A TRANSFER LEARNING METHOD FOR PLDA-BASED SPEAKER VERIFICATION

Qingyang Hong[1], *Jun Zhang*[1], *Lin Li*[1], *Lihong Wan*[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
lilin@xmu.edu.cn

ABSTRACT

Currently, the state-of-the-art speaker verification system is based on i-vector and PLDA. However, PLDA requires tens of thousands of development data from many speakers. This makes it difficult to learn the PLDA parameters for a domain with scarce data. In this paper, we propose an effective transfer learning method based on Bayesian joint probability in which Kullback-Leibler (KL) divergence between the source domain and the target domain is added as a regularization factor. This hypothesis could utilize the development data of source domain to help find a better optimal solution of PLDA parameters for the target domain. Experimental results based on the NIST SRE and Switchboard corpus demonstrate that our proposed method could produce the largest gain of performance compared with the traditional PLDA and the other adaptation approach.

Index Terms— Speaker Verification, PLDA, Transfer Learning, Domain Adaptation

1. INTRODUCTION

The task of speaker verification is to verify the identity of speaker given a speech utterance. Its robustness is affected by many factors (channel, noise, language, duration, etc.), of which the most important one is channel variation. In the past ten years, text-independent speaker verification [1] has achieved great progress to solve the channel problem. Many machine learning approaches, including support vector machine (SVM) [2,3], joint factor analysis (JFA) [4,5], i-vector [6], have been proposed. Currently, the state-of-the-art speaker verification system is based on i-vector and probability linear discriminant analysis (PLDA) [7,8,9]. With the posterior estimation of the hidden variables on the Baum-Welch statistics from the Gaussian components of a universal background model (UBM) [1], each speech utterance can be represented as a low-dimension vector, i.e. i-vector. Length normalization, including centering and whitening, is subsequently conducted to the extracted i-vectors [8]. Furthermore, PLDA is generally adopted to compensate the channel difference in i-vectors.

However, PLDA requires tens of thousands of labeled development data from many speakers. For the NIST evaluation, this is not a problem since there are sufficient data provided. But it will be very difficult for practical applications. Even if we have sufficient development data to get a well-optimized PLDA for a source domain, it is not suitable to use it directly for a new target domain with different channels. Several studies have shown that when the development data and evaluation data are from different domains, the performance of speaker verification will significantly deteriorate due to domain mismatch [10-15].

To minimize the performance gap between different domains, J. Villalba and E. Lleida applied the variational Bayes for two-covariance model [10]. D. Garcia-Romero et al. proposed several adaptation approaches with similar performances, of which PLDA interpolation approach did not require keeping the i-vectors of source domain to retrain the PLDA [11,12]. A. Kanagasundaram et al. proposed an unsupervised inter-dataset variability approach to compensate the mismatch but only linear discriminant analysis (LDA) projection was applied prior to the PLDA modeling [15].

Motivated by the study of face verification [16], we propose in this paper an effective transfer learning method from the source domain to the target domain. Transfer learning has been successfully applied in many fields [17,18]. But to the best of our knowledge, there is still not a transfer learning schedule for PLDA-based speaker verification. Based on a Kullback-Leibler (KL) divergence between the distributions of source domain and target domain, we define a new optimization function to maximize the shared information of two domains. After updating steps of transfer learning based on expectation-maximization (EM) algorithm, we get the new transfer learning schedule of PLDA model based on the same target domain data. Experimental results based on the NIST SRE and Switchboard corpus show that our method could improve the verification performance greatly compared with the traditional PLDA, and is more effective at reducing the performance gap than PLDA interpolation approach.

This paper is organized as follows. Firstly, the general theory of standard Gaussian PLDA is briefly introduced in which the re-estimation and scoring formulas of PLDA are

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

given. After that, we describe in detail the proposed transfer learning method of PLDA, including the objective function and re-estimation formulas. Experiments based on the NIST SRE and Switchboard corpus are then conducted to verify the effectiveness of this proposed method.

2. STANDARD GAUSSIAN PLDA

PLDA has gained popularity as an elegant classification tool to find target classes in recent NIST challenges. In this paper, we use the Gaussian PLDA (G-PLDA) after i-vector length normalization. In Gaussian PLDA, the i-vector x_{ij} for the *j*th utterance of speaker *i* is decomposed as follows.

$$x_{ij} = \mu + \phi\beta_i + \varepsilon_{ij} \tag{1}$$

where μ represents the mean of development data, β_i is an identity variable of speaker *i* having a standard normal prior *N(0,I)*, matrix ϕ constrains the dimension of the speaker subspace, and the residual ε_{ij} contains the session factors following a normal distribution with mean 0 and covariance matrix Σ.

For $i = 1,...,N, j = 1,...,M_i$, let η_{ij} denotes the first order statistic（x_{ij} - μ）. Then the mean value of the first order statistic of speaker *i* is defined as F_i,

$$F_i = \frac{\sum_{j=1}^{M_i} \eta_{ij}}{M_i} \tag{2}$$

where M_i is the number of utterances which belong to speaker *i*. And the posterior distribution of F_i based on the hidden variable β_i is

$$P(F_i \mid \beta_i) = N(\phi\beta_i, \frac{\Sigma}{M_i}) \tag{3}$$

In the E-step, we can calculate the expectation of $P(\beta_i \mid F_i)$ based on Bayes' theorem.

In the M-step, the following log-likelihood function will be maximized.

$$\log\left\{\prod_{i=1}^{N}\prod_{j=1}^{M_i}(P(\eta_{ij},\beta_i))\right\} \tag{4}$$

Finally, the re-estimation formulas of PLDA parameters（ϕ, Σ）are derived as follows.

$$\phi = \left(\sum_{i=1}^{N} M_i F_i E(\beta_i)^T\right)\left(\sum_{i=1}^{N} M_i E(\beta_i \beta_i^T)\right)^{-1} \tag{5}$$

$$\Sigma = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M_i}\left[\eta_{ij}\left(\eta_{ij}^T - E(\beta_i)^T \phi^T\right)\right]}{\sum_{i=1}^{N} M_i} \tag{6}$$

For the scoring of PLDA, supposed two i-vector x_{mod} and x_{tst} for the model and the test utterance respectively, the likelihood ratio between the "same-speaker" hypothesis H_s and "different-speaker" hypothesis H_d is calculated as follows [19,20]:

$$\begin{aligned} s(x_{\mathrm{mod}}, x_{tst}) &= \log\frac{P(x_{\mathrm{mod}}, x_{tst} \mid H_s)}{P(x_{\mathrm{mod}}, x_{tst} \mid H_d)} \\ &= \log N\left(\begin{bmatrix} x_{\mathrm{mod}} \\ x_{tst} \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \Sigma + \phi\phi^T & \phi\phi^T \\ \phi\phi^T & \Sigma + \phi\phi^T \end{bmatrix}\right) \\ &- \log N\left(\begin{bmatrix} x_{\mathrm{mod}} \\ x_{tst} \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \Sigma + \phi\phi^T & 0 \\ 0 & \Sigma + \phi\phi^T \end{bmatrix}\right) \end{aligned} \tag{7}$$

3. TRANSFER LEARNING

Due to the problem of scarce data, the PLDA that is directly optimized on the limited development data of target domain may lead to an over-fitting solution. In view of the distribution similarity between data in the source domain and the target domain, we can utilize the information of the source domain to help find a better optimal solution that adequately reflects both domains and generalizes to the target domain [16]. In this paper, we propose a novel transfer learning method, in which KL regularization factor is added into the objective function of PLDA.

Based on the definition of KL divergence and formula (3), we can derive the representation of KL divergence between the source domain PLDA and the target domain PLDA as follows.

$$KL(P(F_s \mid \beta_i) \| P(F_t \mid \beta_i)) = \left\{ tr(\Sigma_t^{-1}\Sigma_s) + (\phi_t\beta_i - \phi_s\beta_i)^T \Sigma_t^{-1} M_i (\phi_t\beta_i - \phi_s\beta_i) - k + \ln(\frac{\det\Sigma_t}{\det\Sigma_s}) \right\} / 2 \tag{8}$$

where F_s is the mean value of the first order statistic of i-vectors in the source domain, F_t is the mean value of the first order statistic of i-vectors in the target domain.（ϕ_s, Σ_s）are the PLDA parameters of the source domain, and（ϕ_t, Σ_t）are the PLDA parameters of the target domain. *k* is the dimension of F_s and F_t. Additionally, operator tr is to get the trace of matrix $\Sigma_t^{-1}\Sigma_s$.

3.1. Objective Function

Given the PLDA parameters （ϕ_s, Σ_s） of the source domain and the limited development data of the target domain, our objective is to learn the suitable PLDA parameters （ϕ_t, Σ_t） for the target domain. The new optimization objective of transfer learning is defined as follows.

$$\min \sum_{i=1}^{N} (-P(F_t \mid \beta_i) P(\beta_i) + \lambda * KL(P(F_s \mid \beta_i) \parallel P(F_t \mid \beta_i)) \quad (9)$$

where λ is an adjusting weight. In this objective function, the first part is the same as the optimization objective of standard PLDA, and the second part is KL divergence which can be calculated as formula (8). When $\lambda = 0$, this objective function will regress to the original PLDA. With the value of λ increasing, the optimization process will gradually lead to the distribution of the source domain.

3.2. Re-estimation Formula

In formula (9), by setting the derivative of objective function towards ϕ_t to be zero, we have

$$\sum_{i=1}^{N} \sum_{j=1}^{M_i} [\Sigma_t^{-1} \phi_t E(\beta_i \beta_i^T) - \Sigma_t^{-1} \eta_{ij} E(\beta_i)^T \\ + \lambda \Sigma_t^{-1} (\phi_t \beta_i - \phi_s \beta_i) \beta_i^T] = 0 \quad (10)$$

And then

$$\phi_t = \left[\sum_{i=1}^{N} (M_i F_i E(\beta_i)^T + \lambda \phi_s M_i E(\beta_i \beta_i^T)) \right] / \\ \left[(1+\lambda) \sum_{i=1}^{N} M_i E(\beta_i \beta_i^T) \right] \quad (11)$$

The maximum likelihood estimation for Σ_t is obtained through

$$\sum_{i=1}^{N} \sum_{j=1}^{M_i} [-\frac{1+\lambda}{2} \Sigma_t + \frac{1}{2} \phi_t E(\beta_i \beta_i^T) \phi_t^T + \frac{1}{2} \eta_{ij} \eta_{ij}^T - \eta_{ij} E(\beta_i^T) \phi_t^T \\ + \frac{\lambda}{2} \Sigma_s + \frac{\lambda}{2} (\phi_t \beta_i - \phi_s \beta_i)(\phi_t \beta_i - \phi_s \beta_i)^T] = 0 \quad (12)$$

And then,

$$\Sigma_t = \sum_{i=1}^{N} \sum_{j=1}^{M_i} [\eta_{ij} \eta_{ij}^T + \phi_t E(\beta_i \beta_i^T) \phi_t^T - 2 \eta_{ij} E(\beta_i^T) \phi_t^T \\ + \lambda (\Sigma_s + (\phi_t \beta_i - \phi_s \beta_i)(\phi_t \beta_i - \phi_s \beta_i)^T)] / [(1+\lambda) \sum_{i=1}^{N} M_i] \quad (13)$$

Finally, we get the re-estimation formula of ϕ_t and Σ_t as follows.

$$\phi_t = w \phi_s + (1-w) \phi^{'} \quad (14)$$

$$\Sigma_t = w \Sigma_s + (1-w) \Sigma^{'} + w \sigma \quad (15)$$

where $w = \lambda / (1+\lambda)$. $\phi^{'}$ and $\Sigma^{'}$ will be updated in each step based on formula (5) and (6). σ is a new factor, which can be calculated as follows.

$$\sigma = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} \left(\phi_s E\left(\beta_i \beta_i^{T} \right) \phi_s^T - \phi_t E\left(\beta_i \beta_i^{T} \right) \phi_s^T \right)}{\sum_{i=1}^{N} M_i} \quad (16)$$

It can be seen that the new learned PLDA parameters （ϕ_t, Σ_t） are the linear fusion of source domain parameters and target domain parameters, where the fusion coefficient is represented by *w*. This is different with PLDA interpolation [11] in that the proposed re-estimation will be conducted in each EM step. After the transfer learning, the scoring of PLDA is also based on formula (7), which is the same as the standard Gaussian PLDA.

4. EXPERIMENTS

To evaluate the effective performance of the proposed transfer learning method to domain mismatch, experiments were conducted based on the NIST SRE and Switchboard (SWB) corpus. We extracted 32-dimension MFCC with appended delta coefficients from each speech utterance. The total variability subspace of dimension 400 was estimated by the Baum-Welch statistics. And the PLDA was trained with speaker subspace of dimension 120. All the results presented in this paper concentrated on female trials only.

From the SWB corpus, 11,453 utterances from 993 speakers were picked out to train a UBM containing 1,024 Gaussians. And we used the same training data to estimate matrix T and PLDA parameters. In our experiments of domain adaptation, the SWB corpus was used as the data of source domain.

For the performance evaluation, the NIST SRE10 [21] telephone data (condition-5) was used as enroll and test sets, which includes 355 target and 15,958 non-target trials.

For each i-vector, the centering process was based on the mean of its domain, but the whitening process was based on the SWB statistics. Note that this was a key step to produce the best results of domain adaptation. If the whitening process was based on different statistics, the verification performance would deteriorate greatly. The impact of length normalization has been also addressed in [12]. In our case, the system setup was listed in Table 1.

Table 1 System setup of our experiments

Domain	UBM,T	Centering	Whitening	PLDA
Source	SWB	SWB	SWB	SWB
Target	SWB	SRE	SWB	SRE

To conduct domain adaptation, we designed four experiments based on varying amounts of target domain data. From the NIST SRE corpus, we selected 150, 300, 600 and 1,325 speakers to act as the development data of target domain respectively. Four PLDA of the target domain would be learned based on the corresponding dataset respectively. And PLDA of the source domain would be learned based on the SWB data.

We compared the performances of transfer learning with PLDA of the source domain (Source PLDA), PLDA of the target domain (Target PLDA) and PLDA interpolation. For PLDA interpolation, the PLDA parameters of source domain and the PLDA parameters of target domain were fused directly with an interpolation parameter which would be adjusted based on the amount of target domain data [12].

In our experiments, the equal error rate (EER) and the 2010 minimum decision cost function (minDCF) were calculated as evaluation metrics. The EER results of performance evaluation of the NIST SRE10 telephone data were shown as follows.

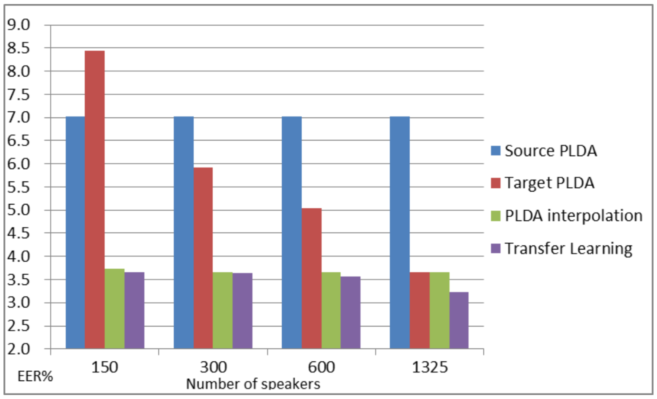

Figure 1 The EER results of source PLDA, target PLDA, PLDA interpolation and transfer learning

With the same test data, Figure 1 illustrates the EER results of different methods when 150, 300, 600 and 1,325 speakers were selected to act as the development data of target domain respectively. For the method of Source PLDA, EER was fixed with the value of 7.02%, since it didn't utilize the development data of target domain. For the method of Target PLDA, with the increasing number of speakers, EER was reduced from the value of 8.45% to 5.92%, 5.04% and 3.66% respectively. When there were only 150 speakers in development data, EER of Target PLDA was even worse than Source PLDA. However, when 1,325 speakers were all used, the parameters of Target PLDA were optimized well and had quite better performance than Source PLDA. This proved the importance of development data for the robust optimization of PLDA. PLDA interpolation remained, basically, constant across the number of speakers and the transfer learning results were also constant until a larger number of speakers were available.

For the methods of PLDA interpolation and transfer learning, the EER results of 150 speakers were 3.73% and 3.66% respectively, which reduced the EER greatly by 55.9% and 56.7% compared with the method of Target PLDA. When there were 300 speakers, the EER was reduced from 5.91% of Target PLDA to 3.66% of PLDA interpolation and 3.64% of transfer learning respectively. This showed that both adapted methods were powerful to reduce the performance gap.

For all four experiments, the EER results of transfer learning were always the lowest, which demonstrated the significant effectiveness of our method. Specifically, the EER could be further reduced by transfer learning even when all 1,325 speakers were used. This showed that the parameters of Target PLDA and PLDA interpolation were not full optimized, but the proposed method could utilize the development data of source domain more effectively to help find a better optimal solution of PLDA parameters for the target domain.

In Table 2, the minDCF results for different number of speakers are further compared. In all cases, our proposed transfer learning method had the smallest minDCF.

Table 2 minDCF results of source PLDA, target PLDA, PLDA interpolation and transfer learning

Number of Speakers	150	300	600	1325
Source PLDA	0.0575	0.0575	0.0575	0.0575
Target PLDA	0.0696	0.0611	0.0499	0.0482
PLDA interpolation	0.0551	0.0428	0.0482	0.0437
Transfer Learning	0.0465	0.0403	0.0446	0.0400

5. CONCLUSION

In this paper, we have successfully designed a transfer learning method from a source domain with sufficient development data to a new target domain. Our proposed method is based on KL divergence, which can effectively utilize the similar information between the source domain and target domain. This could help find a better optimal solution that adequately reflects both domains and generalizes to the target domain. We have conducted four experiments based on varying amounts of target domain data based on the NIST SRE and Switchboard corpus. And experimental results demonstrated that our proposed method could produce the largest gain of performance in EER and minDCF, compared with the traditional PLDA and PLDA interpolation approach.

6. ACKNOWLEDGMENT

This work was partly supported by the National Natural Science Foundation of China (Grant No. 61105026 and No. 11274259).

7. REFERENCES

[1] D. Reynolds, T. Quatieri, R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, 10(1-3), pp. 19-41, 2000.

[2] W. Campbell. "Support vector machines using GMM supervectors for speaker verification," *Signal Processing*

Letters, IEEE, vol.13, no.5, pp.308-311, May 2006.

[3] W. Campbell, D. Sturim, D. Reynolds, A. Solomonoff. "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, 2006.

[4] P. Kenny, 2006, "Joint factor analysis of speaker and session variability: theory and algorithms," *Technical Report CRIM-06/08-14*.

[5] D. Matrouf, N. Scheffer, B. Fauve, et al. "A straightforward and efficient implementation of the factor analysis model for speaker verification," In *Proc. INTERSPEECH*. 2007: 1242-1245.

[6] N. Dehak, P.J. Kenny, R. Dehak, et al. "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, no. 4, pp. 788-798.

[7] S.J.D. Prince and J.H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE ICCV*, Rio de Janeiro, Brazil, Oct. 2007.

[8] D. Garcia-Romero and C.Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. INTERSPEECH*, Florence, Italy, Aug. 2011, pp. 249-252.

[9] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey, The Speaker and Language Recognition Workshop*, Brno, Czech Republic, Jun. 2010.

[10] J. Villalba and E. Lleida, "Bayesian adaptation of PLDA based speaker recognition to domains with scarce development data," in *Odyssey: The Speaker and Language Recognition Workshop*, Singapore, 2012.

[11] D. Garcia-Romero and A. McCree, "Supervised Domain Adaptation for i-vector based speaker recognition," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014.

[12] D. Garcia-Romero, A. McCree, S. Shum, N. Brummer, and C. Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Proc. Odyssey, The Speaker and Language Recognition Workshop*, Joensuu, Finland, Jun. 2014.

[13] O. Glembek, J. Ma, P. Matejka, B. Zhang, O. Plchot, L. Burget and S. Matsoukas, "Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014.

[14] H. Aronowitz, "Inter dataset variability compensation for speaker recognition," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014.

[15] A. Kanagasundaram, D. Dean and S. Sridharan, "Improving out-domain PLDA speaker verification using unsupervised inter-dataset variability compensation approach," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, 2015.

[16] X.D. Cao, D. Wipf, F. Wen, G.Q. Duan, J. Sun, "A practical transfer learning algorithm for face verification," in *Proc. IEEE ICCV*, p 3208-3215, 2013.

[17] S.J. Pan, Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering,* vol. 22, no. 10, pp. 1345-1359, 2010.

[18] L. Deng, X. Li, "Machine Learning Paradigms for Speech Recognition: An Overview," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1060-1089, 2013.

[19] S. Cumani, N. Brummer, L. Burget, P. Laface, O. Plchot, and V. Vasilakakis, "Pairwise discriminative speaker verification in the-vector space," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1217-1227, 2013.

[20] Q.Y. Hong, L. Li, M. Li, L. Huang, L.H. Wan and J. Zhang, "Modified-prior PLDA and Score Calibration for Duration Mismatch Compensation in Speaker Recognition System," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015.

[21] "The NIST year 2010 Speaker Recognition Evaluation plan," (Available at http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf), 2010.