Human Operator Authentication Using Limited Voice Data: A Power Grid Dispatcher Instance

Zheng Wang(✉), Zhen Wang, Yanyu Lu, and Shan Fu

School of Electronic Information and Electrical Engineering,
Shanghai Jiao Tong University, Shanghai 200240, China
letitbe@sjtu.edu.cn

Abstract. Automatic speaker verification (ASV) has the potential to replace the error-prone and expensive human-based security check to protect the ever increasingly interconnected complex systems, such as power grid system. However, state-of-the-art ASV system relies heavily on a large amount of matched development data and adequate long duration test utterance to maintain the acceptable performance. Unfortunately, such large amounts of data are not always feasible to collect in real world application. In this paper, we propose a new method for i-vector extraction by incorporating historical test information to reduce to requirement of long test utterance duration. The historical tests are weighted by a world MAP estimator and then used in the computation of current test's Baum-Welch statistics. Meanwhile, we modify linear discriminant analysis (LDA) to reduce the requirement of matched development data. In modified LDA training, the variability between development and evaluation data is separated and the objective is to simultaneously minimize the within-class variability and domain variability when maximize the between-class variability. Experiments are conducted on data collected from power grid dispatchers. By adding historical test information, we observe consistent improvement over baseline system especially for shorter duration condition. With modified LDA, at least 63% of performance gap is recovered when system parameters are trained with mismatched development data. Finally, we integrate proposed methods in one system and apply it to power grid dispatching room scenario. Experimental results show our proposed methods achieve fair performance with limited voice data and successfully reduce the amount of data required by ASV system.

Keywords: Speaker verification · Power dispatching · I-vector · Linear discriminant analysis · Short utterance · Domain mismatch

1 Introduction

Power grid is essential for today's society as an enabling infrastructure. The efficiency and safety of power system have major consequences for maintaining stable electricity supply, supporting economic growth and ensuring national security. With the rapid development of technology, a lot of sophisticated automation has been introduced into the power system operation. Since this equipment become more complex and start to

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 423–437, 2019.
https://doi.org/10.1007/978-3-030-22507-0_33

affect each other, the risk and potential loss of malicious intrusion or attack also increase. Thus, there is an increasing need for verifying the identity of the person regarding authorized to operate the particular machine.

In this situation, conventional human-based authentication such as passwords, tokens, and manual checks is no longer considered to offer high level security alone because human operators are found one of the biggest sources of errors in complex systems [1]. For example, passwords or pin numbers are easily forgotten or forged. And even the most highly trained and alert operators are prone to fatigue and boredom after a long period of continuous work. Therefore, the biometric identification technology can be a useful supplement to existing authentication techniques.

One of the most promising biometric identification technologies is automatic speaker verification (ASV), which is the task of verifying an individual's identity from their voice samples using machine learning algorithms, without any human intervention. Since voice has been one of the most casual means for natural interactions between humans and machines, voice-based systems are easy and intuitive for human operators to use. Further, voice is inherent to individuals and can neither be lost nor stolen which makes it highly accurate and reliable. The availability of low-cost and portable microphones gives it capability of easy integration. ASV has seen significant advancements over the past few decades, giving rise to the successful introduction for various sectors, such as health care, finance and manufacturing industry etc.

Although state-of-the-art i-vector/PLDA based systems exhibit satisfactory performance with adequate speech data [2], a major challenge in ASV is to improve performance with limited voice segments. On the one hand, to achieve fair performance, ASV systems need to be presented with sufficient long utterance (two or three minutes) for enrollment and test i-vectors extraction [3, 4]. Indeed, it is often difficult to acquire such long speech for practice ASV systems because of background noise, voice overlaps or faulty recording devices. Also, there are difficulties related to speaker himself. In fact, unwilling speakers, the state of health, the character of speakers can all contribute to a reduced available amount of speech data. On the other hand, the systems require a large amount of development data to estimate reliable hyper-parameters. Particularly, the success of PLDA modeling depends on the availability of a large set of labeled in-domain data. In most real-life application, collection of such amount of development data from target domain is infeasible. Hence, it is crucial to maintain ASV performance when it is constrained on limited voice data.

Over the years, considerable research effort has been made to overcome such challenges. In [5], the duration variability is mitigated by propagating the posterior covariance of i-vectors to PLDA. However, scoring is computationally expensive in this method. The work in [6] proposed full posterior distribution PLDA to address short duration issue. The work in [7] attempted to improve short utterance system performance by adaptation for i-vector estimation. Also, many techniques are proposed to deal with inadequate target domain data in PLDA modeling. The work in [8] proposes Bayesian adaptation of PLDA models. In [9], unsupervised clustering of i-vectors for adapting covariance matrices of PLDA models is proposed. The work in [10] proposes inter-dataset variability compensation (IDVC) to find a feature space that is more domain independent. In this paper, we propose a new method by incorporating historical test information for short utterance i-vector extraction. In addition, we modify the conventional LDA projection to compensate the domain mismatch before PLDA modeling. In contrast to the existing works address limited utterance length and limited in-domain development data in separate view, we integrate proposed methods in one system and validate it in a real-life power grid dispatching room scenario.

The rest of the paper is organized as follows. Section 2 describes i-vector/PLDA framework as our baseline ASV system. The proposed method for i-vector extraction and modification for LDA are detailed in Sect. 3. Section 4 presents the experimental setups. Section 5 discusses system implementation and evaluation. Section 6 concludes the paper and outlines future studies.

2 Baseline ASV System Description

2.1 I-Vector Extraction

As mentioned earlier, i-vector based system has become de facto choice for speaker verification and related tasks. I-vector is essentially a low-dimensional representation of the Gaussian mixture model (GMM) super-vector found through a factor analysis process. Specifically, the speaker and channel dependent GMM super-vector M can be generated by

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

where m is the speaker and channel independent super-vector, which is concatenated means of universal background model (UBM), T is a low-rank total variability (TV) matrix, and w is a random latent variable with standard normal distribution. In i-vector approach, the universal background model (UBM) and total variability (TV) matrix are trained with large amount speech data gathered from different speakers. The i-vector x is given by the maximum a posteriori (MAP) point estimate of the hidden variable w which is equal to the mean of the posterior distribution of w conditioned on input utterance:

$$\mathrm{x} = \left(I + T^T\Sigma^{-1}NT\right)^{-1}T^T\Sigma^{-1}N(E - m) \tag{2}$$

where Σ is a diagonal matrix, in which the diagonal blocks are corresponding covariance matrices of Gaussian components of the UBM, N and E are zero and first order Baum-Welch (BW) statistics matrices, respectively. Given an utterance $\mathrm{X} = \{x_1, x_2, \ldots, x_F\}$, the zero and first order BW statistics are computed using UBM as

$$N_i = \sum\nolimits_{j=1}^{F} Pr\left(i|x_j\right) \tag{3}$$

$$E_i(X) = \frac{1}{N_i}\sum\nolimits_{j=1}^{F} Pr\left(i|x_j\right)x_j \tag{4}$$

where $Pr\left(i|x_j\right)$ is posterior probability of generating x_j by corresponding Gaussian component density:

$$Pr(i|x_j) = \frac{\omega_i p_i(x_j)}{\Sigma_{k=1}^{C} \omega_k p_k(x_j)} \tag{5}$$

2.2 Linear Discriminant Analysis (LDA)

After the i-vector extraction, linear discriminant analysis (LDA) is used to compensate within-class variations and reduce the dimensionality prior to probabilistic linear discriminant analysis (PLDA) modeling. In LDA method, we simultaneously maximize the between-class variability and minimize the within-class variability by maximizing the following objective function:

$$\mathrm{J}(v) = \frac{v^T \Sigma_b v}{v^T \Sigma_w v} \tag{6}$$

where v is eigenvector, Σ_b and Σ_w are between-class scatter matrix and within-class scatter matrix, respectively, which are determined by

$$\Sigma_b = \sum\nolimits_{s=1}^{S} n_s(\bar{\mathrm{x}}_s - \bar{\mathrm{x}})(\bar{x}_s - \bar{x})^T \tag{7}$$

$$\Sigma_w = \sum\nolimits_{s=1}^{S} \sum\nolimits_{i=1}^{n_s} \left(\mathrm{x}_i^s - \bar{\mathrm{x}}_s\right)\left(\mathrm{x}_i^s - \bar{\mathrm{x}}_s\right)^T \tag{8}$$

where S is the number of all speakers, n_s is the number of utterances from speaker s, $\bar{\mathrm{X}}_s$ is the average of the i-vectors from speaker s, and $\bar{x}$ is the average of all i-vectors, defined as follows

$$\bar{\mathrm{x}}_s = \frac{1}{n_s} \sum\nolimits_{i=1}^{n_s} x_i^s \tag{9}$$

$$\bar{x} = \frac{1}{N} \sum\nolimits_{s=1}^{S} \sum\nolimits_{i=1}^{n_s} x_i^s \tag{10}$$

where N is the total number of utterances.

The LDA projection matrix is found by solving the following eigenvalue problem:

$$\Sigma_b v = \Lambda \Sigma_w v \tag{11}$$

where Λ is eigenvalue matrix. The projection matrix A is formalized by selecting first k eigenvectors corresponding to the k largest eigenvalues:

$$\mathrm{A} = [v_1, v_2 \ldots v_k] \tag{12}$$

Finally, the LDA compensated i-vectors are calculated as

$$x_{LDA} = A^T x \tag{13}$$

2.3 Probabilistic Linear Discriminant Analysis (PLDA)

Apart from compensating the within-class variations in i-vector space by subspace transformation, probabilistic linear discriminant analysis (PLDA) is widely used to reduce the redundant information such as channels from i-vectors. Here, the generative model for length-normalized i-vectors of s speaker with n_s sessions can be expressed as

$$\mathrm{x_{i,j}} = \mu + \mathrm{V} z_i + \varepsilon_{i,j} \tag{14}$$

where μ is the mean of i-vectors, V defines the eigen-voice subspace, z_i is the speaker factor, and $\varepsilon_{i,j}$ is the residual term.

The verification scores of PLDA system is given as batch likelihood ratio. For projected enrollment and test i-vectors, z_{target} and z_{test}, the batch likelihood ratio is computed as

$$\Lambda\left(z_{target}, z_{test}\right) = \log \frac{p\left(z_{target}, z_{test} | H_1\right)}{p\left(z_{target} | H_0\right) p\left(z_{test} | H_1\right)} \tag{15}$$

where H_1 denotes the hypothesis that i-vectors belong to the same speaker and H_0 denotes the hypothesis that they are from different speakers. Figure 1 shows the process of calculating scores from the enrollment and test utterance in our i-vector/PLDA ASV system.

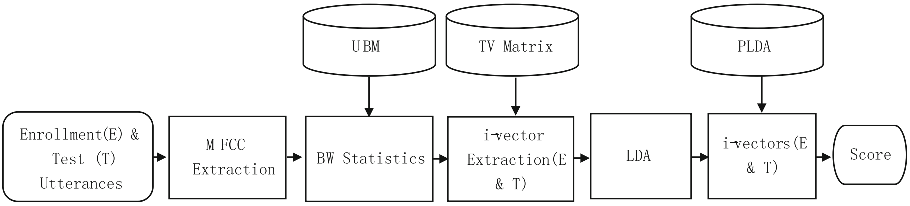

Fig. 1. Block diagram of i-vector/PLDA ASV system

3 Proposed System Modification

3.1 Analysis of I-Vector Estimation for Short Utterance

In i-vector systems, the test utterance and enrolment utterance(s) are represented by test and enrolment i-vectors extracted with pre-trained UBM and TV matrix. Then ASV is addressed by comparing the test i-vector with enrolment i-vector(s) signed by the individual to generate an accepted or rejected decision. Though the requirement of speech duration can somehow be met in enrolment stage, it may not be possible to maintain the same during the verification stage. This seriously limits the implementation of ASV system in real-world applications.

To better understand the effects of test duration variability on system performance, we present a detailed analysis of i-vector extraction pipeline. With short utterance, there is an increased uncertainty of BW statistics estimation due to lack of enough data to compute statistics parameters, which leads to an uncertain i-vector estimation. For i-vector systems, BW statistics totally represent the feature extracted from a test segment. [7, 11] Particularly, the zero-order BW statistics defines the covariance matrix of the posterior distribution given the utterance as

$$w_{\Sigma} = \left(I + T^{T}\Sigma^{-1}NT\right)^{-1} \tag{16}$$

where w_{Σ} is the covariance of the estimated i-vector, T is TV matrix, Σ is the UBM covariance, N is a diagonal matrix, where the diagonal blocks are the zero-order BW statistics of corresponding Gaussian components in UBM. Since the UBM and TV matrix are pre-trained with large quantity of data from different speakers, the higher variability introduced in BW statistics account for the uncertainty in i-vector estimation for short test segment.

3.2 Incorporating Historical Test Information in I-Vector Extraction

In order to improve the i-vector estimation, we propose a new method for adding historical test information in BW statistics computation. Rather than only use current test utterance to compute the BW statistics, we also exploit the weighted historical test utterance statistics to provide additional information. We define the weight γ_i as the estimated probability of current test utterance and historical test utterance i belonging to the same speaker. Then the BW statistics used to extract the current test i-vector is given by

$$N = N_c + \Sigma\gamma_i N_i \tag{17}$$

$$E = E_c + \Sigma\gamma_i E_i \tag{18}$$

where N_c and E_c are BW statistics computed from current test utterance, N_i and E_i are BW statistics computed from historical test utterance, and γ_i is corresponding weight assigned to historical test.

To compute the weight γ_i for historical test utterance, we use a world MAP estimator which was proposed in [12] and successfully applied to unsupervised GMM adaptation thereafter in [13, 14]. We first train a two-class Bayesian classifier based on two score models - target and non-target scores - learned from a development set. [14] Each score distribution is modelled by a 12 components GMM. Given the priori target and non-target score distributions, we can compute the posteriori probability of having a target. Specifically, for every encountered test utterance, ASV system output a raw score. Given current test raw score, s_0, the posteriori probability of this test belonging to the target speaker is defined as

$$P(tar|s_0) = \frac{P(s_0|tar)P_{tar}}{P(s_0|tar)P_{tar} + P(s_0|non)P_{non}} \tag{19}$$

where $P(s_0|tar)$ and $P(s_0|non)$ are the probabilities of the score given the target and non-target score distributions, P_{tar} and P_{non} are the prior probabilities of target and non-target test respectively. Then for historical test utterance i with raw score, s_i, we can compute weight γ_i as follows:

$$\gamma_i = P(tar|s_o)P(tar|s_i) + [1 - P(tar|s_0)][1 - P(tar|s_i)] \tag{20}$$

Note that all scores used are normalized. In proposed method, we do not require access to the historical test utterances as well as i-vectors. To utilize historical test information, only raw score and corresponding BW statistics are needed, which do not put a heavy burden on real-life applications. Figure 2 shows the flow diagram of the proposed method.

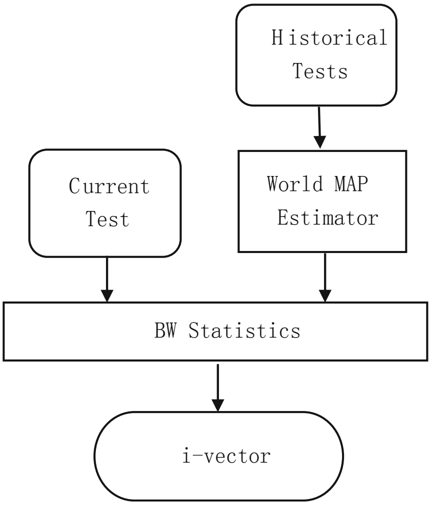

Fig. 2. Flow diagram of the proposed i-vector extraction method

3.3 Modified LDA for Domain Mismatch Compensation

One of the keys to the success of i-vector/PLDA framework is the use of a large quantity of previously collected speech data to characterize and model speaker and channel variability. However, it is unrealistic to assume such a large set of development data for every domain of interest. This is especially true for PLDA modeling, which needs labeled speech data, whereas the training of UBM and TV matrix only need unlabeled data. Studies have found that when PLDA is trained using out-domain data, the ASV system performance degrades rapidly due to the mismatch between development and evaluation data [15].

Conventional LDA projection falls to compensate this domain variability because it captures the domain variability in between-class scatter matrix. Instead of minimizing the domain mismatch in projected i-vectors, LDA maximizes domain variability when training the projection matrix. In order to address such problem, we modify the LDA training to separate domain variability from scatter matrix estimation. For simplicity, we assume the speakers do not overlap across different domains. In our method, the new between-class scatter matrix and within-class scatter matrix are defined as

$$\Sigma'_b = \sum\nolimits_{s=1}^{S_{OUT}} n_s(\bar{\mathrm{x}}_s - \bar{\mathrm{x}}_{out})(\bar{\mathrm{x}}_s - \bar{\mathrm{x}}_{out})^T + \sum\nolimits_{s=1}^{s_{in}} n_s(\bar{\mathrm{x}}_s - \bar{\mathrm{x}}_{out})(\bar{\mathrm{x}}_s - \bar{\mathrm{x}}_{out})^T \tag{21}$$

$$\Sigma'_w = \sum\nolimits_{s=1}^{S_{OUT}} \sum\nolimits_{i=1}^{n_s} \left(\mathrm{x}_i^s - \bar{\mathrm{x}}_s\right)\left(\mathrm{x}_i^s - \bar{\mathrm{x}}_s\right)^T + \sum\nolimits_{s=1}^{s_{in}} \sum\nolimits_{i=1}^{n_s} \left(\mathrm{x}_i^s - \bar{\mathrm{x}}_s\right)\left(\mathrm{x}_i^s - \bar{\mathrm{x}}_s\right)^T \tag{22}$$

where S_{out} and S_{in} are the number of out-domain and in-domain speakers, $\bar{\mathrm{x}}_{out}$ and $\bar{\mathrm{x}}_{in}$ are the average of the out-domain and in-domain i-vectors, respectively. Also, we define inter-domain variability matrix as

$$\Sigma_d = S_{out}(\bar{\mathrm{x}}_{out} - \bar{x})(\bar{\mathrm{x}}_{out} - \bar{x})^T + S_{in}(\bar{\mathrm{x}}_{in} - \bar{x})(\bar{\mathrm{x}}_{in} - \bar{x})^T \tag{23}$$

Finally, the modified LDA projection matrix can be calculated by maximizing the following objective function,

$$\mathrm{J}(v) = \frac{v^T \sum\nolimits'_b v}{v^T \sum\nolimits_{wd} v} \tag{24}$$

where v is eigenvector, and $\Sigma_{wd} = \Sigma'_w \Sigma_d^T$. By maximizing above objective function, we can simultaneously maximize the between-class variability and minimizing both within-class variability and domain variability.

4 Experimental Setups

4.1 Speech Data and Acoustic Features

Audio data are collected by an integrated microphone from power grid dispatching hall and dispatcher training simulator (DTS) room. All speakers are male. The raw data are automatically saved in a memory card every 3 min. The two locations have different room sizes, background noises, telephone channels, and so on. Figure 3 shows different environmental setting of audio data collection. From raw audio data, 19 dimensional Mel-frequency cepstral coefficients (MFCCs) together with energy coefficient are extracted and appended with delta and delta-delta features to form a 60-dimensional vector. The vector is extracted every 10 ms, using a Hamming window of 20 ms. And silence frames are detected and discarded by an energy-based voice activity detector (VAD).

(a)

(b)

Fig. 3. Different locations of audio data collection. (a) Power grid dispatching hall. (b) DTS room

Unless stated otherwise, we partition data gathered from DTS room into two subsets. We use one subset as development data and the other as evaluation data. In order to carry out experiments for short utterance conditions, original speech utterances are split into 2 s, 5 s, 10 s (only contain active frames) duration as short test segments. We randomly select initial frame and create 500 truncated segments for each duration. To test the effectiveness of modified LDA in ASV tasks with limited target domain data, we frame the domain mismatch compensation problem as reducing the mismatch between the data collected from different locations. We regard speech utterances collected from power grid dispatching hall as in-domain data, and utterances collected from dispatcher training simulator (DTS) room are considered as out-domain data. In this case, the speech files from DTS room are used as development data and speech files from power grid dispatching hall are used as evaluation data.

4.2 I-Vector Extraction and PLDA Modeling

To extract i-vector, we train a UBM with 512 Gaussian components on development data and use UBM to estimate the BW statistics. The TV subspace has a dimension of 400 and is trained on same development data. For LDA and modified LDA training, the reduced dimension is kept at 200. Length normalization is applied to LDA projected i-vectors to convert their behavior into Gaussian. Then a PLDA model with 150 latent variables is trained. We train the World MAP estimator on development data. The prior probability used are 0.1 for target and 0.9 for non-target.

4.3 Evaluation Criteria

There are two kind of mistakes in ASV system: a false rejection happens when a genuine speaker is incorrectly rejected and a false alarm when an imposter is accepted. In our experiment, the system performance is evaluated using equal error rate (EER) in which the false rejection rate and false alarm rate are equal. Also, we report experimental results in terms of minimum detection cost function (minDCF).

5 Results and Discussions

5.1 Baseline ASV System Performance

In the first series of experiments, we compare the performance of baseline ASV system in different test durations. The experiments are conducted on speech files collected from DTS room. We use 3 min raw speech for enrollment and three types of truncated segments (contain 2 s, 5 s 10 s active frames respectively) for test i-vector extraction. The results are presented in Fig. 4.

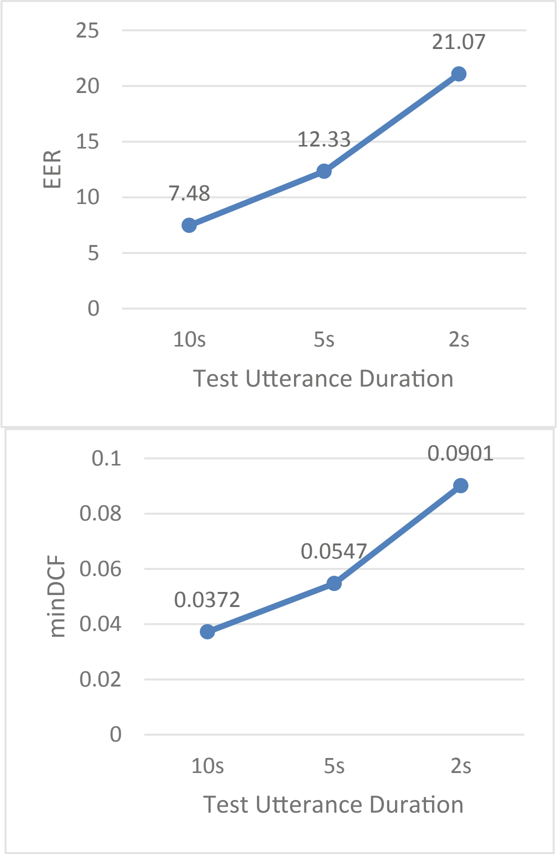

Fig. 4. Baseline ASV system performance for different test duration conditions

It can be observed that system performance in terms of both EER and minDCF degrades monotonically with the decrease in speech duration. When ASV system is presented with 2 s short utterance, the EER and minDCF increase 182% and 142% respectively compared to 10 s test utterance. This illustrates the need for proposed i-vector extraction method.

Next, we use speech files from power grid dispatching hall as evaluation data. Similarly, 3 min raw speech is used for enrollment and 2 s, 5 s, 10 s truncated speech segments are used for testing. This series of experiments aims to show the effect of in-domain development data on the performance of baseline system. The results are presented in Fig. 5.

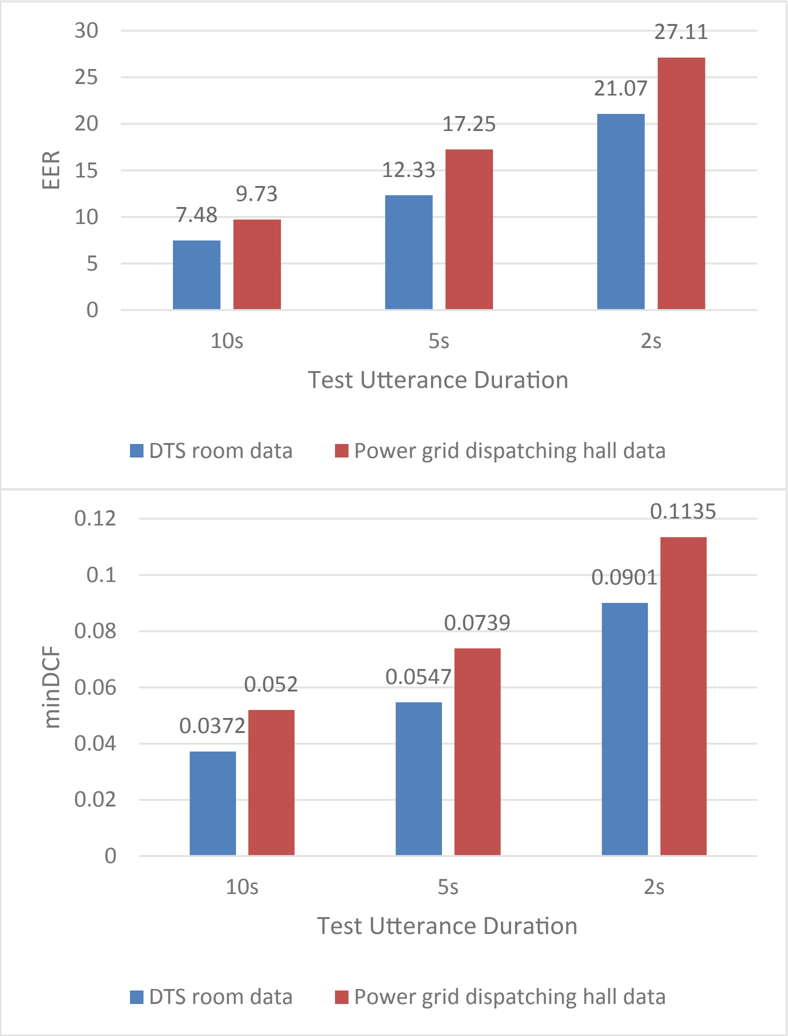

Fig. 5. Performance comparison using in-domain and out-domain development data

As shown in Fig. 5, there is a gap in performance on power grid dispatching hall enroll/test set when hyper-parameters are trained with development data gathered from DTS room. In Sect. 5.3, we employ modified LDA to reduce this performance degradation.

5.2 Proposed Method for I-Vector Extraction

In this section, we conduct experiments to test the effectiveness of incorporating historical test information in short utterance i-vector extraction. We use speech files collected from DTS room as both development and evaluation data. The results are presented in Table 1.

Table 1. Performance comparison of baseline system and system incorporating historical test information (proposed-1) in i-vector extraction

	EER			minDCF		
	10 s	5 s	2 s	10 s	5 s	2 s
Baseline (matched)	7.48	12.33	21.07	0.0372	0.0547	0.0901
Proposed-1	7.09	11.45	19.08	0.0359	0.0518	0.0827
Relative improvement	5.2%	7.1%	9.4%	3.4%	5.3%	8.2%

Experimental results reported in Table 1 show when enough historical information is inserted, the proposed method could achieve noticeable improvement in terms of EER and minDCF compared with the baseline i-vector system in different short duration conditions. We observe that the relative improvement increases with the decrease in test utterance duration. This suggests that incorporating historical information is useful for short utterance.

To analyze the behavior of our method more precisely, we investigate the system performance in terms of EER for each newly added test utterance. We conduct the experiment on 10 random draws from the entire truncated speech segments pool and

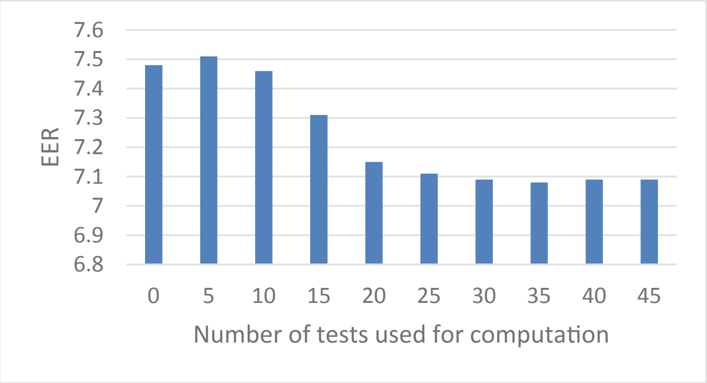

Fig. 6. Average EER of the 10 s test utterance condition

evaluate the performance individually. The results are averaged over 10 random draws for statistical significance. We notice that a minimum amount of data should be presented for proposed system to obtain stable gain. The average EER of 10 s test utterance condition are presented in Fig. 6. In 2 s and 5 s utterance conditions, the patterns are similar.

5.3 Modified LDA

As shown in Sect. 5.1, when ASV system is developed using data which is outside the target domain, it significantly affects the performance due to the mismatch between development and evaluation data. To investigate this situation, we use speech files collected from DTS room as development data and speech files collected from power grid dispatching hall as evaluation data. We use modified LDA projection to replace the conventional LDA in baseline system. System performance in terms of EER and minDCF are presented in the Table 2.

Table 2. Performance comparison of baseline system, system with modified LDA (proposed-2)

	EER			minDCF		
	10 s	5 s	2 s	10 s	5 s	2 s
Baseline (matched)	7.48	12.33	21.07	0.0372	0.0547	0.0901
Baseline (mismatched)	9.73	17.25	27.11	0.052	0.0739	0.1135
Proposed-2	7.74	14.15	22.67	0.0398	0.0588	0.0927

From Table 2, a relative gain of at least 16.4% in EER and 18.3% in minDCF is observed after applying modified LDA. In terms of bridging the performance gap between a matched baseline (DTS room data for both development and evaluation) and a mismatched baseline (DTS room data for development, power grid dispatching hall data for evaluation) system, we are able to recover at least 63% of the performance gap for different duration conditions. It demonstrates that modified LDA is quite successful in reducing the volume of in-domain development data.

Finally, we conduct experiment on system integrating the proposed i-vector extraction method and modified LDA. We develop system on speech data collected from DTS room and evaluate performance on data collected from power grid dispatching hall. From Table 3, it can be observed that further improvement is achieved with combined approach. Compared to baseline, it shows at least 20% improvement for different test segment durations.

Table 3. Performance of system using combined approach (proposed-3)

	EER			minDCF		
	10 s	5 s	2 s	10 s	5 s	2 s
Baseline (mismatched)	9.73	17.25	27.11	0.052	0.0739	0.1135
Proposed-3	7.3	13.19	21.63	0.0367	0.0546	0.0903

6 Conclusions and Future Work

The performance of i-vector/PLDA ASV systems depends on a large quantity of in-domain development data for PLDA training. During the evaluation, it is also critical that the speech duration is long enough to reduce the uncertainty in i-vector estimation. In many practical applications, the speaker verification performance is affected due to the difficulty in collecting significant amount of speech data. In this study, we propose modification for i-vector ASV system to address the issue of performance degradation with limited voice data. With the aid of historical test information, we observe a relative improvement of 9.4% in EER for 2 s test duration condition. When system is trained on mismatched development dataset, we are able to recover at least 63% of performance gap using modified LDA projection. The best performance is achieved with combined method, where we obtain relative improvement in the range of 20–29% over baseline system.

Despite the promising results, there are still some problems to study in the future. For example, currently world MAP estimator assumes the prior probabilities when the corresponding scores are not encountered in the score GMM training data. While it is anticipated that this situation is rare, we intend to investigate its effect on system performance. In addition, speakers can overlap in different domains and the data in one domain can be multi-modal. Such multi-modality can lead to misrepresentation of the speaker and non-speaker information [16]. We intend to extend our modified LDA method to compensate for speaker population difference among different portions of training data. Also, we intend to investigate the relationship between system performance and different sizes of in-domain data used for LDA training. In our future work, we intend to explore applying proposed methods onto deep neural networks (DNN) based systems. Using DNN instead of GMM to derive speaker specific information is a very promising direction to look at.

References

1. Yang, F., Wu, C., Wang, F., et al.: Review of studies on human reliability researches during 1998 to 2008. Sci. Technol. Rev. **27**(8), 87–94 (2009)
2. Dehak, N., Kenny, P., Dehak, R., et al.: Front-end factor analysis for speaker verification. IEEE/ACM Trans. Audio Speech Lang. Process. **19**(4), 788–798 (2011)
3. Cai, W., Li, M., Li, L., et al.: Duration dependent covariance regularization in PLDA modeling for speaker verification. In: Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 1027–1031 (2015)
4. Kanagasundaram, A., Dean, D., Sridharan, S., et al.: A study on the effects of using short utterance length development data in the design of GPLDA speaker verification systems. Int. J. Speech Technol. **20**(2), 247–259 (2017)
5. Kenny, P., Stafylakis, T., Ouellet, P., et al.: PLDA for speaker verification with utterances of arbitrary duration. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7649–7653 (2013)
6. Cumani, S., Plchot, O., Laface, P.: On the use of i-vector posterior distributions in probabilistic linear discriminant analysis. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(4), 846–857 (2014)

7. Poddar, A., Sahidullah, M., Saha, G.: Improved i-vector extraction technique for speaker verification with short utterances. Int. J. Speech Technol. **21**(3), 473–488 (2018)
8. Villalba, J., Lleida, E.: Unsupervised adaptation of PLDA by using variational bayes methods. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 744–748 (2014)
9. Shum, S., Reynolds, D.A., Garcia-Romero, D., et al.: Unsupervised clustering approaches for domain adaptation in speaker recognition systems. In: Odyssey: The Speaker and Language Recognition Workshop, pp. 266–272 (2014)
10. Aronowitz, H.: Compensating inter-dataset variability in PLDA hyper-parameters for robust speaker recognition. In: Odyssey: The Speaker and Language Recognition Workshop, pp. 280–286 (2014)
11. Kenny, P., Ouellet, P., Dehak, N., et al.: A study of interspeaker variability in speaker verification. IEEE/ACM Trans. Audio Speech Lang. Process. **16**(5), 980–988 (2008)
12. Fredouille, C., Bonastre, J.F., Merlin, T.: Bayesian approach-based decision in speaker verification. In: Odyssey: The Speaker and Language Recognition Workshop, pp. 77–81 (2001)
13. Preti, A., Bonastre, J.F., Matrouf, D.: Confidence measure based unsupervised target model adaptation for speaker verification. In: Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 754–757 (2007)
14. Mclaren, M., Matrouf, D., Vogt, R., et al.: Applying SVMs and weight-based factor analysis to unsupervised adaptation for speaker verification. Comput. Speech Lang. **25**(2), 327–340 (2011)
15. Rahman, M.H., Kanagasundaram, A., Himawan, I., et al.: Improving PLDA speaker verification performance using domain mismatch compensation techniques. Comput. Speech Lang. **47**, 240–258 (2017)
16. Glembek, O., Ma, J., Matejka, P., et al.: Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4060–4064 (2014)