CROSS-LINGUAL TEXT-INDEPENDENT SPEAKER VERIFICATION USING UNSUPERVISED ADVERSARIAL DISCRIMINATIVE DOMAIN ADAPTATION

Wei Xia[1], Jing Huang[2], John H.L. Hansen[1]

[1] Center for Robust Speech Systems, UT-Dallas, TX, USA
[2] JD AI Research, Mountain View, CA, USA

ABSTRACT

Speaker verification systems often degrade significantly when there is a language mismatch between training and testing data. Being able to improve cross-lingual speaker verification system using unlabeled data can greatly increase the robustness of the system and reduce human labeling costs. In this study, we introduce an unsupervised Adversarial Discriminative Domain Adaptation (ADDA) method to effectively learn an asymmetric mapping that adapts the target domain encoder to the source domain, where the target domain and source domain are speech data from different languages. ADDA, together with a popular Domain Adversarial Training (DAT) approach, are evaluated on a cross-lingual speaker verification task: the training data is in English from NIST SRE04-08, Mixer 6 and Switchboard, and the test data is in Chinese from AISHELL-I. We show that with the ADDA adaptation, Equal Error Rate (EER) of the x-vector system decreases from 9.331% to 7.645%, relatively 18.07% reduction of EER, and 6.32% reduction from DAT as well. Further data analysis of ADDA adapted speaker embedding shows that the learned speaker embeddings can perform well on speaker classification for the target domain data, and are less dependent with respect to the shift in language.

Index Terms— Speaker Verification, Adversarial Training, Domain Adaptation, Speaker Representation

1. INTRODUCTION

Speaker verification (SV) offers a natural and flexible option for biometric authentication. The text-independent SV system, which does not require the fixed input voice content, is a flexible and challenging task. In real-world scenarios, however, speaker verification systems may degrade significantly when training on one language and test it on another. Language mismatch falls into two scenarios that include (i) the speaker verification system is trained on one language, but the enrollment and test data for speakers are in a second language, and (ii) the enrollment data is in one language, but the test data is in a second language. This study focused on the first scenario where the speaker model is trained on English data, but the enrollment and test materials for speakers are in a new language, Chinese. Since it is not desirable to re-train the speaker model on a new language, the challenge is to find an alternative solution which would allow such an existing system to maintain performance when enrollment and test speaker data are from a new language.

Recently, the speaker representation models have moved from the commonly used i-vector model [1, 2, 3], with a probabilistic linear discriminant (PLDA) back-end [4, 5] to a new paradigm: speaker embedding trained from deep neural networks. Various speaker embeddings based on different network architectures [6, 7] , attention mechanism [8, 9], loss functions [10, 11], noise robustness [12, 13], and training paradigms [14, 15] have been proposed and greatly improve the performance of speaker verification systems. Snyder et al. [6] recently proposed the x-vector model, which is based on a Time-Delay Deep Neural Network (TDNN) architecture that computes speaker embeddings from variable-length acoustic segments. This x-vector model has become very successful in various speaker recognition tasks. We use it as the baseline in this study.

However, models trained with these deep neural networks may not generalize well to other datasets in different domains. To alleviate the domain mismatch problem, we can use domain adaptation methods to reduce the domain shift. We can compensate the mismatch by estimating the compensation model [16, 17, 18, 19] using unlabeled data and source domain data. Adversarial adaptation methods [20, 21, 12, 22] were also applied to ensure that the network cannot distinguish the distributions of training and testing examples. Wang et al. [23] proposed an unsupervised approach based on Domain Adversarial Training (DAT) to address speaker recognition problem in domain mismatched conditions.

In this study, we introduce the unsupervised Adversarial Discriminative Domain Adaptation (ADDA) [24] approach. It was originally tested on image classification tasks. We adapt the ADDA approach to the cross-lingual unsupervised adaptation for text-independent speaker verification. Unsupervised adaptation without requiring target domain labels largely reduces labeling costs and utilizes a large amount of publicly available online data. Our approach only requires source and unlabeled target domain data to learn an asymmetric mapping that adapts the target domain feature encoder to the source domain. Furthermore, the ADDA uses separate encoders for the source and target domain without assuming that source and target domain data has a similar class distribution. We show that ADDA is more effective yet considerably simpler than other domain-adversarial methods: the source data is in English from NIST SRE04-08, Mixer 6 and Switchboard, and the target data is in Chinese from AISHELL-I. We show that with the ADDA adaptation, Equal Error Rate (EER) of the x-vector system decreases from 9.331% to 7.645%, relatively 18.07% reduction on EER. ADDA also has 12.54% relative reduction of EER compared to DAT.

In the following sections, we describe the ADDA approach and corresponding baseline systems in Section 2. We provide detailed explanations of our experiments in Section 3, as well as results and discussions in Section 4. Finally we conclude in Section 5 with future work.

1.1. Related work

A number of domain adaptation approaches have been proposed to alleviate the domain shift problem. For example, Wang et al. [23] apply the DAT technique to alleviate the i-vectors mismatch across different domains. They use a multi-task learning framework to jointly

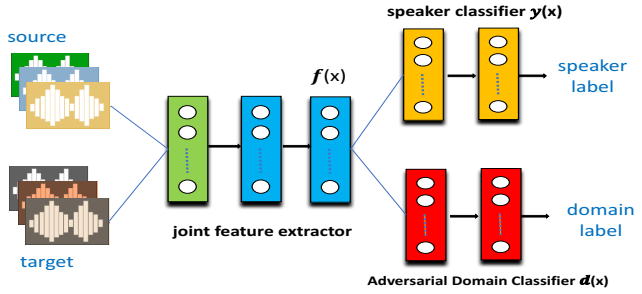


Fig. 1: Overview of the Domain Adversarial Training (DAT) framework. Adversarial domain classifier has a gradient reversal layer. Speaker classifier and domain classifier both take input from the joint feature extractor, are optimized to excel in their own tasks.

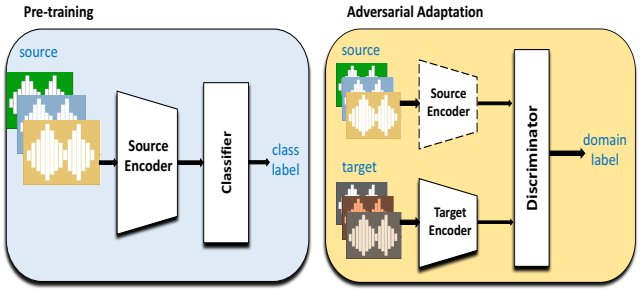


Fig. 2: Overview of the proposed Adversarial Discriminative Domain Adaptation (ADDA) approach. Source DNN encoder is fixed during the adversarial adaptation.

learn a shared feature extractor and two classifiers. With a gradient reversal layer in the domain classifier, the shared feature extractor can extract domain-invariant and speaker-discriminative features. In [16, 17], the authors proposed an Inter-Dataset Variability Compensation (IDVC) technique to remove the mismatch using Nuisance Attribute Projection (NAP). First, a subspace is computed representing all different data-sets and then NAP is used to remove that subspace as an i-Vector pre-processing step. All these work were on i-vectors for speaker verification, while our work is on the recently proposed x-vectors and shows very promising results.

2. SPEAKER VERIFICATION SYSTEMS

2.1. The X-vector system

We use a recently proposed successful speaker model called X-vector [6], to extract speaker representations, and a Probabilistic Linear Discriminant Analysis (PLDA) back-end to compare pairs of enrollment and test speaker embeddings. The X-vector model is based on a Time-Delay Deep Neural Network (TDNN) architecture that computes speaker embeddings from variable-length acoustic segments. The network consists of layers that operate on speech frames, a statistics pooling layer that aggregates over the frame-level representations, additional layers that operate at the segment-level, and finally a softmax output layer. The embeddings are extracted after the statistics pooling layers.

2.2. Cross-lingual adversarial training baseline

In order to address the cross-lingual speaker verification problem, we first implement a Domain Adversarial Neural Network (DANN) [23] using Domain Adversarial Training (DAT) [20] to transfer speaker information from labeled English data to another language where only unlabeled data exists, for example, Chinese. DANN in Fig. 1 is a Y-shaped network with two discriminative branches: a speaker recognizer and an adversarial language classifier. Both branches take input from a shared feature extractor that aims to learn hidden representations that capture the underlying information of the speaker and are independent of languages.

We can implement the language independent speaker verification system assuming that DANN can learn features that perform well on speaker classification for the source and target language data, are independent with respect to the shift in language. This can be done by minimizing the speaker classification loss and maximizing the domain classification loss with a gradient reversal layer. DANN mainly has two components: 1) a speaker recognizer y for the source data; 2) an adversarial language classifier d that predicts a scalar indicating whether the input speech is from the source language or the target language. The two classifiers take input from the shared feature extractor f, which operates on the average of the speaker embeddings. The loss function of DANN is a multi-task loss which combines the loss of the speaker classifier and the domain classifier with a weight λ. Training DANN consists in optimizing,

$$\mathbb{E}(\theta_f,\theta_y,\theta_d) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_y^i(\theta_f,\theta_y) - \lambda[\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_d^i(\theta_f,\theta_d) + \frac{1}{n'}\sum_{i=n+1}^{N}\mathcal{L}_d^i(\theta_f,\theta_d)], \quad (1)$$

where $\theta_f, \theta_y, \theta_d$ are parameters of the joint feature extractor and two classifiers, and L_y, L_d are the prediction and the domain loss functions. n and n' are the number of samples of the source and target domain data respectively. We can optimize this loss function using stochastic gradient descent to get the parameters, Using this DAT approach, we are able to minimize the divergence between the source and target feature distributions. Therefore, the learned embeddings are less dependent on the shift in language.

2.3. Adversarial discriminative domain adaptation

Different from the DAT method which applies a gradient reversal layer to confuse the domain classifier, we apply the Adversarial Discriminative Domain Adaptation (ADDA) approach to directly learn an asymmetric mapping, in which we modify the target model in order to match the source distribution. A summary of this entire training process is provided in Fig. 2. Unlike the DAT method which uses a shared feature encoder, our proposed ADDA approach uses separate encoders for the source and target domain data. When there is a significant domain shift, the DAT method may not work well since it inherently assumes that source and target domain data has a similar class distribution.

We define input samples $\mathbf{x} \in \mathbf{X}$ with data labels $y \in \mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ are input space and output space, respectively. In our speaker verification experiments, $\mathbf{x}$ and y are x-vectors and speaker labels. The probabilistic distribution $\mathcal{D}(\mathbf{x}, y)$, however, might be different between training and evaluation dataset due to various domain mismatch such as language mismatch. We denote $\mathcal{S}(\mathbf{x}, y)$ and $\mathcal{T}(\mathbf{x}, y)$ as source domain and target domain distribution respectively. Our goal is to minimize the distance between the empirical

source and target mapping distributions. We firstly learn a source mapping M_s, along with a source classifier $\mathcal{C}$, and then learn to map the target domain encoder to the source domain.

We train the source classification model using a standard cross entropy loss defined below,

$$\min_{M_s,\mathcal{C}} \mathcal{L}_{cls}(\mathbf{X}_s, Y_s) = - \mathbb{E}_{(\mathbf{x}_s,y_s)\sim(\mathbf{X}_s,Y_s)} \sum_{k=1}^{K} \mathbb{1}_{[k=y_s]} \log \mathcal{C}(M_s(\mathbf{x}_s)), \tag{2}$$

In order to minimize the source and target representation distances, we use a domain discriminator $\mathcal{D}$ to classify whether a data point is drawn from the source or the target domain. We optimize $\mathcal{D}$ using an adversarial loss $\mathcal{L}_{adv_D}(X_s, X_t, M_s, M_t)$, defined below:

$$\min_{D} \mathcal{L}_{adv_D}(\mathbf{X}_s, \mathbf{X}_t, M_s, M_t) = - \mathbb{E}_{\mathbf{x}_s\sim\mathbf{X}_s}[\log \mathcal{D}(M_s(\mathbf{x}_s))] - \mathbb{E}_{\mathbf{x}_t\sim\mathbf{X}_t}[\log(1 - \mathcal{D}(M_t(\mathbf{x}_t)))], \tag{3}$$

The DAT method uses a gradient reversal layer [20] to learn the mapping by maximizing the discriminator loss directly, where its adversarial loss $\mathcal{L}_{adv_M} = -\mathcal{L}_{adv_D}$. Different from DAT, in order to train the mapping, we use the loss function $\mathcal{L}_{adv_M}$ defined below. This objective has the same fixed-point properties as the minimax loss but provides stronger gradients to the target mapping.

$$\min_{M_s,M_t} \mathcal{L}_{adv_M}(\mathbf{X}_s, \mathbf{X}_t, \mathcal{D}) = -\mathbb{E}_{\mathbf{x}_t\sim\mathbf{X}_t}[\log \mathcal{D}(M_t(\mathbf{x}_t))]. \tag{4}$$

We can optimize this objective function in two steps. First, we need to train a discriminative source classification model, we choose to use a three-layer Deep Neural Network (DNN) and the input features are x-vectors. We start optimizing classification loss $\mathcal{L}_{cls}$ over source domain mapping function M_s and classifier $\mathcal{C}$ by training with the labeled source English data, $\mathbf{X}_s$ and $\mathbf{Y}_s$. Because we make M_s fixed while learning M_t, we can then optimize $\mathcal{L}_{adv_D}$ and $\mathcal{L}_{adv_M}$ without revisiting the first objective term.

Through this unsupervised adversarial discriminative domain adaptation approach, we can adapt the target encoder to the source domain. In the next section, we will present promising results on cross-lingual text-independent speaker verification tasks using ADDA.

3. EXPERIMENTAL SETUP

3.1. English Corpora

We use Speaker Recognition Evaluation (SRE) 04-08, Mixer 6, and Switchboard (SWBD) to train the x-vector model. SRE corpus is part of the Mixer 6 project, which was designed to support the development of robust speaker recognition technology by providing carefully collected speech across numerous microphones. Switchboard is a collection of about two-sided telephone conversations among thousands of speakers from all areas of the United States.

3.2. Chinese Corpora

AISHELL-1 [25] is a subset of the AISHELL-ASR0009 corpus, which is a 500 hours multi-channel mandarin speech corpus designed for various speech/speaker processing tasks. Speech utterances are recorded at 44.1kHz via microphones, 16kHz via Android phones and 16kHz via iPhones.

There are 360 participants in the recording, and speakers' gender, accent, age, and birth-place are recorded as meta-data. About 80 percent of the speakers are from age 16 to 25. Most speakers come from the Northern area of China. The entire corpus includes training and test sets, without speaker overlap. Though the training data provides speaker labels, we do not use any speaker label information of the training data or include it in training our x-vector model. We only use it for unsupervised domain adaptation. We call it AISHELL unlabeled training set.

The training set contains 120,098 utterances from 340 speakers; Test set contains 7,176 utterances from 20 speakers. For each speaker, around 360 utterances (about 26 minutes of speech in total) are released. In order to test our proposed unsupervised ADDA approach, we don't use any speaker labels of the training data. We train our x-vector based speaker model on the SRE04-08, Mixer 6, and switchboard dataset, and evaluate on the Chinese AISHELL test 143520 trials.

3.3. Evaluation setup

We use SRE04-08, Mixer6 and Switchboard data to train the TDNN based x-vector model. We follow the Kaldi SRE16 recipe to augment the training data by adding noises and reverberations. We use an energy based VAD and the raw feature to train the model are 23-dimensional MFCCs. Having established the x-vector system using English data, we now try to address the challenge of evaluation enrollment and test speakers for a mismatched language, Chinese. To accomplish this, A set of unlabeled data for the new language is needed. We use the target domain AISHELL unlabeled training data. We extract x-vectors on source domain SRE and SWBD data and target domain AISHELL unlabeled data to train the adaptation network.

We train the Adversarial Domain Adaptation Network (ADAN) in two steps. First, we train a DNN encoder and classifier on SRE and SWBD x-vectors. Next, we use the pre-trained source model as an initialization for the target DNN encoder and perform adversarial adaptation to learn a target domain mapping on the AISEHLL unlabeled x-vectors.

During testing, we use AISHELL evaluation set enrollment x-vectors and test x-vectors as the input to the ADDA, and extract the new vectors $\hat{x}_e$, $\hat{x}_t$ using the trained target encoder of ADDA. Adapted embeddings $\hat{x}_e$, $\hat{x}_t$ are therefore expected to be domain-invariant and speaker discriminative representations which stay in the same subspace. We apply mean and length normalization on the adapted embeddings. For the back-end, we train a Probabilistic Linear Discriminant Analysis (PLDA) model on combined SRE clean and noise augmented data, and compute log-likelihood ratio scores of enrollment and test trials. We also perform unsupervised PLDA adaptation using Kaldi to utilize the AISHELL unlabeled data.

3.4. Model configuration

For this experiment, our base architecture is a three-layer Deep Neural Network which is fine-tuned on the source domain for 100 epochs using a batch size of 128. When training ADDA, the adversarial discriminator consists of three additional fully connected layers: 2 hidden layers and an adversarial discriminator output. With the exception of the output, these additionally fully connected layers use a ReLU activation function. ADDA target encoder training then proceeds for another 100 epochs with a batch size of 128. For the DAT training, the shared feature encoder is a three-layer DNN. We use an Adam optimizer with a learning rate 10^{-4}. The speaker classi-

fier and the language classifier are two-layer DNNs. To confuse the language domain classifier, the language classifier has a gradient reversal layer. We use a multi-task loss with equal weights to combine the two cross entropy losses.

4. RESULTS AND DISCUSSIONS

4.1. Results

In this section, we show experimental results using x-vector, x-vector with DAT and x-vector with ADDA training with and without PLDA adaptations in Table 1. We use Linear Discriminant Analysis (LDA) to reduce all three embeddings to 256 dimension for comparison. Also, we concatenate the DAT embedding with the x-vector since we find it always performs better than a single DAT embedding. From Table 1, we observe that our proposed method, ADDA, greatly improves Equal Error Rate (EER) on AISHELL test trials. After ADDA adaptation, EER of the x-vector system decreases from 9.331% to 7.645%, relatively 18.07%. The ADDA approach also achieves relatively 12.54% improvement compared with the concatenated x-vector and DAT embedding. The major reason that ADDA works better might be that it uses an adversarial discriminator to adapt the target encoder to the source domain. Also, by initializing the target representation space with the pre-trained source model, we can effectively learn the asymmetric mapping function.

Fig. 3 shows the Detection Error Trade-off (DET) curve of our speaker recognition system at three different settings without PLDA adaptation. From the figure, we see after DAT or ADDA adaptation, the overall speaker verification system performance improves significantly compared with the x-vector system. Further, both False Positive Rate (FPR) and False Negative Rate (FNR) of the ADDA embedding system reduce by a large margin compared with the x-vector+DAT embedding system. It indicates that ADDA embedding has more invariance to language shift.

	EER(%)	MinDCF
x-vector	9.331	0.7755
x-vector + DAT	8.741	0.7475
ADDA embedding	7.645	0.7257
x-vector + PLDA adaptation	9.162	0.7095
x-vector + DAT + PLDA adaptation	7.799	0.6989
ADDA embedding + PLDA adaptation	7.504	0.7062

Table 1: Speaker verification results using different models with a PLDA back-end.

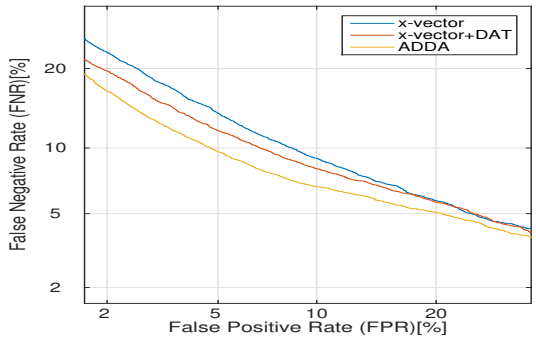
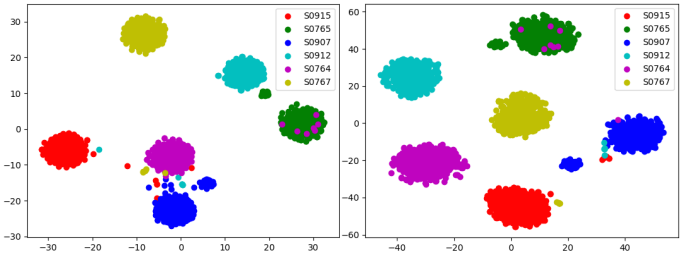


Fig. 3: DET curve results with different speaker representations.



(a) x-vector (b) ADDA embedding

Fig. 4: Visualizations of x-vector and ADDA speaker embeddings using t-SNE

4.2. Visualization of speaker embeddings

To investigate the effect ADDA has on speaker verification, we further assess the quality of the learned speaker features, using t-SNE [26], we plot embeddings after LDA from same K speakers of the AISHELL test set. The results are presented in Fig. 4. Fig. 4 (a) is the visualization of x-vectors, and Fig. 4 (b) is the visualization of ADDA embedding. It can be seen that the ADDA embeddings have more discriminative ability to separate different speakers. However, for x-vectors, we observe that some utterances from different speakers are grouped together and not well separated in the embedding space. Also, for speaker "0764", it is difficult to separate it from speaker "0765" using both methods. It is probably because these two speakers have very similar speaker information.

4.3. Clustering analysis

In order to quantitatively analyze the quality of adapted speaker representations, we also perform clustering on the adapted embeddings. Since t-SNE cannot maintain distance information, which is necessary to apply most clustering algorithms, we perform K-means clustering after LDA transformed x-vectors and ADDA embeddings. Given the knowledge of the ground truth speaker labels, we compute the Normalized Mutual Information (NMI) [27] of the K-means clustering assignment. NMI is a metric that measures the agreement of the ground truth labels and the clustering results. The NMI score of x-vectors is 0.787, and the NMI score of ADDA embeddings is 0.802, relatively 1.9% higher. This result is consistent with the visualization using t-SNE. Therefore, we can conclude that with the ADDA adaptation, we can learn more speaker discriminative and language independent speaker embeddings.

5. CONCLUSIONS AND FUTURE WORK

We presented a discriminative adversarial unsupervised adaptation method in this paper. By exploiting how to alleviate the domain mismatch problem in an English-Chinese cross-lingual speaker verification task, we showed that our proposed unsupervised ADDA approach can perform well on speaker classification for the target domain data. Additional data analysis indicated that the representations learned via ADDA can be well separated and are less dependent with respect to the shift in language.

In the future, we would like to investigate the influence of phonetic content on cross-lingual text-independent speaker verification. We intend to use a phoneme decoder to analyze the linguistic factor of speaker models.

6. REFERENCES

[1] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[2] Pavel Matějka, Ondřej Glembek, Fabio Castaldo, Md Jahangir Alam, Oldřich Plchot, Patrick Kenny, Lukáš Burget, and Jan Černocky, "Full-covariance ubm and heavy-tailed plda in i-vector speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2011, pp. 4828–4831.

[3] John HL Hansen and Taufiq Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[4] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors.," in *Odyssey*, 2010, p. 14.

[5] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision (ICCV), IEEE International Conference on*, 2007, pp. 1–8.

[6] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2018.

[7] Daniel Michelsanti and Zheng-Hua Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *INTERSPEECH*, 2017.

[8] FA Rezaur rahman Chowdhury, Quan Wang, Ignacio Lopez Moreno, and Li Wan, "Attention-based models for text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2018, pp. 5359–5363.

[9] Shi-Xiong Zhang, Zhuo Chen, Yong Zhao, Jinyu Li, and Yifan Gong, "End-to-end attention based text-dependent speaker verification," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 171–178.

[10] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2018, pp. 4879–4883.

[11] Chunlei Zhang, Kazuhito Koishida, and John HL Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 9, pp. 1633–1644, 2018.

[12] Hong Yu, Zheng-Hua Tan, Zhanyu Ma, and Jun Guo, "Adversarial network bottleneck features for noise robust speaker verification," in *INTERSPEECH*, 2017, pp. 1492–1496.

[13] Wei Xia and John HL Hansen, "Speaker recognition with nonlinear distortion: Clipping analysis and impact," in *INTERSPEECH*, 2018, pp. 746–750.

[14] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer, "End-to-end text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2016, pp. 5115–5119.

[15] Hee-soo Heo, Jee-weon Jung, IL-ho Yang, Sung-hyun Yoon, and Ha-jin Yu, "Joint Training of Expanded End-to-End DNN for Text-Dependent Speaker Verification," in *INTERSPEECH*, 2017, pp. 1532–1536.

[16] Hagai Aronowitz, "Inter dataset variability compensation for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2014, pp. 4002–4006.

[17] Ahilan Kanagasundaram, David Dean, and Sridha Sridharan, "Improving out-domain plda speaker verification using unsupervised inter-dataset variability compensation approach," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2015, pp. 4654–4658.

[18] Abhinav Misra and John HL Hansen, "Maximum-likelihood linear transformation for unsupervised domain adaptation in speaker verification," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 9, pp. 1549–1558, 2018.

[19] Abhinav Misra and John HL Hansen, "Modelling and compensation for language mismatch in speaker verification," *Speech Communication*, vol. 96, pp. 58–66, 2018.

[20] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[21] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang, "Multi-adversarial domain adaptation," in *AAAI Conference on Artificial Intelligence*, 2018.

[22] Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger, "Adversarial deep averaging networks for cross-lingual sentiment classification," *arXiv preprint arXiv:1606.01614*, 2016.

[23] Qing Wang, Wei Rao, Sining Sun, Leib Xie, Eng Siong Chng, and Haizhou Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2018, pp. 4889–4893.

[24] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," in *Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 1, p. 4.

[25] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5.

[26] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[27] Nguyen Xuan Vinh, Julien Epps, and James Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *Journal of Machine Learning Research*, vol. 11, no. Oct, pp. 2837–2854, 2010.