

Cross-domain Adaptation with Discrepancy Minimization for Text-independent Forensic Speaker Verification

Zhenyu Wang, Wei Xia, John H.L. Hansen

Center for Robust Speech Systems (CRSS), University of Texas at Dallas, TX 75080

{zhenyu.wang, wei.xia, john.hansen}@utdallas.edu

Abstract

Forensic audio analysis for speaker verification offers unique challenges due to location/scenario uncertainty and diversity mismatch between reference and naturalistic field recordings. The lack of real naturalistic forensic audio corpora with ground-truth speaker identity represents a major challenge in this field. It is also difficult to directly employ small-scale domain-specific data to train complex neural network architectures due to domain mismatch and loss in performance. Alternatively, cross-domain speaker verification for multiple acoustic environments is a challenging task which could advance research in audio forensics. In this study, we introduce a CRSS-Forensics audio dataset collected in multiple acoustic environments. We pre-train a CNN-based network using the VoxCeleb data, followed by an approach which fine-tunes part of the high-level network layers with clean speech from CRSS-Forensics. Based on this fine-tuned model, we align domain-specific distributions in the embedding space with the discrepancy loss and maximum mean discrepancy (MMD). This maintains effective performance on the clean set, while simultaneously generalizes the model to other acoustic domains. From the results, we demonstrate that diverse acoustic environments affect the speaker verification performance, and that our proposed approach of cross-domain adaptation can significantly improve the results in this scenario.

Index Terms: speaker verification, cross-domain adaptation, discrepancy loss, maximum mean discrepancy, forensics, distribution alignment

1. Introduction

The need for forensic speaker recognition arises when an individual contributes his or her voice as evidence such from telephone recordings, wiretaps, audio surveillance, or informant recordings [1]. The use of technology for forensic speaker recognition has been considered as early as 1926 based on speech waveform analysis [2]. Today, forensic speaker recognition is commonly performed by human expert phoneticians, fully or partially, who generally have backgrounds in linguistics and statistics. Full or assisted automatic approaches are also considered as an efficient tool for forensic speaker recognition.

Speaker verification determines whether a test utterance belongs to the target person. Previously, the segment-level vector that represents the speech signal, called *i*-vector, with a probabilistic linear discriminant analysis (PLDA) backend, has dominated the text-independent speaker verification research field [3]. Additionally, *i*-vector variants have been widely used in multiple fields of paralinguistic speech attribute recognition [4, 5]. With the emergence of the large speaker labeled audio datasets and growing computational resources, there is increased interest in applying more effective networks such as *x*-vector and other neural network architecture to speaker verification tasks [6, 7, 8, 9, 10].

As noted above, intrinsic and extrinsic characteristics of audio samples may affect the performance of speaker verification systems in discriminating speakers. Here, we temporarily freeze intrinsic variation of the speaker, and place primary research interest on diversity of acoustic environments as extrinsic variation. This represents a core challenge in most forensic voice analysis investigations. In this context, there does exist shared domain-invariant information across these acoustic environments. In this study, we employ cross-domain adaptation to learn the domain-invariant information across domains containing extrinsic variations of the speaker identity. Firstly, we use a large-scale dataset to train a common feature extractor for speaker identities. The well-trained model is fine-tuned with the clean set of our small in-domain data. Finally, we perform the proposed cross-domain adaptation with discrepancy minimization to extend the fine-tuned model to all other target sets with learned domain-specific subnets. It can significantly improve the speaker verification performance on other sets in different acoustic environments, and still keep an excellent performance on the clean set, alleviating the catastrophic forgetting problem [24]. Experimental results show that the speaker verification performance in a specific dataset with multiple acoustic domains is improved by a large margin using the proposed cross-domain adaptation approach.

A fundamental assumption in machine learning is that training data and test data are drawn from the same distribution [11, 12]. Since this assumption can be undermined by many factors, such as deploying a well-trained model from a large-scale dataset to specific tasks with emotion, duration, language mismatch, etc. [13, 14, 15, 16, 17]. In general, effective audio data set collection under multiple acoustic environments can help, where a portion of the data encompasses source/reference information, while other data can be obtained as field, unknown, or target domain with diverse acoustic mismatch [18]. The source domains and target domain can be similar, but are generally not identical. In domain adaptive training, maximum mean discrepancy (MMD) [19] and triplet discrepancy loss [20, 21] are used for discrepancy evaluation in some domain adaptation motivated tasks [22, 23].

Speech is a highly variable phenomenon. Intrinsic speaker characteristics represent speech traits which are controlled/dependent on the speaker, vs. extrinsic characteristics which are dependent on audio capture and environmental factors [1]. Intrinsic properties include the speaker's age, sex/gender, vocal effort, emotional and physical state (e.g. angry, sad, stressed, distracted, etc.). Extrinsic properties include the recording conditions such as microphone type and placement, background noise and reverberation due to environmental scenarios. The multifaceted sources of variation pose the greatest challenge to accurately model and recognize a speaker regardless of the approaches employed.

As noted above, intrinsic and extrinsic characteristics of audio samples may affect the performance of speaker verification systems in discriminating speakers. Here, we temporarily freeze intrinsic variation of the speaker, and place primary research interest on diversity of acoustic environments as extrinsic variation. This represents a core challenge in most forensic voice analysis investigations. In this context, there does exist shared domain-invariant information across these acoustic environments. In this study, we employ cross-domain adaptation to learn the domain-invariant information across domains containing extrinsic variations of the speaker identity. Firstly, we use a large-scale dataset to train a common feature extractor for speaker identities. The well-trained model is fine-tuned with the clean set of our small in-domain data. Finally, we perform the proposed cross-domain adaptation with discrepancy minimization to extend the fine-tuned model to all other target sets with learned domain-specific subnets. It can significantly improve the speaker verification performance on other sets in different acoustic environments, and still keep an excellent performance on the clean set, alleviating the catastrophic forgetting problem [24]. Experimental results show that the speaker verification performance in a specific dataset with multiple acoustic domains is improved by a large margin using the proposed cross-domain adaptation approach.

2. Overview of the Adaptation Framework

Since the pre-trained system is trained on a large-scale dataset, the model adaptation should be applied to alleviate the impact of data mismatch between training data and data in specific domains. This adaptation includes two stages: adaptation on the clean set and the cross-domain adaptation among different acoustic environments.

2.1. Pre-trained systems

We use an i-vector system as one of our pre-trained systems. The other pre-trained system is based on neural networks, by employing a multiple-layer residual neural network (ResNet) [25] to build a front-end main body as a common feature extractor, followed by a learnable dictionary encoding (LDE) layer [9] to encode a variable-length temporal sequence into a fixed-dimensional representation at the utterance level.

2.2. Adaptation to clean data

The first-stage adaptation method of the GMM-based i-vector system is MLLR, which uses a linear transformation of Gaussian model parameters on top of the UBM [26].

In the first step of our CNN-based system adaptation, the new model is initialized by weights of the pre-trained CNN-based model, freezing the first three layers of ResNet, then it continues to be trained with clean data. The learning rate of this adaptation is 10x smaller than that used in the training process. The inputs across domains will share the domain-independent adapted feature extractor to generate embeddings for further cross-domain adaptation, and the process through the entire common feature extractor is symbolized as Φ_0 .

2.3. Cross-domain adaptation

The objective of our second-stage adaptation is to learn domain-invariant representations in the embedding space for all domains. With the fine-tuned model on the clean set in the first step, following a domain adaptation setup, we take the clean data as the ‘‘source’’ domain data. LENA-booth, far-field, and noisy data (CRSS-Forensics corpus explained in Sec. 3.1.2) are taken as the ‘‘target’’ domain data, which also have speaker labels. At first, a discrepancy loss will be used to reduce the distinction of clean sample representations across domain-specific embedding extractors, and this process is symbolized as Φ_1 . Next, for process Φ_2 , it is easier to extract domain-invariant representations for each source/target pair and then subsequently align the distributions of source and target domain afterwards. In the end, domain-specific classifiers are employed to optimize verification performance for individual domains.

2.3.1. Cross-domain discrepancy minimization

The embedding extractors learned from target domains are likely to make wrong decisions on clean samples near the class boundaries. Based on this fact, a discrepancy loss is employed to compute the absolute values of differences among all pairs of representations from domain-specific embedding extractors. The discrepancy loss function is formulated as,

$$Loss_{dis} = \frac{2}{N(N-1)} \times \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathbb{E}_{\mathbf{x} \in \mathcal{D}_s} \left| \Phi_1^i(\Phi_0(\mathbf{x}^s)) - \Phi_1^j(\Phi_0(\mathbf{x}^s)) \right| \quad (1)$$

In Eq. 1, \mathbf{x}^s is the input feature vector of the clean set. Φ_1^i is an intermediate process between the common feature extractor and

discrepancy evaluation in domain i , N domains are included.

2.3.2. Pair-wise distribution alignment

We employ MMD as a pair-wise discrepancy evaluation, which is a kernel two-sample test for the hypothesis $\mathcal{D}_s = \mathcal{D}_t$ (target and source domains are the same) based on the observed samples. Given that generated distributions are identical, MMD assumes that all the corresponding statistics are the same. The definition is given in Eq. 2, which estimates the discrepancy between each pair of source and target domain,

$$MMD(\mathcal{D}_s, \mathcal{D}_t)_h \triangleq \left\| \mathbb{E}_{\mathcal{D}_s} [\Phi_2^h(\Phi_1^h(\Phi_0(\mathbf{x}^s)))] - \mathbb{E}_{\mathcal{D}_t} [\Phi_2^h(\Phi_1^h(\Phi_0(\mathbf{x}^t)))] \right\|_h^2 \quad (2)$$

Where h is a specific target domain, \mathbf{x}^s and \mathbf{x}^t are sequence input features in the source and target domains. Φ_2^h is an intermediate process between the discrepancy evaluation and MMD evaluation in domain h . Given N target domains, $\widehat{MMD}(\mathcal{D}_s, \mathcal{D}_t)_h$ is an unbiased estimator of $MMD(\mathcal{D}_s, \mathcal{D}_t)_h$, where $\mathbb{E}_{\mathbf{x} \in \mathcal{D}} = 1/n \sum_{i=1}^n \Phi_2^h(\Phi_1^h(\Phi_0(\mathbf{x}_i)))$. Domain-invariant representations for each paired source and target domain can be learned by minimizing the $Loss_{mmd}$,

$$Loss_{mmd} = \sum_{t=1}^N \widehat{MMD}(\mathcal{D}_s, \mathcal{D}_t)_h \quad (3)$$

2.3.3. Domain-specific classification

Each domain-specific embedding extractor is followed by a softmax classifier. We add this classification loss for each classifier to ensure that the embedding distribution performance is improved in each target domain. We use a cross-entropy \mathcal{J} loss as the following,

$$Loss_{cls} = \sum_{t=1}^N \mathbb{E}_{\mathbf{x} \in \mathcal{D}_t} \mathcal{J}(\Phi_2^t(\Phi_1^t(\Phi_0(\mathbf{x}^t))), y_i^t) \quad (4)$$

Given N target domains, the classification losses are computed for each domain-specific classifiers.

2.3.4. Cross-domain adaptation framework

Given a sequence data input, the common feature extractor projects a sequence data into a temporal orderless embedding in a shared process Φ_0 . Φ_1 and Φ_2 processes of the cross-domain adaptation method have three independent subnets corresponding to specific target domains. The domain-invariant representations are learned in this unshared process Φ_1 to minimize discrepancies among target domains. The final embeddings for new target domains are generated through the unshared process Φ_2 from domain-specific classifiers. Also, distributions of both target and source samples are aligned among each domain in Φ_2 . Assuming the cross-domain adaptation can effectively learn a domain-invariant representation, we compute the clean set embedding as the average of target domain embeddings here. The framework of the cross adaptation in each step is illustrated in Fig. 1. The multi-task loss function is formulated as,

$$Loss_{total} = \mu(Loss_{mmd} + Loss_{dis}) + Loss_{cls} \quad (5)$$

Where μ is a variant adaptation factor with a progressive schedule from 0 to 1. in order to stabilize parameter sensitivity in the early adaptation stage.

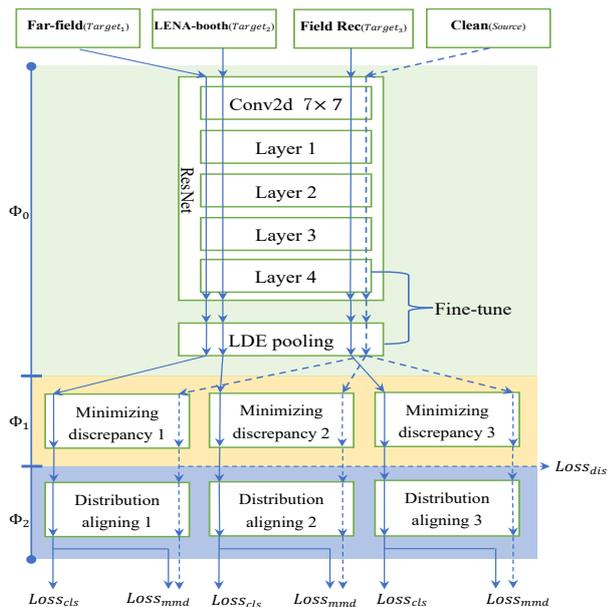


Figure 1: Cross-domain adaptation framework.

3. Experiments

3.1. Data description

3.1.1. VoxCeleb

We use VoxCeleb 1 and 2 [27] dataset which is extracted from videos on YouTube as the training data for our pre-trained systems. Videos included in the dataset are shot in a large number of challenging visual and auditory environments including background chatting, laughter, overlapping speech, and varying room acoustics [28]. The training dataset contains over 1,261,189 utterances from 7273 identities.

3.1.2. CRSS-Forensics

The CRSS-Forensics corpus¹ contains read speech, prompted speech, and spontaneous speech in three conditions: (i) clean (sound booth), (ii) noisy (field recordings), and (iii) high stress (actual police interviews). Two phases are included in the recording process. Phase-1 contains speech data recorded in different acoustic environments such as a sound booth and public environments. Speech in Phase-2 is collected in a law enforcement facility, using an interview room with an actual police officer/detective. Figure 2 shows sample recording environments for the sound booth, public environments, and police interview room.

For the sound booth of Phase-1, speech data is recorded using microphones (sample rate: 44.1 kHz) and a participant body-worn mobile data collection platform called LENA unit (sample rate: 16 kHz). Microphones are positioned at 4 different locations in the booth; the distances from each microphone to the speaker are 10 inches and 0.8, 4, and 8 feet, as shown in Fig. 2 (a). In field environments, speech is collected by the LENA unit worn by the participant, and the recording environments include seven indoor and outdoor places such as hallway (Fig.2 (b)), cafeteria (Fig.2 (c)), walking path (Fig.2 (d)), parking lot (Fig.2 (e)), etc. In Phase-2, a detective interviews the participant concerning a specific investigative scenario while following standard procedures in an interview room (Fig. 2 (f)).

¹The CRSS-Forensics corpus will be released with a license.

Table 1: Data statistics for the CRSS-Forensics corpus

Phase	Session type	Duration	Spk	Utt
1	Clean	64 h	75	6788
	Far-field	64 h	75	6788
	LENA-booth	33.9 h	75	4060
	Field Recording	99.4 h	75	11965
2	Police Interview	40.8 h	58	4615

Table 1 summarizes the specific data size for each session. For the 75 speakers in the corpus, there are 65 English native speakers and 10 non-native speakers, with 27 male speakers, and 48 female speakers. Each participant was allowed to opt-out of Phase-2 (i.e. IRB protocol due to high stress level exposure), so there are 17 speakers fewer from the police interview set.

In this current study, data from Phase-1 is used for cross-domain adaptation. We note that various recording environments are considered as the extrinsic characteristics for audio samples, while speech from speakers under stress in police interviews are intrinsic variations. Consequently, data in Phase-2 is not compatible with data in Phase-1 for domain-invariant information extraction. We split data in Phase-1 at a ratio of 7:1:2 for training, enrollment, and test data. The duration of each speech record is segmented into 30 seconds. The number of test trials in each set is 47684, 47684, 24564, 47778 for the clean, far-field, LENA-booth, and field-recording, respectively. For sound booth data, speech data with mics close to the speaker (10 inches & 0.8 feet) are viewed as clean in the target domain for first-step adaptation, and speech data from the remaining two distant mics are viewed as far-field data. Data captured by the LENA body-worn unit was used to explore channel mismatch influence.

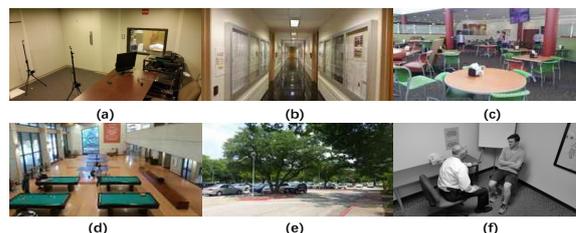


Figure 2: Forensic recording environments - (a) sound booth, (b-e) field recording, (f) police interview.

3.2. i-vector system configuration

Gender independent i-vector extractors were trained on the VoxCeleb 1 and 2 to produce the 400-dimensional i-vectors. 20-dimensional MFCCs were augmented with their delta and double-delta coefficients, composing 60-dimensional MFCC feature vectors. MLLR was used for adapting the basic i-vector model to the clean data.

3.3. CNN-based system configuration

A 30-dimensional Fbank was extracted with a frame length of 25 ms and frame shift of 10 ms as the acoustic feature for each utterance. Kaldi energy-based VAD was also applied to remove the silent segments in each utterance. Then the raw acoustic feature was input into a widely-used ResNet-34 architecture with Angular-Softmax [29]. The corresponding channel sizes for conv1, layer1, layer2, layer3, and layer4 were 16, 16, 32, 64, and 128 respectively. The number of dictionary components in the LDE layer was 64. Adam optimizer was employed for the network training with betas of (0.9, 0.98), weight decay of 1e-4, and the AMSGrad variant. The learning rate was adjusted with the warm-up scheduling named “Noam” in [30].

In the fine-tuning stage, each input data was cropped to 1200 frames, which approximated the mean length of an utterance after VAD. The parameters in layer 4 and the LDE layer were updated, and parameters for all other layers were frozen.

3.4. Cross-domain adaptation configuration

We did cross-domain adaptive training for layer 4 in ResNet, LDE layer, and the cross-domain adaptation block via back propagation. The cross-domain subnet of each domain was composed of four fully-connected layers: (4096,1024), (1024,1024), (1024,256), and (256,256); the first two and last two layers were applied for minimizing discrepancy and distribution alignment, respectively. Since the cross-domain adaptation block was trained from scratch, we set its learning rate to be 10x larger than that of the layer 4 in ResNet and LDE layers. The learning rate of the cross-domain adaptation block was adjusted using the formula,

$$\eta_p = \frac{\eta_0}{(1 + \alpha p)^\beta} \quad (6)$$

Where $\eta_0 = 0.01$, $\alpha = 10$, $\beta = 0.75$ and p linearly changes from 0 to 1 corresponding to the training steps. Since there exist no parameter-wise differences between each subnet in the early adaptation stage, $Loss_{cmd}$, $Loss_{dis}$ may result in noisy activations. To stabilize parameter sensitivity, a progress strategy [31] is used for $Loss_{total}$:

$$\mu = \frac{2}{\exp(-\theta p)} - 1 \quad (7)$$

Where $\theta = 10$ makes μ change from 0 to 1.

4. Results and Discussion

We compare a GMM-based i-vector system with a system based on neural networks for the speaker verification task. Results of the pre-trained model, the fine-tuning step, and the cross-domain adaptation (CDA) process are presented respectively for the two systems. In the cross-domain setup, there are channel, reverberation, and noise mismatches between datasets in various domains, which can be used to analyze how extrinsic variation in multiple acoustic environments impacts the speaker verification performance. Practically, it is very inefficient and computational expensive to maintain multiple models for different domains. The cross-domain adaptation method can help us train a single model to keep an excellent performance on the source-domain clean set and improve the verification result of the fine-tuned model on each target domain by a large margin. Table 2 shows systems' testing results in terms of equal error rate (EER) for each set. Due to the small absolute values shown in each set, we use a relative decrease (RD%) of EER to evaluate system improvement in each step. MLLR and fine-tuning are used by i-vector and CNN-based systems respectively in the clean set adaptation step.

Before any adaptation is applied, embeddings extracted from the raw common feature extractor are employed for speaker verification. Two systems show good results with low EER in general. Testing results for diverse domains vary a lot for two systems. The noisy data collected by the LENA unit in public environments shows the worst testing performance. Clean data and LENA-booth data are both collected in the sound booth, where the former one is recorded by microphones and achieves a better result than the latter recorded with the LENA unit. This indicates that the noise and channel mismatches considerably affect the speaker verification performance. The far-field condition generally has less of an influence on speaker verification. Far-field data in our corpus is recorded by mics in a

sound booth, where the voice is less likely to be distorted by the given distance as it travels in the wild.

Table 2: Equal error rate of systems in various domains

	Clean	LENA-booth	Far-field	Noisy
① Raw i-vector	0.58%	0.83%	0.69%	2.73%
② i-vector MLLR	0.22%	0.73%	0.34%	2.37%
① → ② RD	62.07% ↓	12.05% ↓	50.72% ↓	13.19% ↓
③ i-vector CDA	0.21%	0.63%	0.23%	2.28%
② → ③ RD	4.55% ↓	13.70% ↓	32.35% ↓	3.80% ↓
④ Raw CNN	1.15%	1.83%	1.03%	2.43%
⑤ CNN fine-tune	0.113%	1.49%	0.20%	2.06%
④ → ⑤ RD	90.17% ↓	18.58% ↓	80.58% ↓	15.23% ↓
⑥ CNN CDA	0.108%	1.01%	0.16%	2.02%
⑤ → ⑥ RD	4.42% ↓	32.21% ↓	20.00% ↓	1.94% ↓

For the clean data adaptation, the i-vector model employs MLLR to accomplish adaptation, and the CNN-based model is fine-tuned with clean data. Fine-tuning shows a more significant improvement than MLLR, especially in the clean and far-field sets. Furthermore, the system verification performance gains a limit improvement for LENA-booth and noisy set in each system. The i-vector system with MLLR and the fine-tuned CNN-based system achieved effective improvements with an averaged EER RD of +34.5% and +51.1% respectively.

Based on models processed by the first-step adaptation, the cross-domain adapted i-vector system obtains a better result with a lower EER of 0.63% in the LENA-booth set, and the CNN-based system with cross-domain adaptation achieves better results in the other three conditions, which are 0.108%, 0.16%, and 2.02% in the form of EER for the clean, LENA-booth, and far-field respectively. By means of generating nearly identical distributions for all statistics in the CNN CDA, the model has a good performance on the clean set (+4.42% RD), while it significantly improves the results on domains of LENA-booth and far-field for +32.21% and +20% respectively. However, the model of CNN CDA achieves limited improvement in the noisy domain. Cross-domain adaptation contributes to speaker verification performance improvement in both systems, and is more effective for the CNN-based system. The average RD of EER for the i-vector system is +13.6%, and that for the CNN-based system is +14.6%. Therefore, the proposed cross-domain adaptation methods have been shown to be promising/effective for new environment mismatch data scenarios.

5. Conclusions

In this study, we explored a cross-domain method to adapt i-vector and CNN-based systems to data from new various acoustic domains. The adaptation included two steps of the clean set adaptation and cross-domain adaptation. In the clean set adaptation, an i-vector system was used with an MLLR method with clean data, and the CNN-based system employed clean data to fine-tune the model, which achieved the best result by applying fine-tuning to layer 4 in ResNet and LDE layer. Based on the first-stage adaptation, the cross-domain adaptation approach was employed to learn domain-invariant information so that the model could be greatly generalized to the other acoustic domains without degrading verification performance on the clean set. In conclusion, diverse or unknown acoustic environments which are common in field data for forensics were shown to affect speaker verification system performance, and the proposed cross-domain adaptation method was able to contribute to verification performance improvement for multiple domains.

6. References

- [1] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.
- [2] J. H. Wigmore, "New mode of identifying criminals, a," 1926.
- [3] N. Dehak, P. J. 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, 2010.
- [4] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Twelfth annual conference of the international speech communication association*, 2011.
- [5] M. Li and W. Liu, "Speaker verification and spoken language identification using a generalized i-vector framework with phonetic tokenizations and tandem features," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [6] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.
- [7] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 165–170.
- [8] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.
- [9] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," *arXiv preprint arXiv:1804.05160*, 2018.
- [10] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1695–1699.
- [11] S. Ben-David, T. Lu, T. Luu, and D. Pál, "Impossibility theorems for domain adaptation," in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 129–136.
- [12] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: Learning bounds and algorithms," *arXiv preprint arXiv:0902.3430*, 2009.
- [13] Y. Xi, P. Li, Y. Song, Y. Jiang, and L. Dai, "Speaker to emotion: Domain adaptation for speech emotion recognition with residual adapters," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 513–518.
- [14] W. Lin, M.-W. Mak, N. Li, D. Su, and D. Yu, "Multi-level deep neural network adaptation for speaker verification using mmd and consistency regularization," 2020.
- [15] W. Xia and J. H. Hansen, "Speaker recognition with nonlinear distortion: Clipping analysis and impact," *Proc. Interspeech 2018*, pp. 746–750, 2018.
- [16] W. Xia, J. Huang, and J. H. Hansen, "Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5816–5820.
- [17] K. Veselý, S. Watanabe, K. Žmolíková, M. Karafiát, L. Burget, and J. H. Čermocký, "Sequence summarizing neural network for speaker adaptation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5315–5319.
- [18] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," *arXiv preprint arXiv:1702.05374*, 2017.
- [19] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.
- [20] Y. Zhu, F. Zhuang, and D. Wang, "Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 5989–5996.
- [21] R. Xu, Z. Chen, W. Zuo, J. Yan, and L. Lin, "Deep cocktail network: Multi-source unsupervised domain adaptation with category shift," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3964–3973.
- [22] W.-W. Lin, M.-W. Mak, L. Li, and J.-T. Chien, "Reducing domain mismatch by maximum mean discrepancy based autoencoders," in *Odyssey*, 2018, pp. 162–167.
- [23] 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.
- [24] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [26] J. Ganitkevitch, "Speaker adaptation using maximum likelihood linear regression," in *Rheinish-Westfälische Technische Hochschule Aachen, the course of Automatic Speech Recognition*, www-i6.informatik.rwth-aachen.de/web/Teaching/Seminars/SS05/ASR/Juri_Ganitkevitch_Ausarbeitung.pdf. Citeseer, 2005.
- [27] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.
- [28] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.
- [29] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.
- [30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [31] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014.