Robust speaker recognition using library of cross-domain variation compensation transforms

Houjun Huang, Shengyu Yao, Ruohua Zhou✉ and Yonghong Yan

Although the state-of-the-art i-vector-based probabilistic linear discriminant analysis systems resulted in promising performances in the National Institute of Standards and Technology speaker recognition evaluations, the impact of domain mismatch when the system development data and the evaluation data are collected from different sources remains a challenging problem. This issue was a focus of the Johns Hopkins University 2013 speaker recognition workshop where a domain adaptation challenge (DAC13) corpus was created to address it. The cross-domain variation compensation (CDVC) approach has been recently proposed to address it when in-domain development data are available. The work reported by the present authors addresses this issue when in-domain development data are unavailable using a library of CDVC transforms. This approach is evaluated on the DAC13 corpus and is shown to be more powerful than nuisance attribute projection-based inter-dataset variability compensation and the whitening library.

Introduction: The introduction of i-vectors [1] and probabilistic linear discriminant analysis (PLDA) [2, 3] provides promising results in the recent National Institute of Standards and Technology (NIST) speaker recognition evaluations (SREs). However, the success of i-vectors-based PLDA systems is dependent on the availability of a large development set containing thousands of multi-session speakers. Furthermore, for maximum effectiveness, the development set must match the evaluation set. Mismatches due to domain variation can lead to significant degradation in speaker recognition performance, as has been demonstrated by recent studies on the domain adaptation challenge (DAC13) corpus created in the JHU 2013 speaker recognition workshop [4] to address the impact of domain mismatch.

Many effective approaches for cross-domain speaker recognition have been reported in recent years. Domain adaptation approaches for PLDA based i-vector speaker recognition systems that aim at adapting out-domain PLDA models to match the evaluation domain were recently proposed in [5–8]. The source normalisation (SN) method reported in [9] computes the inter-speaker scatter matrix as the average of individual subdomain inter-speaker scatter matrices, thereby removing the inter-domain bias from the final matrix and avoiding an overestimation of inter-speaker variability. The within-class covariance correction approach proposed in [10] relies on a conventional computation of the between speaker scatter matrix, but adds a correction term to the within scatter matrix using the cross-domain scatter obtained from a source-labelled dataset. The inter-dataset variability compensation (IDVC) approach via nuisance attribute projection (NAP) to compensate dataset mismatch in the i-vector space directly was reported in [11]. The nearest neighbour-based i-vector normalisation method was proposed in [12] to compensate for the shift due to the dataset mismatch by normalising each i-vector with a set of its nearest neighbours from the development set. A cross-domain variability compensation method using a bank of whitening transforms was reported in [13] to address domain mismatch when development data are derived from a combination of sources. We have recently reported the cross-domain variation compensation (CDVC) approach for cross-domain speaker recognition in [14]. This CDVC approach was shown to be more powerful than the nearest neighbour-based i-vector normalisation method and the IDVC approach when an in-domain development dataset is available.

When a speaker recognition system is applied in unseen scenarios, matched development data are not always available. However, mismatched development data are almost always adequate. This work focuses on cross-domain speaker recognition using multi-source out-domain development data. In this case, the IDVC [11] approach and the library of whitening transforms [13] achieved great performance improvements. This Letter presents an alternative approach in that rather than relying on a single CDVC transform, a library of CDVC transforms is created by partitioning a development set into subsets to train a bank of CDVC parameters that can be applied on an every i-vector of evaluation set using a maximum-likelihood selection process.

Related work: IDVC aims at directly estimating and removing cross-dataset variability in the i-vector space. Given a development dataset, it is split into subsets according to different sources. For each subset all i-vectors are averaged to get its centre. Given the set of source specific centres, principal component analysis is used to find a basis, which is treated as the inter-dataset variability subspace, for the subspace spanned by these centres.

In the whitening library approach, rather than relying on one whitener, a library of whitening transforms is created by dividing a development set into partitions to train a bank of whitening parameters that are applied on an utterance-by-utterance basis using an automatic selection process.

CDVC transforming library: In the CDVC approach, the inter-domain variation is modelled as an additive factor with normal distribution in the i-vector space. For an evaluation domain i-vector $\boldsymbol{\omega}$ cross-domain variability is compensated as follows:

$$\tilde{\boldsymbol{\omega}} = \boldsymbol{\omega} - \boldsymbol{\varepsilon} \tag{1}$$

where $\tilde{\boldsymbol{\omega}} \sim \mathcal{N}(\boldsymbol{\mu}_{\text{ori}}, \boldsymbol{\Sigma}_{\text{ori}})$ is the compensated i-vector and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\mu}_{\varepsilon}, \boldsymbol{\Sigma}_{\varepsilon})$ is the inter-dataset variation. Hyper parameters $(\boldsymbol{\mu}_{\text{ori}}, \boldsymbol{\Sigma}_{\text{ori}})$ are pre-estimated using the PLDA training set. Ideally, hyper parameters $(\boldsymbol{\mu}_{\varepsilon}, \boldsymbol{\Sigma}_{\varepsilon})$ should be trained using in-domain development data. When $\boldsymbol{\omega}$ is given, CDVC transforming can be done with the maximum-a-posteriori algorithm

$$\begin{aligned}\tilde{\boldsymbol{\omega}} &= \underset{\tilde{\boldsymbol{\omega}}}{\arg\max}\,\{\ln f(\tilde{\boldsymbol{\omega}}|\boldsymbol{\omega})\} \\ &= \left(\sum_{\text{ori}}^{-1} + \sum_{\varepsilon}^{-1}\right)^{-1}\left(\sum_{\varepsilon}^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu}_{\varepsilon}) + \sum_{\text{ori}}^{-1}\boldsymbol{\mu}_{\text{ori}}\right)\end{aligned} \tag{2}$$

In practice, however, the strictly matched data may be unavailable to estimate the hyper parameters $(\boldsymbol{\mu}_{\varepsilon}, \boldsymbol{\Sigma}_{\varepsilon})$. If we have a large development set derived from a combination of sources at hand, it can be split into subsets according to different sources. i-vectors of each subset are used to train source-specific CDVC parameters $(\boldsymbol{\mu}_{\varepsilon,j}, \boldsymbol{\Sigma}_{\varepsilon,j})$ to build a bank of CDVC transforms with hyper parameters $(\boldsymbol{\mu}_{\text{ori}}, \boldsymbol{\Sigma}_{\text{ori}})$ shared by all sources. Each i-vector $\boldsymbol{\omega}$ in the evaluation set is compared against a source specific Gaussian model as $\mathcal{N}(\boldsymbol{\omega}|\boldsymbol{m}_j, \boldsymbol{S}_j)$, where parameters $(\boldsymbol{m}_j, \boldsymbol{S}_j)$ are estimated from a source-specific subset. The CDVC hyper parameters $(\boldsymbol{\mu}_{\varepsilon,i}, \boldsymbol{\Sigma}_{\varepsilon,i})$ of the source whose Gaussian model produces the maximum likelihood are chosen to do domain mismatch compensation as follows:

$$i = \underset{j}{\arg\max}\,\{\mathcal{N}(\boldsymbol{\omega}|\boldsymbol{m}_j, \boldsymbol{S}_j)\} \tag{3}$$

$$\tilde{\boldsymbol{\omega}} = \left(\sum_{\text{ori}}^{-1} + \sum_{\varepsilon,i}^{-1}\right)^{-1}\left(\sum_{\varepsilon,i}^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu}_{\varepsilon,i}) + \sum_{\text{ori}}^{-1}\boldsymbol{\mu}_{\text{ori}}\right) \tag{4}$$

The maximum-likelihood selection process ensures each i-vector of the evaluation set will be transformed using the most matched source specific CDVC parameters in the library. For each i-vector, if there are one or more known sources that well match it, its cross-domain variability could be compensated somewhat.

Experiments and results: Experiments were conducted on the DAC13 corpus. In the DAC13 corpus, the SWB dataset (which consists of telephone calls from the Switchboard-I and -II corpora) was provided as the speaker-labelled development set, the SRE dataset (which consists of telephone calls from NIST SREs 2004–2008) was provided to do domain adaptation, and telephone calls of NIST SRE-2010 were used as the evaluation set. More details about the DAC13 corpus can be found in [4]. In this Letter, the SRE dataset is assumed to be unavailable.

600-dimension i-vectors and evaluation trials officially provided by the workshop were used for our experiments. i-vectors and trial lists are available online [4]. The length normalisation approach [15] was first applied on all i-vectors. The CDVC compensated i-vectors were also be normalised to unit length.

The Gaussian PLDA [3, 15] system with a gender-independent PLDA model trained using all SWB data was taken as our baseline system. The effectiveness of the proposed approach was compared with that of the IDVC [11] method and the whitening library [13].

The IDVC approach was applied as the same as in [11], where the SWB dataset is divided into 12 gender-dependent subsets according to the collection types (SWB1; SWB2 phases 1–3; and SWB2 cellulars 1–2). The first ten principal components of the resulting scatter matrix were treated as nuisance directions and removed from all i-vectors for both the development and the evaluation dataset. The whitening library approach was conducted as in [13], where a library of six whitening transforms was built by splitting the SWB dataset into six gender-independent partitions using the labelled collection types. The CDVC transforming library was constructed using gender-independent data from the six source-specific subsets of the SWB. In Table 1, results are presented for the male trials, female trials and pooled trials in metrics of equal error rates (EERs) and the minimum detection cost function (DCF) (old and new) as described in [16].

Table 1: Comparison of effectiveness of different compensation approaches

Trials	Compensation	EER (%)	DCF_{old}	DCF_{new}
Male	None (baseline)	5.50	0.44	0.61
	IDVC	3.02	0.29	0.45
	Library of whitener	3.05	0.27	0.43
	Library of CDVC	**2.59**	**0.24**	**0.40**
Female	None (baseline)	7.36	0.47	0.70
	IDVC	4.06	0.35	0.59
	Library of whitener	3.93	0.32	0.58
	Library of CDVC	**3.38**	**0.30**	**0.56**
Pooled	None (baseline)	6.49	0.46	0.67
	IDVC	3.60	0.33	0.54
	Library of whitener	3.60	0.30	0.52
	Library of CDVC	**3.02**	**0.27**	**0.50**

The results in Table 1 show that when a large development dataset collected from all kinds of sources is available, the IDVC approach, the whitening library and the proposed library of CDVC transforms could provide promising performance improvements when the i-vector/PLDA-based speaker recognition system is applied in new scenarios without any matched development data. Among them, the IDVC approach and the whitening library show comparable performance, while the CDVC transforming library performs the best. On combination of the male and female trials, a library of CDVC transforms helps the speaker recognition system achieve a relative performance improvement of 53.5% in EER, 41.3% in DCF_{old} and 25.4% in DCF_{new} over the baseline system.

Conclusion: This Letter presents an approach to perform domain mismatch compensation for cross-domain speaker recognition where in-domain development data are assumed to be unavailable. Results indicate that a CDVC transforming strategy that uses an individual CDVC model selected from a bank of various domain-specific CDVC transforms is useful for reducing domain mismatches between system development data and evaluation data. The effectiveness of the approach was evaluated on the DAC13 corpus, and was shown to be more powerful than the NAP-based inter-dataset variability compensation [11] and the whitening library approach [13].

Acknowledgments: This work was partially supported by the National Natural Science Foundation of China (nos. 11461141004, 91120001, 61271426), the Strategic Priority Research Program of the Chinese Academy of Sciences (grant nos. XDA06030100, XDA06030500), the National 863 Program (no. 2012AA012503) and the CAS Priority Deployment Project (no. KGZD-EW-103-2).

Submitted: *8 September 2015* E-first: *7 January 2016*
doi: 10.1049/el.2015.3174

Houjun Huang, Shengyu Yao, Ruohua Zhou and Yonghong Yan (*Institute of Acoustics, Chinese Academy of Sciences, Beijing, People's Republic of China*)

✉ E-mail: zhouruohua@hccl.ioa.ac.cn

References

1 Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., and Ouellet, P.: 'Front-end factor analysis for speaker verification', *IEEE Trans. Audio Speech Lang. Process.*, 2011, **19**, (4), pp. 788–798, doi: 10.1109/TASL.2010.2064307

2 Prince, S., and Elder, J.: 'Probabilistic linear discriminant analysis for inferences about identity'. IEEE 11th Int. Conf. on Computer Vision, 2007, Rio de Janeiro, Brazil, October 2007, pp. 1–8, doi: 10.1109/ICCV.2007.4409052

3 Kenny, P.: 'Bayesian speaker verification with heavy-tailed priors'. Odyssey: The Speaker and Language Recognition Workshop, Brno, Czech Republic, June 2010

4 JHU 2013 speaker recognition workshop, Available on http://www.clsp.jhu.edu/workshops/archive/ws13-summer-workshop/groups/spk-13/, 2013

5 Villalba, J., and Lleida, E.: 'Bayesian adaptation of PLDA based speaker recognition to domains with scarce development data'. Odyssey: The Speaker and Language Recognition Workshop, Singapore, June 2012

6 Garcia-Romero, D., and McCree, A.: 'Supervised domain adaptation for i-vector based speaker recognition'. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2014, Florence, Italy, May 2014, pp. 4047–4051, doi: 10.1109/ICASSP.2014.6854362

7 Shum, S.H., Reynolds, D.A., Garcia-Romero, D., and Mc-Cree, A.: 'Unsupervised clustering approaches for domain adaptation in speaker recognition systems'. Odyssey: The Speaker and Language Recognition Workshop, Joensuu, Finland, June 2014

8 Garcia-Romero, D., McCree, A., Shum, S., Brummer, N., and Vaquero, C.: 'Unsupervised domain adaptation for i-vector speaker recognition'. Odyssey: The Speaker and Language Recognition Workshop, Joensuu, Finland, June 2014

9 McLaren, M., and van Leeuwen, D.: 'Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources', *IEEE Trans. Audio Speech Lang. Process.*, 2012, **20**, (3), pp. 755–766, doi: 10.1109/TASL.2011.2164533

10 Glembek, O., Ma, J., Matejka, P., Zhang, B., Plchot, O., Burget, L., and Matsoukas, S.: 'Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems'. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2014, Florence, Italy, May 2014, pp. 4032–4036, doi: 10.1109/LSP.2015.2451591

11 Aronowitz, H.: 'Inter dataset variability compensation for speaker recognition'. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2014, Florence, Italy, May 2014, pp. 4002–4006, doi: 10.1109/ICASSP.2014.6854353

12 Sadjadi, S.O., Pelecanos, J.W., and Zhu, W.: 'Nearest neighbor based i-vector normalization for robust speaker recognition under unseen channel conditions'. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2015, Brisbane, Australia, April 2015, pp. 4684–4688

13 Singer, E., and Reynolds, D.: 'Domain mismatch compensation for speaker recognition using a library of whiteners', *IEEE Signal Process. Lett.*, 2015, **22**, (11), pp. 2000–2003, doi: 10.1109/LSP.2015.2451591

14 Huang, H., Zhou, R., and Yan, Y.: 'Cross-domain variation compensation for robust speaker verification', *Electron. Lett.* 2015, **51**, (21), pp. 1706–1707

15 Garcia-Romero, D., and Espy-Wilson, C.Y.: 'Analysis of *i*-vector length normalization in speaker recognition systems'. 12th Annual Conf. of the Int. Speech Communication Association, INTERSPEECH 2011, Florence, Italy, August 2011, pp. 249–252

16 NIST Year 2010 Speaker Recognition Evaluation Plan, Available on http://www.nist.gov/itl/iad/mig/upload/NIST_SRE10_evalplanr6.Pdf, 2010