INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Transfer Learning for Speaker Verification on Short Utterances

Qingyang Hong[1], *Lin Li*[1*], *Lihong Wan*[1], *Jun Zhang*[1], *Feng Tong*[2]

[1]School of Information Science and Technology, Xiamen University, China
[2]Key Lab of Underwater Acoustic Communication and Marine Information Technology of MOE, Xiamen University, China

[*]Corresponding to:lilin@xmu.edu.cn

Abstract

Short utterance lacks enough discriminative information and its duration variation will propagate uncertainty into a probability linear discriminant analysis (PLDA) classifier. For speaker verification on short utterances, it can be considered as a domain with limited amount of long utterances. Therefore, transfer learning of PLDA can be adopted to learn discriminative information from other domain with a large amount of long utterances. In this paper, we explore the effectiveness of transfer learning based PLDA (TL-PLDA) on the NIST SRE and Switchboard (SWB) corpus. Experimental results showed that it could produce the largest gain of performance compared with the traditional PLDA, especially for short utterances with the duration of 5s and 10s.

Index Terms: speaker verification, transfer learning, PLDA, short utterance

1. Introduction

For commercial applications of biometric authentication, speaker verification on short utterances is more preferred for the users. Currently, the state-of-the-art speaker verification system is based on i-vector [1] and probability linear discriminant analysis (PLDA) [2][3]. An i-vector is the low dimension representation of a Gaussian mixture model (GMM) mean supervector from a given speech utterance, which is obtained based on zero and first order Baum-Welch statistics based on universal background model (UBM) [1][4]. Since the statistics are accumulated over-time, the current speaker recognition feature spaces reach high relative entropy level with long duration more than 20 seconds [5]. For those short utterances with sparse statistics, the performance of speaker verification will deteriorate greatly due to limited discriminative information.

Many studies have been conducted to investigate the influence of duration [6][7][8][9][10][11][12][13][14], which include the improved methods of score calibration and duration modeling. In [6][7], quality measure function (QMF) of duration is adopted to counteract the duration variability problem and improves the calibration performance of speaker recognition system. For the methods of duration modeling, most studies focus on the PLDA, which has gained popularity as an elegant classification tool to find target classes in recent NIST challenges. However, duration variation might propagate uncertainty into a PLDA classifier, especially for those short utterances. In [9][10][14], the duration variability of i-vector were compensated in the PLDA model and performance improvement had been achieved. In [15], we also proposed an effective modified-prior PLDA framework to deal with the duration variation. As shorter utterances tend to have large covariance, the probability distribution function of i-vector can be modified with duration scaled covariance matrix during the PLDA training process. Then the formulation of the likelihood for standard Gaussian PLDA model is revised according to the duration-dependent posterior distribution of the i-vector. Overall, these works mostly deal with the problem of duration variation, but not consider and compensate the limited discriminate information of short utterances.

In [16], we have proposed a novel transfer learning method for target domain with limited amount of speakers and sessions, in which Kullback Leibler (KL) regularization factor is added into the objective function of PLDA to measure the similarity between the source domain and the target domain. Experimental results showed that our proposed transfer learning based PLDA (TL-PLDA) could produce the largest gain of performance compared with the traditional PLDA and the PLDA interpolation approach.

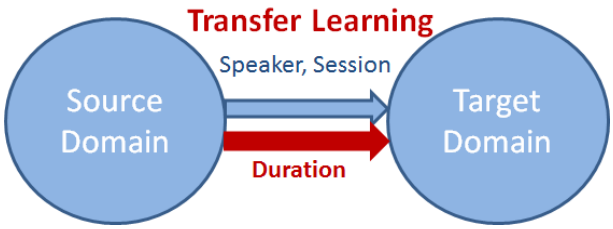


Figure 1: *Transfer learning for duration compensation.*

Motivated by our former successful deployment of transfer learning for speaker and session variations, this paper further investigates TL-PLDA for duration compensation (Figure 1). For speaker verification on short utterances, it can be viewed as a domain with limited amount of long utterances with enough linguistic content. The differences in the linguistic content of short utterances can be learned using the development data with full-length i-vectors [14]. Therefore, transfer learning can be also adopted to learn valuable information from other domain with long utterances. To explore the effectiveness of TL-PLDA for short utterances, several evaluation tasks will be designed on varying duration conditions with full length and randomly truncated test utterances with the duration of 5s, 10s, 20s and 40s respectively.

2. Methods

In the state-of-the-art i-vector speaker verification system, an i-vector $\boldsymbol{x}$ is a fixed-length vector, which is decomposed by a total variability matrix $\boldsymbol{T}$ into a single low dimensional subspace.

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{x} \tag{1}$$

http://dx.doi.org/10.21437/Interspeech.2016-432

where $\boldsymbol{m}$ is a UBM-based supervector, $\boldsymbol{x}$ is a hidden variable which can be defined as the mean of posterior distribution of the Baum-Welch statistics for an utterance.

Given a UBM with C components and each component is a Gaussian mixture characterized by $\lambda_c = \left\{w_c, \mu_c, \sigma_c^2\right\}, c = 1, 2, \ldots, C$, with the weight w_c, the mean μ_c and variance σ_c. For an utterance $\boldsymbol{y}$ with L feature vectors $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_l, \ldots, \boldsymbol{y}_L\}$, the zero and first order Baum-Welch statistics, represented as N_c and F_c respectively, for each component c are obtained as follows.

$$N_c = \sum_{l=1}^{L} P(c \mid \boldsymbol{y}_l, \lambda_c) \tag{2}$$

$$F_c = \frac{1}{N_c} \sum_{l=1}^{L} P(c \mid \boldsymbol{y}_l, \lambda_c)(\boldsymbol{y}_l - \mu_c) \tag{3}$$

Based on the supervector $\boldsymbol{F}$ which is concatenated with each component F_c, we can get the *i*-vector $\boldsymbol{x}$. Since long duration of L is important to get the sufficient statistics, short utterance will not provide reliable discriminative information for *i*-vector extraction. To measure the discriminative information and similarity of *i*-vector, we can use Cosine distance score [1].

2.1. Standard Gaussian PLDA

Generally, the Gaussian PLDA (G-PLDA) is adopted after *i*-vector length normalization [3]. In Gaussian PLDA, the *i*-vector $\boldsymbol{x}_{ij}$ for the *j*th utterance of speaker i is decomposed as follows.

$$\boldsymbol{x}_{ij} = \boldsymbol{\mu} + \boldsymbol{\Phi}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_{ij} \tag{4}$$

where $\boldsymbol{\mu}$ represents the mean of development data, $\boldsymbol{\beta}_i$ is an identity variable of speaker i having a standard normal prior $N(0, \boldsymbol{I})$, matrix $\boldsymbol{\Phi}$ constrains the dimension of the speaker subspace, and the residual $\boldsymbol{\varepsilon}_{ij}$ contains the session factors following a normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}$. Generally, the standard Gaussian PLDA only models the speaker and session variations.

2.2. Transfer learning for short utterances

For the i-vector/PLDA system, short utterances not only lack enough discriminative information, but also might cause duration variations under mismatched condition. Our objective is to learn the suitable PLDA parameters $(\boldsymbol{\Phi}_t, \boldsymbol{\Sigma}_t)$ for the target domain with short utterances.

Given the PLDA parameters $(\boldsymbol{\Phi}_s, \boldsymbol{\Sigma}_s)$ of the source domain and the short-time development data of the target domain, the KL based optimization objective of transfer learning is defined as follows.

$$\begin{aligned} \min \sum_{i=1}^{N} (&-P(\boldsymbol{F}_t \mid \boldsymbol{\beta}_i) P(\boldsymbol{\beta}_i) \\ &+ \lambda KL(P(\boldsymbol{F}_s \mid \boldsymbol{\beta}_i) \parallel P(\boldsymbol{F}_t \mid \boldsymbol{\beta}_i))) \end{aligned} \tag{5}$$

where $P(\boldsymbol{F}_t \mid \boldsymbol{\beta_i})$ and $P(\boldsymbol{F}_s \mid \boldsymbol{\beta_i})$ is the posterior distribution of $\boldsymbol{F}_i$ based on the hidden variable $\boldsymbol{\beta}_i$ in the target domain and the source domain respectively, and $\boldsymbol{F}_i$ denotes the mean i-vector value of the first order statistic $(\boldsymbol{x}_{ij} - \boldsymbol{\mu})$ of speaker i of N speakers in the target domain (represented as $\boldsymbol{F}_t$) and the source domain (represented as $\boldsymbol{F}_s$) respectively. λ is an adjusting weight. When $\lambda = 0$, this objective function will regress to the original standard PLDA, i.e. $\min \sum_{i=1}^{N} (-P(\boldsymbol{F}_t \mid \boldsymbol{\beta}_i) P(\boldsymbol{\beta}_i))$. With the value of λ increasing, the optimization process will gradually lead to the distribution of the source domain. In the experiments, we'll discuss it in more detail.

For the TL-PLDA of target domain, by setting the derivative of objective function in (5) towards $\boldsymbol{\Phi}_t$ or $\boldsymbol{\Sigma}_t$ to be zero, we can get the final re-estimation formula of $\boldsymbol{\Phi}_t$ and $\boldsymbol{\Sigma}_t$ as follows[16].

$$\boldsymbol{\Phi}_t = w\boldsymbol{\Phi}_s + (1 - w)\boldsymbol{\Phi}' \tag{6}$$

$$\boldsymbol{\Sigma}_t = w\boldsymbol{\Sigma}_s + (1 - w)\boldsymbol{\Sigma}' + w\boldsymbol{\Delta} \tag{7}$$

where $w = \lambda/(1 + \lambda)$. $\boldsymbol{\Phi}'$ and $\boldsymbol{\Sigma}'$ will be updated in each step using the development data of target domain based on the re-estimation formula of standard Gaussian PLDA. $\boldsymbol{\Delta}$ is a new factor, which can be calculated as follows.

$$\boldsymbol{\Delta} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} (\boldsymbol{\Phi}_s E(\boldsymbol{\beta}_i \boldsymbol{\beta}_i^T) \boldsymbol{\Phi}_s^T - \boldsymbol{\Phi}_t E(\boldsymbol{\beta}_i \boldsymbol{\beta}_i^T) \boldsymbol{\Phi}_s^T)}{\sum_{i=1}^{N} M_i} \tag{8}$$

where M_i is the number of utterances for speaker i in the target domain. With the EM algorithm, the parameters are estimated with the termination condition when increment value of the objective function is less than the threshold 0.1 or the iteration number of EM steps exceeds 10.

3. Experiments

Experiments were conducted based on the NIST SRE10 and Switchboard (SWB) corpus. From the SWB corpus, 11453 utterances from 993 female speakers and 9813 utterances from 844 male speakers were picked out to train the gender-dependent UBM containing 1024 Gaussians. The total variability subspace of dimension 400 was estimated by the Baum-Welch statistics based on the same data. For each i-vector, the centering process was based on the mean of its own domain, but the whitening process was based on the SWB statistics [16]. The PLDA was trained with speaker subspace of dimension 120. The results presented in this paper included female and male trials.

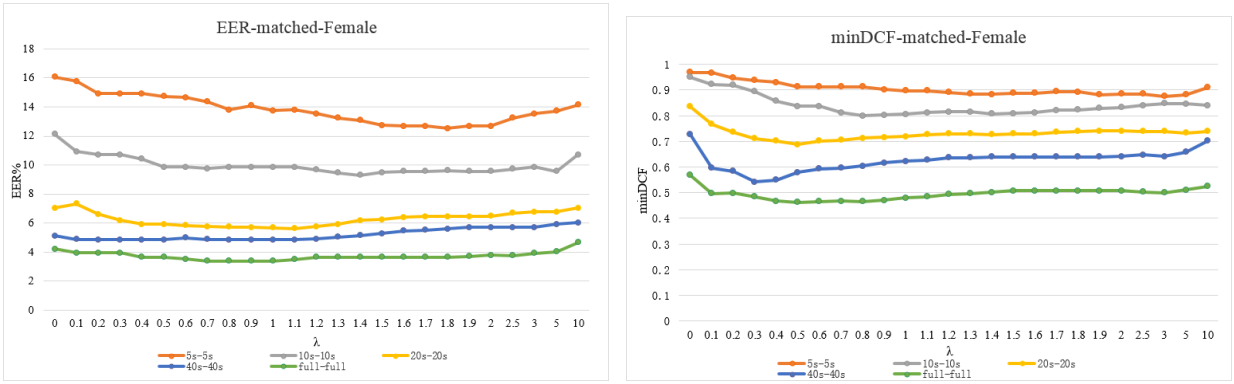
In our experiments of transfer learning, the SWB corpus was used as the data of source domain. From the NIST SRE10 corpus, we selected 1325 female speakers and 1024 male speakers to act as the gender-dependent development data of target domain. Since this work is focused on duration compensation, we use enough number of speakers to avoid the variations of speakers and sessions. For the performance evaluation, the NIST SRE10 telephone data (condition-5) was used as enrolment and test sets. Each enrolment utterance remains full length.

To analyze the effective performance of TL-PLDA for the domain with short utterances, several evaluation tasks were designed on varying duration conditions with full length and randomly truncated utterances of target domain development data (including 16904 utterances for female and 15254 utterances for male) and test data (including 16313 trials for female and 14060 trials for male in condition-5) with the duration of 5s, 10s, 20s and 40s respectively.

The experiments of short utterances included the cases of matched duration and mismatched duration. For the evaluation of matched duration, we designed five versions of duration group (duration of development data - duration of test data): 5s-5s, 10s-10s, 20s-20s, 40s-40s and full-full. For those of mismatched duration, we had four versions of duration group

Table 1: *Performance of Cosine distance score for different duration groups.*

Condition	Duration Group	Female				Male			
		Mean	Var	Min	Max	Mean	Var	Min	Max
Matched	5s-5s	**0.06**	**0.004**	-0.095	0.26	**0.08**	**0.004**	-0.099	0.33
	10s-10s	**0.07**	**0.004**	-0.085	0.28	**0.10**	**0.005**	-0.031	0.42
	20s-20s	**0.11**	**0.005**	-0.052	0.39	**0.13**	**0.005**	-0.045	0.40
	40s-40s	**0.14**	**0.007**	-0.028	0.49	**0.17**	**0.006**	-0.001	0.49
	Full-Full	**0.24**	**0.011**	0.0006	0.75	**0.26**	**0.010**	0.016	0.58
Mismatched	Full-5s	**0.10**	**0.005**	-0.075	0.37	**0.12**	**0.006**	-0.048	0.46
	Full-10s	**0.12**	**0.006**	-0.046	0.50	**0.15**	**0.0070**	-0.032	0.51
	Full-20s	**0.15**	**0.007**	-0.050	0.54	**0.18**	**0.0076**	-0.002	0.54
	Full-40s	**0.18**	**0.008**	-0.016	0.60	**0.21**	**0.0081**	-0.003	0.56



(a) EER (b) minDCF

Figure 2: *The performance of TL-PLDA for different adjusting weight under matched condition for female trials.*

(duration of development data - duration of test data): full-5s, full-10s, full-20s and full-40s. In our experiments, the equal error rate (EER) and the 2010 minimum decision cost function (minDCF) were calculated as evaluation metrics.

3.1. Cosine distance of short utterances

To investigate the influence of short utterances on the reliability of i-vector feature, we first conducted the experiment based on Cosine distance score(CDS) for female and male trials in condition-5 task, since it can be viewed as the measure of similarity and it is more computationally efficient than PLDA. Table 1 presents the results for different duration groups (duration of model utterance - duration of test utterance), in which the scores of target have been analyzed in statistical method.

It is shown that when shorter test utterances were used for i-vector extraction, it would result in a marginal drop in Cosine distance score (CDS). With the increasing of duration of test utterances, the mean of Cosine distance score increased subsequently, which indicated that the enrolment and test utterance become more similar. This demonstrated that the discriminate capability of i-vector was heavily dependent on the duration.

3.2. TL-PLDA for matched short utterances

The experiments in this part focused on duration matched condition of the development data of target domain and the test data. The data in target domain consisted of the truncated short utterances of SRE10, and we trained the corresponding PLDA of 5s, 10s, 20s, 40s and full length respectively. Based on long utterances of SWB, we conducted the transfer learning of PLDA to learn discriminative information. For the female trial, we firstly investigated the influence of adjusting weight λ in (5), as shown in Figure 2.

Figure 2(a) compares the EER results for different adjusting weight λ with the value from 0 to 10. When the value is equal to 0, the TL-PLDA regressed to the original standard PLDA of the target domain (Target PLDA). Therefore, we can directly compare the performance of TL-PLDA with the standard PLDA. With the increasing of adjusting weight λ, the EER could be reduced but had some fluctuation. Figure 2(b) further compares the corresponding minDCF results for different adjusting weight λ.

In Figure 3, the best results of TL-PLDA were compared with the target domain PLDA. It's obvious to find that with the reduction of duration, the EER values and minDCF value of the two methods increased subsequently. It can be seen that transfer learning was more effective to reduce the EER of shorter utterances. For the duration group of 10s-10s for female trials, the EER was reduced from 11.775% of Target PLDA to 9.2958% of TL-PLDA. For the duration group of 5s-5s for female trials, the proposed TL-PLDA obtained 27% reduction of EER value compared to Target PLDA. For all cases, TL-PLDA had the lowest EER results and gained the comparable performance of minDCF value, which demonstrated the effectiveness of our method.

3.3. TL-PLDA for mismatched short utterances

It is known that duration mismatch will cause uncertainty into the PLDA classifier. In this experiment, we evaluated the effectiveness of TL-PLDA under different conditions: full-5s, full-10s, full-20s and full-40s. In Table 2, the best results of TL-PLDA were listed and compared with Target PLDA and Source PLDA (optimized based on the development data of source do-

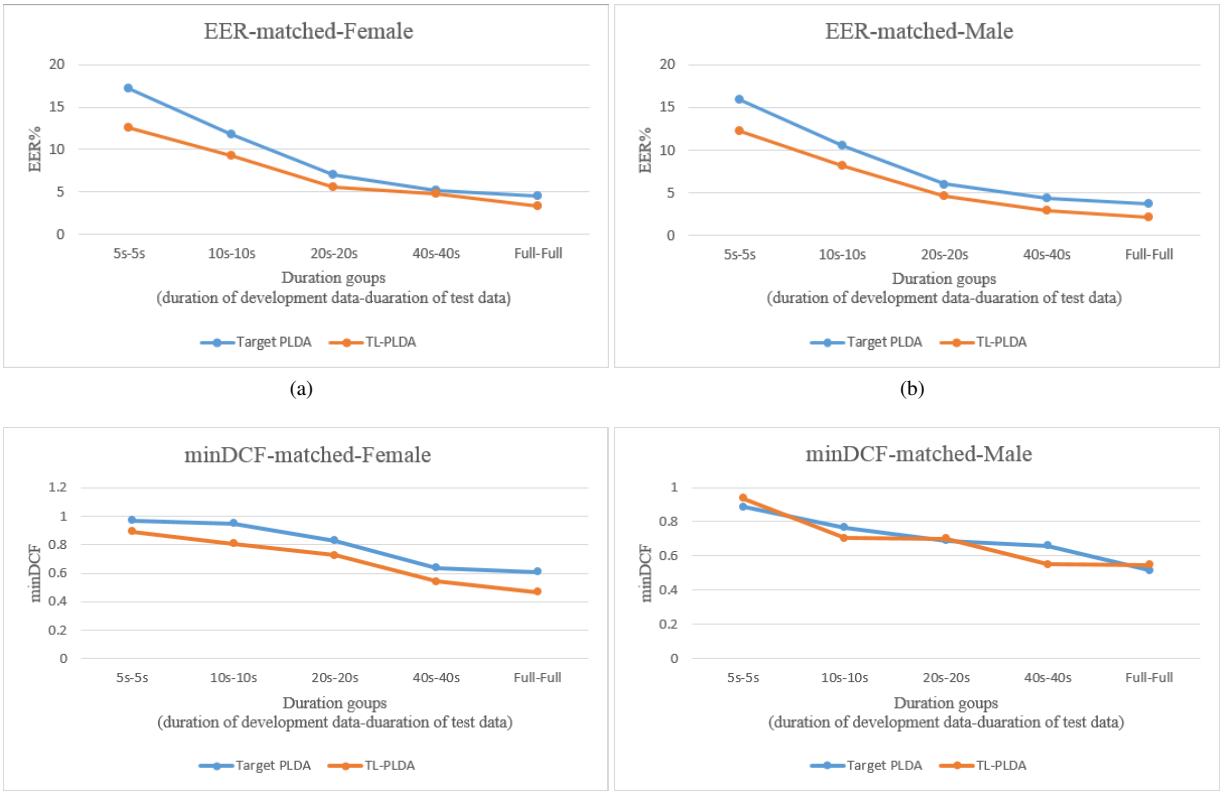


Figure 3: Performance comparison under matched condition.

Table 2: *Performance comparison under mismatched condition.*

Gender	Duration Group	Source PLDA		Target PLDA		TL-PLDA	
		EER%	minDCF	EER%	minDCF	EER%	minDCF
Female	Full-5s	15.775	0.9380	12.157	0.9211	**11.268**	**0.8513**
	Full-10s	13.235	0.9042	9.1052	0.8689	**8.169**	**0.7465**
	Full-20s	9.8592	0.8598	7.0811	0.7675	**5.7213**	**0.6795**
	Full-40s	8.547	0.8468	5.0302	0.7253	**4.8433**	**0.5647**
Male	Full-5s	13.314	0.983	10.482	0.8130	**9.915**	**0.8839**
	Full-10s	9.5207	0.8385	7.332	0.6459	**6.5156**	**0.7337**
	Full-20s	9.1413	0.9207	5.3257	0.5836	**4.2493**	**0.6925**
	Full-40s	7.2254	0.8370	4.3353	0.6532	**3.1792**	**0.5202**

main). In female trials, we can see that the EER was reduced from 12.157% of Target PLDA to 11.268% of TL-PLDA for the case of Full-5s. And the EER was reduced by 10.3% from 9.1052% of Target PLDA to 8.169% of TL-PLDA for the case of Full-10s. In male trials, the EER was reduced from 10.482% of Target PLDA to 9.915% of TL-PLDA for the case of Full-5s. And the EER was reduced by 11.1% from 7.332% of Target PLDA to 6.5156% of TL-PLDA for the case of Full-10s. Compared with Source PLDA and Target PLDA, TL-PLDA always had the lowest EER results. But as shown in some cases of male trials, the minDCF value didn't achieved a better one while adjusting the weight λ to obtain the minimum of EER value.

4. Conclusions

In this paper, we have successfully applied transfer learning method for the target domain with short utterances. For those short utterances with sparse statistics, the performance of PLDA-based speaker verification will deteriorate greatly due to limited discriminative information and duration mismatch. Based on the similarity measure of KL divergence, transfer PLDA could learn linguistic content information from other domain with long utterances and thus improve the robustness. We have conducted experiments for varying durations of target domain data based on the NIST SRE10 and Switchboard corpus. The results showed that the proposed TL-PLDA method could outperform the traditional PLDA, especially for short utterances with the duration of 5s and 10s.

5. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 61105026, 11274259.

6. References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20*, pp. 1–8, 2007.

[3] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31*, 2011.

[4] P. Kenny, "Joint factor analysis of speaker and session variability : Theory and algorithms," in *Tech. Rep., Centre de recherche informatique de Montr'eal (CRIM)*, 2005.

[5] A. Nautsch, C. Rathgeb, R. Saeidi, and C. Busch, "Entropy analysis of i-vector feature spaces in duration-sensitive speaker recognition," in *40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, Brisbane, Australia, April 19-24*, 2015.

[6] M. I. Mandasari, R. Saeidi, and D. A. V. Leeuwen, "Quality measures based calibration with duration and noise dependency for speaker recognition," *Speech Communication*, vol. 72, pp. 126–137, 2015.

[7] M. I. Mandasari, R. Saeidi, M. Mclaren, and D. A. Van Leeuwen, "Quality measure functions for calibration of speaker recognition systems in various duration conditions," *IEEE Transactions on Audio Speech & Language Processing*, vol. 21, no. 11, pp. 2425–2438, 2013.

[8] A. Nautsch, C. Rathgeb, C. Busch, H. Reininger, and K. Kasper, "Towards duration invariance of i-vector-based adaptive score normalization," in *Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014.

[9] T. Hasan, R. Saeidi, J. H. L. Hansen, and D. A. Van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2013, Vancouver, Canada, May 26-30*, 2013, pp. 7663–7667.

[10] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2013, Vancouver, Canada, May 26-30*, 2013, pp. 7649 – 7653.

[11] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances." in *9th Annual Conference of the International Speech Communication Association, Brisbane, Australia Sep 22-26*, 2008, pp. 853–856.

[12] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-vector based speaker recognition on short utterances." in *12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31*, 2011.

[13] A. K. Sarkar, D. Matrouf, P. M. Bousquet, and J. F. Bonastre, "Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification," in *13th Annual Conference of the International Speech Communication Association, Portland, USA, Sep 9-13*, 2012.

[14] A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, and D. Ramos, "Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques," *Speech Communication*, vol. 59, no. 2, pp. 69–82, 2014.

[15] Q. Hong, L. Li, M. Li, L. Huang, L. Wan, and J. Zhang, "Modified-prior plda and score calibration for duration mismatch compensation in speaker recognition system," in *16th Annual Conference of International Speech Communication Association, Germany, Dresden, Sep 6-10*, 2015.

[16] Q. Hong, J. Zhang, L. Li, L. Wan, and F. Tong, "A transfer learning method for plda-based speaker verification," in *41th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016, Shanghai, China, March 20-25*, 2016.