SEMI-SUPERVISED NUISANCE-ATTRIBUTE NETWORKS FOR DOMAIN ADAPTATION

Weiwei Lin, Man-Wai Mak and Youzhi Tu

The Hong Kong Polytechnic University
Dept. of Electronic and Information Engineering, Hong Kong SAR

Jen-Tzung Chien

National Chiao Tung University
Dept. of Electrical and Computer Engineering, Taiwan

ABSTRACT

How to overcome the training and test data mismatch in speaker verification systems has been a focus of research recently. In this paper, we propose a semi-supervised nuisance attribute network (SNAN) to reduce the domain mismatch in i-vectors and x-vectors. SNANs are based on the idea of nuisance attribute removal in inter-dataset variability compensation (IDVC). But instead of measuring the domain variability through the dataset means, SNANs use the maximum mean discrepancy (MMD) as part of their loss function, which enables the network to find nuisance directions in which domain variability is measured up to infinite moment. The architecture of SNANs also allows us to incorporate the out-of-domain speaker labels into the semi-supervised training process through the center loss and triplet loss. Using SNANs as a preprocessing step for PLDA training, we achieve a relative improvement of 11.8% in EER on NIST 2016 SRE compared to PLDA without adaptation. We also found that the semi-supervised approach can further improve SNANs' performance.

Index Terms— Speaker verification; x-vectors; i-vectors; domain adaptation; maximum mean discrepancy

1. INTRODUCTION

In the past few years, we have witnessed the significant advances in speaker verification (SV), especially in text-independent speaker verification. However, the current state-of-the-art SV systems still lack the robustness against the mismatch in training and test data. There are a lot of realistic scenarios in which the training speech data and test speech data have severe mismatch. Ideally, we want the trained system conformed to the distribution of test data. However, often we do not have enough data from the test environment or these data do not have labels for supervised training. It is desirable to use the existing training data and a small amount of data from the test environment to modify the system to meet the need, which is essentially what domain adaptation (DA) does.

This work was supported by RGC of Hong Kong, Grant 152518/16E and 152137/17E, and Taiwan MOST, Grant 107-2634-F-009-003.

Earlier attempts in i-vector based DA [1, 2] require the in-domain data to have speaker labels. Over the years, several unsupervised DA techniques have been proposed, including inter-dataset variability compensation (IDVC) [3, 4], source normalization (SN) [5] and discriminative multi-domain PLDA [6]. In particular, it has been shown [3] that IDVC is able to reduce the the mismatch between NIST telephone data and Switchboard data, and in several NIST 2016 submissions [7, 8], IDVC was found to be very helpful in boosting system performance.

Despite the success, IDVC merely transforms the i-vectors to a space with less domain variability. It does not consider any information in the in-domain data. To address this limitation, Rahman *et al.* [9] proposed a domain-invariant i-vector extraction method that takes the in-domain prior information into account by incorporating the mean and covariance of in-domain data into the prior of i-vectors. Instead of transforming the i-vectors, we may adapt the PLDA backend [10]. For example, Alam *et al*. [11] borrowed a covariance transformation technique – called correlation alignment (CORAL) [12] – from the computer vision community to align the covariance of the out-of-domain and in-domain features in an unsupervised manner. Zhang *et al*. [13] investigated how to use transfer learning for domain adaptation to improve the performance of in-domain speaker verification task. Recently, domain adversarial training [14] has also been applied to enhance the domain robustness of speaker verification systems [15].

To better utilize the statistics of multi-source data, this paper uses maximum mean discrepancy (MMD) as an objective function for measuring multi-source mismatch. Maximum mean discrepancy is a nonparametric method for measuring the distance between two distributions [16, 17]. With a properly chosen kernel, MMD can utilize all moments of data. In [18, 19], we generalized MMD to measure the discrepancies among multiple distributions and incorporated the measure into the objective function for training autoencoders. This paper is an extension of this earlier work for x-vector adaptation. Specifically, we introduce a new network structure that enables the use of MMD to find the nuisance directions of i-vectors [20] and x-vectors [21]. In addition, we add

triplet loss and center loss into the objective function of the network, which enable us to leverage the speaker labels in the out-of-domain data. As a result, the network can exploit both supervised learning (through triplet loss and center loss) and unsupervised learning (through the MMD loss) to remove the nuisance attributes of i-vectors and x-vectors. We refer to the network as semi-supervised nuisance-attribute network (SNAN).

2. DOMAIN MISMATCH AND MAXIMUM MEAN DISCREPANCY

2.1. Inter-dataset Variability Compensation

Inter-dataset variability compensation (IDVC) [3] follows the subspace removal approach proposed in [22]. It aims to find the directions in the i-vector space with the largest inter-dataset variability and removes the i-vector variability in these directions. This is achieved by projecting the i-vectors $\mathbf{x}$'s as follows:

$$\hat{\mathbf{x}} = (\mathbf{I} - \mathbf{W}\mathbf{W}^{\mathsf{T}})\mathbf{x}, \tag{1}$$

where the columns of $\mathbf{W}$ span the subspace of unwanted variability. $\mathbf{W}$ comprises the eigenvectors of the covariance matrix of the subsets' means. Therefore, in IDVC the domain mismatch is defined by the variances and covariances of subsets' means. However, the mismatch of datasets may not only manifest in the dataset means, but also in the higher-order statistics of these datasets.

2.2. Maximum Mean Discrepancy

The theoretical studies in DA [23] suggest that it is important to have a good measurement of the divergence between the data distributions of different domains. Maximum mean discrepancy is a distance measure on the space of probability. Given two sets of samples $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{y}_j\}_{j=1}^M$, MMD computes the difference in the means of two distributions in a high-dimensional space:

$$\mathcal{D}_{\text{MMD}} = \left\| \frac{1}{N}\sum_{i=1}^{N}\phi(\mathbf{x}_i) - \frac{1}{M}\sum_{j=1}^{M}\phi(\mathbf{y}_j) \right\|^2, \tag{2}$$

where ϕ is a feature map. When ϕ is the identity function, the MMD distance simply computes the discrepancy between the sample means. Eq. 2 can be expanded as:

$$\mathcal{D}_{\text{MMD}} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N}k(\mathbf{x}_i,\mathbf{x}_{i'}) - \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}k(\mathbf{x}_i,\mathbf{y}_j) + \frac{1}{M^2}\sum_{j=1}^{M}\sum_{j'=1}^{M}k(\mathbf{y}_j,\mathbf{y}_{j'}), \tag{3}$$

where $k(\cdot,\cdot)$ is a kernel function. In the case of quadratic (Quad) kernels, we have

$$k(\mathbf{x},\mathbf{y}) = (\mathbf{x}^{\mathsf{T}}\mathbf{y} + c)^2. \tag{4}$$

Then, the MMD becomes

$$\mathcal{D}_{\text{MMD}} = 2c\left\| \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i - \frac{1}{M}\sum_{j=1}^{M}\mathbf{y}_j \right\|_F^2 + \left\| \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}} - \frac{1}{M}\sum_{j=1}^{M}\mathbf{y}_j\mathbf{y}_j^{\mathsf{T}} \right\|_F^2, \tag{5}$$

where $\|\cdot\|_F$ is the Frobenius norm. We can see from Eq. 5 that with a quadratic kernel, MMD can match up to the second-order statistics and c can be adjusted to control the contribution the first-order and the second-order moments to the match. Eq. 5 is very similar to CORAL loss proposed in [12]. The major difference is that Eq. 5 penalizes mismatch in both means and second-moments while CORAL only penalizes mismatch in the second-moments.

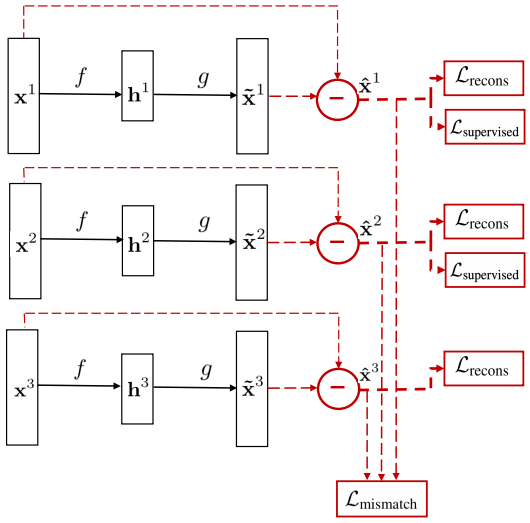

Fig. 1. Architecture of a semi-supervised nuisance-attribute network (SNAN) when data are from three different domains. Solid black arrows represent the connections between neurons. Dashed red arrows represent the signal pathways for computing the domain-mismatch loss $\mathcal{L}_{\text{mismatch}}$, reconstruction loss $\mathcal{L}_{\text{recons}}$ or supervised loss $\mathcal{L}_{\text{supervised}}$ such as center loss or triplet loss. Note that for $\hat{\mathbf{x}}^3$, we do not have $\mathcal{L}_{\text{supervised}}$, which shows the semi-supervised nature of SNAN.

3. SEMI-SUPERVISED NUISANCE-ATTRIBUTE NETWORKS

3.1. Unsupervised Nuisance Attribute Removal

Recall from Eq. 1 that IDVC aims to remove the nuisance information in i-vectors by subtracting $\mathbf{W}\mathbf{W}^{\mathsf{T}}\mathbf{x}$ from the i-

vectors. Our SNAN extends this idea by replacing $\mathbf{W}\mathbf{W}^{\mathsf{T}}\mathbf{x}$ with the output of a network. Specifically,

$$\hat{\mathbf{x}} = \mathbf{x} - \tilde{\mathbf{x}} = \mathbf{x} - g(f(\mathbf{x})), \tag{6}$$

where $\mathbf{h} = f(\mathbf{x})$ is an encoder that maps $\mathbf{x}$ to a latent space and $g(\mathbf{h})$ is a decoder that maps from the latent space back to the input space. Note that in IDVC, $\mathbf{W}$ defines the subspace in which domain variability is the largest. By subtracting out the components of $\mathbf{x}$'s in this subspace, $\hat{\mathbf{x}}$'s will be less domain-dependent. Similarly, $g(f(\mathbf{x}))$ contains all of the domain-specific information. By subtracting out the domain information, $\hat{\mathbf{x}}$'s will become more domain-invariant.

When the data come from multiple sources, we want the transformed data to be as similar to each other as possible. To this end, we define a domain-wise MMD measure. Specifically, given D sets of data $\{\mathbf{x}_i^d\}_{i=1}^{N_d}$, where $d = 1, 2, \ldots, D$, we want the transformed data $\{\hat{\mathbf{x}}_i^d\}_{i=1}^{N_d}$ in Eq. 6 to have small MMD loss:

$$\mathcal{L}_{\text{mismatch}} = \sum_{d=1}^{D} \sum_{\substack{d'=1 \\ d' \neq d}}^{D} \left(\frac{1}{N_d^2} \sum_{i=1}^{N_d} \sum_{i'=1}^{N_d} k(\hat{\mathbf{x}}_i^d, \hat{\mathbf{x}}_{i'}^d) - \frac{2}{N_d N_{d'}} \sum_{i=1}^{N_d} \sum_{j=1}^{N_{d'}} k(\hat{\mathbf{x}}_i^d, \hat{\mathbf{x}}_j^{d'}) + \frac{1}{N_{d'}^2} \sum_{j=1}^{N_{d'}} \sum_{j'=1}^{N_{d'}} k(\hat{\mathbf{x}}_j^{d'}, \hat{\mathbf{x}}_j^{d'}) \right). \tag{7}$$

Eq. 7 can be used as an objective function to measure the discrepancies among multiple domains.

Of course, we also want to retain as much non-domain related information as possible. To this end, we may enforce the network to produce vectors $\hat{\mathbf{x}}$'s that are as close to the original i-vectors $\mathbf{x}$'s as possible. This can be achieved by minimizing the reconstruction loss:

$$\mathcal{L}_{\text{recons}} = \frac{1}{2} \sum_{d=1}^{D} \sum_{i=1}^{N_d} \left\| \mathbf{x}_i^d - \hat{\mathbf{x}}_i^d \right\|^2. \tag{8}$$

Both objectives can be achieved by an antoencoder comprising an encoder network f and a decoder network g, with the total loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mismatch}} + \alpha \mathcal{L}_{\text{recons}}, \tag{9}$$

where α is a parameter controlling the importance of the reconstruction loss.

3.2. Supervised Loss

SNANs can leverage the speaker labels in the out-of-domain data through the center loss [24] or triplet loss [25].

Center loss is motivated by the notion that the softmax loss (also known as cross-entropy loss) can only encourage the deep features of different classes to stay apart [24]. By introducing a center for each class and minimizing the distances between data and their class centers, center loss can help the netwok to learn more discriminative features. To apply center loss to train an SNAN, we consider the network outputs $\hat{\mathbf{x}}_i$'s in Eq. 6 as the feature vectors, where i indexes the training samples in a mini-batch. Denote $y_i \in \{1, \ldots, K\}$ as the class label for the i-th training sample. Then, the center loss is

$$\mathcal{L}_{\text{center}} = \frac{1}{2} \sum_{i=1}^{B} \left\| \hat{\mathbf{x}}_i - \mathbf{c}_{y_i} \right\|^2, \tag{10}$$

where $\mathbf{c}_{y_i}$ is the center of the y_i-th class and B is the number of samples in a mini-batch. Note that $\mathbf{c}_{y_i}$ should be updated using the $\hat{\mathbf{x}}_i$'s in the mini-batch.

Similar to center loss, triplet loss is also based on the idea of maximizing inter-class distance and minimizing intra-class distance. The key components are triplets. A triplet consists of three samples, namely, anchor sample $\hat{\mathbf{x}}_a$, positive sample $\hat{\mathbf{x}}_p$ and negative sample $\hat{\mathbf{x}}_n$. Positive sample shares the same class with the anchor while the negative sample comes from different classes. To learn discriminative features, we need to maximize inter-class distance while minimizing intra-class distance. This can be achieved by the following loss function:

$$\mathcal{L}_{\text{triplet}} = \max \left\{ \left\| \hat{\mathbf{x}}_a - \hat{\mathbf{x}}_p \right\|^2 - \left\| \hat{\mathbf{x}}_a - \hat{\mathbf{x}}_n \right\|^2 + m,\ 0 \right\}, \tag{11}$$

where m is a margin term.

We can also incorporate supervised loss such as center loss and triplet loss into the total loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mismatch}} + \alpha \mathcal{L}_{\text{recons}} + \beta \mathcal{L}_{\text{supervised}}, \tag{12}$$

where β is a parameter controlling the importance of supervised loss. Eq. 12 enables the SNAN to leverage both labeled and unlabeled data. For labeled data, all of the three terms will be involved in the minimization, whereas for unlabeled data, the supervised loss is disabled by setting β to 0. In this way, we can make use of both labeled out-domain data and unlabeled in-domain data. Fig. 1 shows the architecture of SNAN for the case of three domains ($D = 3$).

4. EXPERIMENTAL SETUP

4.1. Speech Data and Acoustic Features

Speech files from NIST 2004–2010 Speaker Recognition Evaluation (hereafter, referred to as SRE04–SRE10)[1] and the development set of SRE16 (SRE16-dev) were used as development data and speech files from the evaluation set of SRE16 (SRE16-eval) were used as test data. For the i-vector system, we used Kaldi [26] to extract 20-dimensional MFCCs plus their delta and double delta coefficients, followed by energy-based voice activity detection (VAD). For the x-vector systems, we extracted 23-dimensional MFCCs using Kaldi's SRE16 recipe, followed by energy-based VAD.

[1] https://www.nist.gov/itl/iad/mig/speaker-recognition

Feature	Adaptation	EER(%)	mCprim	aCprim
i-vector	No Adapt	12.78	0.74	0.94
	IDVC	12.17	0.73	0.90
	SNAN	11.95	0.72	0.87
x-vector	No Adapt	10.74	0.65	0.86
	IDVC	11.24	0.65	0.89
	SNAN	10.35	0.61	0.81

Table 1. The performance of PLDA without adaptation, IDVC, and SNANs without supervised loss ($\beta = 0$ in Eq. 12) on the SRE16 evaluation set. "mCprim" and "aCprim" are the minimum detection cost and the actual detection cost as specified in the evaluation plan of SRE16.

Feature	Supervised Loss	EER(%)	mCprim	aCprim
i-vector	None	11.95	0.72	0.87
	Softmax	11.61	0.71	0.87
	Softmax+Center	11.76	0.72	0.86
	Triplet	11.67	0.72	0.85
x-vector	None	10.35	0.61	0.81
	Softmax	10.28	0.61	0.81
	Softmax+Center	10.31	0.61	0.81
	Triplet	10.57	0.62	0.80

Table 2. The performance of SNANs using different supervised loss functions.

4.2. I/X-vector Extraction and PLDA Model Training

The i-vector/PLDA system is based on a gender-independent UBM with 2048 mixtures and a gender-independent total variability (TV) matrix with 600 total factors. The TV matrix and the UBM were trained on SRE04–SRE10 data. They were used for extracting i-vectors from the speech files in SRE04–SRE10, SRE16-dev and SRE16-eval. X-vectors [21] were extracted using Kaldi's pre-trained model.[2]

I/X-vectors derived from SRE04–SRE10 and the SRE16 development set were used for training the SNANs and the projection matrices in IDVC. The adapted i/x-vectors from SRE04–SRE10 were used to train a gender-independent PLDA model with 200 latent variables. During scoring, the mean of the unlabeled SRE16-dev data was used for centering the adapted enrollment and test i/x-vectors before presenting them to the PLDA model. PLDA scores were normalized by S-norm [27] using SRE16-dev data as the cohort set.

4.3. Configurations of SNANs and IDVC

Each SNAN contains a linear hidden layer with 20 units. Quadratic kernels were used for MMD. We divided the data in SRE04–10 and SRE16 into gender- and language-homogenous subsets to train the projection matrix in IDVC and the SNANs. The rank of $\mathbf{W}$ in Eq. 1 is 6. The size of a mini-batch is 2048.

[2]http://kaldi-asr.org/models.html

5. RESULTS AND DISCUSSIONS

5.1. General Performance Analysis

Table 1 shows the performance of IDVC, SNANs with unsupervised losses only, and an i-vector system without domain adaptation (No Adapt). All systems use PLDA as the back-end. The SNANs use a quadratic kernel with $c = 1$ and α in Eq. 9 was set to 1.

For the i-vector systems, we can see from Table 1 that both of IDVC and SNANs improve the performance in term of EER, although in terms of minimum Cprimary and actual Cprimary, the improvement is minor. We can also observe that the SNANs perform better than IDVC by a small margin.

For the x-vector systems, surprisingly IDVC degrades the performance. Although the SNANs still outperform the baseline system, the improvement is marginal. It could be that the x-vectors are more robust to domain mismatch, which diminishes the benefit of domain adaptation.

5.2. Impacts of Supervised Loss Functions

We have also investigated the impact of the supervised losses on the performance of SNANs. There are three supervised losses, namely, softmax loss (or cross-entropy loss), center loss (Eq. 10) and triplet loss (Eq. 11). Table 2 shows the results of SNANs with these losses together with the SNAN with unsupervised loss only (None). For i-vector based systems, we can see that, in general, adding supervised losses indeed improves the performance of SNANs. However, there is no clear winner among the three supervised losses. For x-vector based systems, adding supervised losses does not give significant improvement in performance.

6. CONCLUSIONS

We proposed semi-supervised nuisance-attribute networks for multiple-source i-vector/x-vector domain adaptation. Compared with IDVC, SNANs can better utilize data statistics and speaker information. Results on SRE16 show that SNANs can improve SV performance under domain mismatch. Results also suggest that the semi-supervised approach can further improve SNANs' performance.

7. REFERENCES

[1] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *Proc. IEEE ICASSP*, pp. 4047–4051, 2014.

[2] J. Villalba and E. Lleida, "Bayesian adaptation of PLDA based speaker recognition to domains with scarce development data," in *Proc. Odyssey*, pp. 47–54, 2012.

[3] H. Aronowitz, "Inter dataset variability compensation for speaker recognition," in *Proc. IEEE ICASSP*, pp. 4002–4006, 2014.

[4] H. Aronowitz, "Compensating inter-dataset variability in PLDA hyper-parameters for robust speaker recognition," in *Proc. Odyssey*, pp. 282–286, 2014.

[5] M. McLaren and D. Van Leeuwen, "Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 755–766, 2012.

[6] A. Sholokhov, T. Kinnunen, and S. Cumani, "Discriminative multi-domain PLDA for speaker verification," in *Proc. IEEE ICASSP*, pp. 5030–5034, 2016.

[7] K. A. Lee *et al.*, "The I4U mega fusion and collaboration for NIST speaker recognition evaluation 2016," in *Proc. Interspeech*, pp. 1328–1332, 2017.

[8] O. Plchot, P. Matějka, A. Silnova, *et al.*, "Analysis and description of ABC submission to NIST SRE 2016," in *Proc. Interspeech*, pp. 1348–1352, 2017.

[9] M. H. Rahman, I. Himawan, D. Dean, C. Fookes, and S. Sridharan, "Domain-invariant i-vector feature extraction for PLDA speaker verification," in *Proc. Odyssey*, pp. 155–161, 2018.

[10] D. Garcia-Romero, A. McCree, S. Shum, N. Brummer, and C. Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Proc. Odyssey*, pp. 260–264, 2014.

[11] M. J. Alam, G. Bhattacharya, and P. Kenny, "Speaker verification in mismatched conditions with frustratingly easy domain adaptation," in *Proc. Odyssey*, pp. 176–180, 2018.

[12] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. 13th AAAI Conf. on Artificial Intelligence*, pp. 2058–2065, 2016.

[13] C. Zhang, S. Ranjan, and J. Hansen, "An analysis of transfer learning for domain mismatched text-independent speaker verification," in *Proc. Odyssey*, pp. 181–186, 2018.

[14] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[15] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *Proc. IEEE ICASSP*, pp. 4889–4893, 2018.

[16] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *Proc. Advances in Neural Information Processing systems*, pp. 513–520, 2007.

[17] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. International Conference on Machine Learning*, pp. 97–105, 2015.

[18] W.-W. Lin, M.-W. Mak, L. Li, and J.-T. Chien, "Reducing domain mismatch by maximum mean discrepancy based autoencoders," in *Proc. Odyssey*, pp. 162–167, 2018.

[19] W.-W. Lin, M.-W. Mak, and J.-T. Chien, "Multisource i-vectors domain adaptation using maximum mean discrepancy based autoencoders," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 12, pp. 2412–2422, 2018.

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

[21] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE ICASSP*, 2018.

[22] A. Solomonoff, C. Quillen, and W. M. Campbell, "Channel compensation for SVM speaker recognition," in *Proc. Odyssey*, vol. 4, pp. 219–226, 2004.

[23] S. B. David, T. Lu, T. Luu, and D. Pál, "Impossibility theorems for domain adaptation," in *Proc. International Conference on Artificial Intelligence and Statistics*, pp. 129–136, 2010.

[24] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. European Conference on Computer Vision*, pp. 499–515, Springer, 2016.

[25] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[27] P. Matejka, O. Novotnỳ, O. Plchot, L. Burget, and J. Cernockỳ, "Analysis of score normalization in multilingual speaker recognition," in *Proc. Interspeech*, pp. 1567–1571, 2017.