dimension is 52. The sizes of i-vectors and the dimension of speaker-specific subspace in PLDA are 600 and 150, respectively. Simple weighted linear summation is adopted here as the score level fusion.

In Table 2, the English Phonemes-MFCC system outperformed the i-vector baseline (3.65 %→3.10 % EER) by using only 123 phoneme tokens which supports our claim that phonetic tokens help. Since majority of the NIST SRE data samples are from English, other language based phoneme tokens are not as effective as the English one. So the accuracy of phoneme decoder indeed could impact the SV performance. By combining systems with phoneme tokens from multiple languages improved the result. This is very useful in the multi-lingual or multi-dialects speaker verification scenarios. Furthermore, in system ID 8 and 9, we adopt the tandem-GMM components as the tokens and evaluated different features for the first-order statistics calculation. Results show that MFCC feature is better than tandem feature in this case. When applying GMM on top of

Table 2 Performance of the proposed methods on the NIST SRE 2010 core condition 5 female part task (original trials).

ID	Methods	Tokens	Token language	Token number	Feature for first order statistics	EER %	norm 08/10 minDCF
1	i-vector baseline	MFCC-GMM		1024	MFCC	3.65	0.19/0.58
2	Phonemes-MFCC	Monophone states	English	123	MFCC	3.10	0.16/0.55
3	Phonemes-MFCC	Monophone states	Mandarin	537	MFCC	4.88	0.22/0.59
4	Phonemes-MFCC	Monophone states	Czech	138	MFCC	5.97	0.25/0.55
5	Phonemes-MFCC	Monophone states	Hungarian	186	MFCC	5.56	0.24/0.64
6	Phonemes-MFCC	Monophone states	Russian	159	MFCC	6.23	0.24/0.61
7	Fusion of methods 2+3+4+5+6					2.65	0.15/0.49
8	Tandem-GMM-MFCC	Tandem-GMM	English	1024	MFCC	2.82	0.14/0.31
9	Tandem-GMM-Tandem	Tandem-GMM	English	1024	Tandem	3.34	0.16/0.40
10	Trigrams-MFCC	Trigrams	English	1024	MFCC	5.07	0.24/0.63
11	Hybrid-GMM-Hybrid	Hybrid-MFCC	English	1024	Hybrid	1.71	0.11/0.19

the tandem features, the number of tokens become comparable to the baseline GMM size which leads to the significant performance enhanced by 22.7 % relative EER reduction. Trigrams tokens based system did not improve the performance which might be because its scale is too large compared to those monophone states and the zero-order statistics vectors are sparse.

Finally, the Hybrid-GMM-Hybrid single system has achieved 1.71 % EER and 0.19 norm new 2010 minDCF, which outperformed the i-vector baseline by relatively 53 % and 67 %, respectively. This is very promising since in this setup the entire GMM i-vector framework remains the same, only features are enhanced to the hybrid ones.

3.2 Results on Text Dependent Speaker Verification

For the text dependent speaker verification task, we used the Part I female portion of the RSR2015 database as our evaluation dataset [13]. In the RSR2015 database, the number of speakers in the background, development and evaluation sets are 47, 47 and 49, respectively. We used the same front end, UBM and PLDA configuration as for our text independent experiments but the UBM, i-vector as well as the PLDA models that we tested were trained on the Part I background data. This consists of parallel recordings of 30 TIMIT phrases uttered by 47 female speakers, each of whom participated in 9 recording sessions on 3 different recording devices. We used the same development and evaluation data in [13] to demonstrate the system performance and we did not use the development data for training.

The number of trials for each of the four text dependent speaker verification scenarios on the Part I of the RSR 2015 database is shown in Table 3. We can see that only the target speaker uttering the correct lexicon content is considered as the true trial, the other cases are all non-target trials. In order to show the results for all three types of non-target trials, we evaluate the system performance separately for each type of trials the same way as in [13]. The gender-dependent GMM UBM consists of 1024 mixture components. Token numbers are shown in Table 2 and the tandem feature dimension is 52. The sizes of i-vectors and the dimension of speaker-specific subspace in PLDA are 400 and 150, respectively.

Table 4 shows the performance of the proposed systems on the development set of Part I for different definitions of target and non-target trials in terms of EER, 08 norm min DCF and 10 norm min DCF. We can see that the proposed generalized i-vector representation (system 2 and 3) outperformed the i-vector baseline (system 1) dramatically for both type 1 and type 3 trials. This is because the introduction of the phonetic level phoneme constraints introduced by the tandem features for the zero-order statistics calculation help the text dependent speaker verification system to reject wrong password trials. Furthermore, the results of system 3 is better than system 2 for both type 1 and type 3 trials. This might be because the features for calculating the first-order statistics in system 3 are the feature level fused hybrid features. The tandem feature could provide more lexicon content information which leads the performance improvement for type 1 and type 3 trials. However, since more content wise text dependent information is embedded in the hybrid feature, the system is less robust to the type 2 trials where the target and imposter speakers utter the same lexicon contents. One possible solution is to classify the type of trials first and then apply different systems for different trials [12]. In this work, we propose an alternative solution by just simply fusing system 1 and 3 at the score level which not only maintains the error reduction for type 1 and type 3 trials but also improve the performance on type 2 trials. From Table 5, we can see that the proposed fusion approach reduced the EER significantly from 23 % to 90 % relatively for different types of trials.

Table 3 Number of trials for each of the four text dependent speaker verification scenarios on the Part I of the RSR 2015 database.

Speaker	Lexical content	Trial type	Female	
			development	evaluation
Target	correct	true	8419	8631
Target	wrong	false	244123	250229
Imposter	correct	false	387230	414249
Imposter	wrong	false	5612176	6006596

Table 4 Performance of the proposed systems on the development set of Part I for different definitions of target and non-target trials in terms of EER, 08 norm min DCF and 10 norm min DCF (EER/08 norm min DCF/10 norm min DCF).

Speaker	type	Target		Imposter		System1:MFCC-	System2:Hybrid-	System3:Hybrid-
Text		Correct	Wrong	Correct	Wrong	GMM-MFCC	GMM-MFCC	GMM-Hybrid
Trials	1	tar	non	-	-	0.77 %/0.054/0.3	0.36 %/0.021/0.1	0.01 %/0/0
	2	tar	-	non	-	6.26 %/0.324/0.8	7.58 %/0.388/0.9	7.48 %/0.381/0.9
	3	tar	-	-	non	0.1 %/0.005/0	0.05 %/0.003/0	0 %/0/0

Table 5 Performance of the proposed fusion system and two reference systems on the development set of Part I for different definitions of target and non-target trials in terms of EER, 08 norm min DCF and 10 norm min DCF (EER/08 norm min DCF/10 norm min DCF).

Speaker	Target		Imposter		System 1+3	HiLAM	i-vector baseline
Text	Correct	Wrong	Correct	Wrong		[13]	[13]
Trials	tar	non	-	-	0.05 %/0.002/0	1.77 %/0.074/-	3.05 %/0.173/-
	tar	-	non	-	5.36 %/0.27/0.7	3.24 %/0.154/-	7.87 %/0.405/-
	tar	-	-	non	0 %/0/0	0.45 %/0.018/-	0.94 %/0.046/-

Table 6 Performance of the proposed systems on the evaluation set of Part I for different definitions of target and non-target trials in terms of EER, 08 norm min DCF and 10 norm min DCF (EER/08 norm min DCF/10 norm min DCF).

Speaker	type	Target		Imposter		System1:MFCC-	System2:Hybrid-	System3:Hybrid-
Text		Correct	Wrong	Correct	Wrong	GMM-MFCC	GMM-MFCC	GMM-Hybrid
Trials	1	tar	non	-	-	0.23 %/0.009/0	0.1 %/0.004/0	0.02 %/0/0
	2	tar	-	non	-	3.85 %/0.192/0.6	4.62 %/0.236/0.7	4.59 %/0.240/0.7
	3	tar	-	-	non	0.08 %/0.003/0	0.02 %/0.001/0	0 %/0/0

Table 7 Performance of the proposed fusion system and two reference systems on the evaluation set of Part I for different definitions of target and non-target trials in terms of EER, 08 norm min DCF and 10 norm min DCF (EER/08 norm min DCF/10 norm min DCF).

Speaker	Target		Imposter		System 1+3	HiLAM	i-vector baseline
Text	Correct	Wrong	Correct	Wrong		[13]	[13]
Trials	tar	non	-	-	0.02 %/0.001/0	0.61 %/0.034/-	1.91 %/0.106/-
	tar	-	non	-	2.94 %/0.151/0.5	2.96 %/0.156/-	6.61 %/0.327/-
	tar	-	-	non	0 %/0/0	0.14 %/0.008/-	0.75 %/0.036/-

Similar results are shown in Tables 6 and 7 for the evaluation set. Comparing with the state-of-the-art approaches (HiLAM and convectional i-vector) in Tables 5 and 7, our proposed method achieves significant performance improvement on type 1 and type 3 trials and comparable results on type 2 trials.

4 Conclusions and Future Works

This paper presents a generalized i-vector representation framework with phonetic tokenizations and tandem features for text independent and text dependent speaker verification tasks. First, the tokens for calculating the zero-order statistics is extended from the MFCC trained GMM components to phonetic phonemes, 3-grams and tandem feature trained GMM components using phoneme posterior probabilities. We show that the Tandem-GMM tokens are superior than the phonemes and trigrams in terms of performance. Since the features for extracting tokens and the features for calculating the first-order statistics are not necessary the same , we show that in terms of first-order statistics calculation, MFCC is superior than tandem features for speaker verification. We further explore the hybrid features which concatenate the acoustic MFCC and the phonetic tandem features at the frame level for both purposes. This setup not only achieves better performance but also fit the conventional i-vector framework. We also demonstrate that the phonetic level phoneme constraints introduced by the tandem features help the text dependent speaker verification system to reject wrong password trials and improve the performance dramatically. Score level fusion of systems with different tokens and features further improves the overall system performance.

Future work includes applying tandem feature extraction on the triphone states, increasing the GMM size and exploring other types of tokens (e.g. articulatory attributes and pattern classes).

Acknowledgments This research is funded in part by the National Natural Science Foundation of China (NSFC 61401524),Natural Science Foundation of Guangdong Province (2014A030313123), SYSU-CMU Shunde International Joint Research Institute and CMU-SYSU Collaborative Innovation Research Center.

References

1. Campbell, W., Sturim, D., & Reynolds, D. (2006). Support vector machines using gmm supervectors for speaker verification. *IEEE Signal Processing Letters*, *13*(5), 308–311.
2. Cumani, S., Brummer, N., Burget, L., & Laface, P. (2011). Fast discriminative speaker verification in the i-vector space. In *Proceedings ICASSP* (pp. 4852–4855): IEEE.
3. Dehak, N., Kenny, P., 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.
4. Dehak, N., Torres-Carrasquillo, P., Reynolds, D., & Dehak, R. (2011). Language recognition via i-vectors and dimensionality reduction. In *Proceedings INTERSPEECH* (pp. 857–860).
5. D'Haro, L.F., Cordoba, R., Salamea, C., & Echeverry, J.D. (2014). Extended phone log-likelihood ratio features and acoustic-based i-vectors for language recognition. In *Proceedings ICASSP* (pp. 5379–5383): IEEE.
6. Ellis, D.P., Singh, R., & Sivadas, S. (2001). Tandem acoustic modeling in large-vocabulary recognition, (Vol. 1 pp. 517–520): Proceedings ICASSP.
7. Hatch, A., Kajarekar, S., & Stolcke, A. (2006). Within-class covariance normalization for SVM-based speaker recognition, (Vol. 4 pp. 1471–1474): Proceedings INTERSPEECH.
8. Hébert, M. (2008). Text-dependent speaker recognition. *Springer Handbook of Speech Processing*, 743–762.
9. Hermansky, H., Ellis, D.P., & Sharma, S. (2000). Tandem connectionist feature extraction for conventional hmm systems. In *Proceedings ICASSP*, (Vol. 3 pp. 1635–1638).
10. Kenny, P., Boulianne, G., & Dumouchel, P. (2005). Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing*, *13*(3), 345–354.
11. Kenny, P., Stafylakis, T., Ouellet, P., & Alam, M.J. (2014). Jfa-based front ends for speaker recognition. In *Proceedings ICASSP* (pp. 1724–1728).
12. Larcher, A., Lee, K.A., Ma, B., & Li, H. (2014). Imposture classification for text-dependent speaker verification. In *Proceedings ICASSP* (pp. 739–743).
13. Larcher, A., Lee, K.A., Ma, B., & Li, H. (2014). Text-dependent speaker verification: Classifiers, databases and rsr2015. *Speech Communication*, *60*, 56–77.
14. Lei, Y., Scheffer, N., Ferrer, L., & McLaren, M. (2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *Proceedings ICASSP*.
15. Li, H., Ma, B., & Lee, C. (2007). A vector space modeling approach to spoken language identification. IEEE Transactions on Audio. *Speech, and Language Processing*, *15*(1), 271–284.
16. Li, M., & Narayanan, S. (2014). *Simplified supervised i-vector modeling with application to robust and efficient language identification and speaker verification*: Computer speech and language.
17. Li, M., Tsiartas, A., Van Segbroeck, M., & Narayanan, S.S. (2013). Speaker verification using simplified and supervised i-vector modeling. In *Proceedings ICASSP* (pp. 7199–7203): IEEE.
18. Li, M., Zhang, X., Yan, Y., & Narayanan, S. (2011). Speaker verification using sparse representations on total variability i-vectors. In *Proceedings INTERSPEECH* (pp. 4548–4551).
19. Matejka, P., Glembek, O., Castaldo, F., Alam, M., Plchot, O., Kenny, P., Burget, L., & Cernocky, J. (2011). Full-covariance ubm and heavy-tailed plda in i-vector speaker verification. In *Proceedings ICASSP* (pp. 4828–4831).
20. (2010). NIST: The NIST 2010 Speaker Recognition Evaluation Plan. www.itl.nist.gov/iad/mig/tests/spk/2010/index.html.
21. Novoselov, S., Pekhovsky, T., Shulipa, A., & Sholokhov, A. (2014). Text-dependent gmm-jfa system for password based speaker verification. In *Proceedings ICASSP* (pp. 729–733).
22. Pinto, J., Garimella, S., Hermansky, H., Bourlard, H., & et al. (2011). Analysis of mlp-based hierarchical phoneme posterior probability estimator. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(2), 225–241.
23. Prince, S., & Elder, J. (2007). Probabilistic linear discriminant analysis for inferences about identity (pp. 1–8): Proceedings ICCV.

24. Schwarz, P., Matejka, P., & Cernocky, J. (2006). Hierarchical structures of neural networks for phoneme. In *Proc. ICASSP*. Software available at http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context (pp. 325–328).
25. Stolcke, A., et al. (2002). Srilm-an extensible language modeling toolkit. In *Proceedings INTERSPEECH*.
26. Variani, E., Lei, X., McDermott, E., Moreno, I.L., & Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *Proceedings ICASSP* (pp. 4080–4084).
27. Wang, H., Leung, C.C., Lee, T., Ma, B., & Li, H. (2013). Shifted-delta mlp features for spoken language recognition. *IEEE Signal Processing Letters*, *20*(1), 15–18.
28. Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (1997). *The HTK book, vol. 2*: Entropic Cambridge Research Laboratory Cambridge.
29. Zhu, Q., Stolcke, A., Chen, B.Y., & Morgan, N. (2005). Using mlp features in sris conversational speech recognition system. In *Proc. INTERSPEECH*.

Ming Li received his B.S. degree in communication engineering from Nanjing University, China, in 2005 and his M.S. degree in signal processing from the Institute of Acoustics, Chinese Academy of Sciences, in 2008. He joined the Signal Analysis and Interpretation Laboratory (SAIL) at USC on a Provost fellowship in 2008 and received his Ph.D. in Electrical Engineering in May 2013. He is currently an assistant professor with SYSU-CMU Joint Institute of Engineering, an associate professor with school of mobile information processing at Sun Yat-sen University. He is also an adjunct professor with Department of Electrical and Computer Engineering at Carnegie Mellon University. His research interests are in the areas of speech and language processing, multimodal human behavior signal processing, affective computing and ultrasound based structure health monitoring.

Lun Liu is an undergraduate student in School of Mobile Information Engineering at Sun Yat-sen University. Her research interests are speech recognition and speaker verification.

Weicheng Cai is a master student in School of Information Science and Technology at Sun Yatsen University. His research interests are speech recognition and speaker verification.

Wenbo Liu is a dual-degree PhD student in Sun Yat-sen University and Carnegie Mellon University, co-advised by Prof. Ming Li and Prof Bhiksha Ramakrishnan. She obtained both her bachelor and master's degree from School of Electronic and Information Engineering, South China University of Technology. Between Feb. 2012 and Jun. 2012, she was an exchange student at Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology. Her research interests mainly focus on machine learning with applications to computer vision and speech processing. She is the winner of the best student paper in International Symposium on Chinese Spoken Language Processing (ISCSLP) 2014.