ADVERSARIAL SPEAKER VERIFICATION

Zhong Meng, Yong Zhao, Jinyu Li, Yifan Gong

Microsoft Corporation, Redmond, WA, USA

{zhme, yonzhao, jinyl, ygong}@microsoft.com

ABSTRACT

The use of deep networks to extract embeddings for speaker recognition has proven successfully. However, such embeddings are susceptible to performance degradation due to the mismatches among the training, enrollment, and test conditions. In this work, we propose an adversarial speaker verification (ASV) scheme to learn the *condition-invariant* deep embedding via adversarial multi-task training. In ASV, a speaker classification network and a condition identification network are jointly optimized to minimize the speaker classification loss and simultaneously mini-maximize the condition loss. The target labels of the condition network can be categorical (environment types) and continuous (SNR values). We further propose multi-factorial ASV to simultaneously suppress multiple factors that constitute the condition variability. Evaluated on a Microsoft Cortana text-dependent speaker verification task, the ASV achieves 8.8% and 14.5% relative improvements in equal error rates (EER) for known and unknown conditions, respectively.

Index Terms— adversarial learning, speaker verification, deep neural network

1. INTRODUCTION

Speaker verification (SV) is the task of authenticating the claimed identity of an utterance, based on a speaker's known recordings. An SV system is text-dependent (TD) if the content of the test utterance is a fixed or prompted text phrase and is text-independent if the test utterance is unconstrained speech. Over the years, the i-vector followed by probabilistic linear discriminant analysis (PLDA) [1] has been the dominant approach for SV.

Recently, with the advent of deep learning, deep embeddings learned from a deep network have achieved great success and have become the state-of-the-art in speaker recognition. In [2, 3], a deep neural network (DNN) is trained to discriminate between speakers and the outputs from a hidden layer are averaged over frames in an utterances as the embeddings to represent the enrolled speakers and test utterances. In [4], pooling over hidden layer activations in the training time is introduced for speaker classification. In [5, 4], long short-term memory (LSTM) model is optimized using an end-to-end criterion. A triplet loss is further introduced in [6] to learn speaker embeddings. In [7], the attention mechanism was introduced to determine the weights for combining the frame-level features, instead of just equally averaging all the frames. Although these methods have greatly advanced the performance of SV, the performance degradation due to mismatched conditions is still a significant barrier for deploying speaker recognition technologies. The major adverse conditions causing mismatches are different channel and noise environments, and the mismatches exist in two folds, not only between the training and test conditions but also the enrollment and test conditions [8, 9].

Recently, adversarial learning [10] has achieved great success in estimating generative models. In speech area, it has been applied to speech enhancement [11, 12, 13], voice conversion [14], acoustic model adaptation [15, 16, 17], noise-robust [18, 19], speaker-invariant [20, 21] automatic speech recognition, speaker model adaptation [22] and speech enhancement [23, 11, 24] using gradient reversal layer (GRL) [25]. In these works, adversarial learning is used to learn an intermediate representation in a DNN that is invariant to the shift among different conditions (e.g., environments, speakers, SNRs, etc.). To benefit from this, in this work, we propose adversarial speaker verification (ASV) to suppress the effects of condition variability in speaker modeling for robust SV. In ASV, a speaker classification network and a condition identification network are jointly trained to minimize the speaker classification loss and to mini-maximize the condition loss through adversarial multi-task learning. The target labels of the condition network can be categorical (environment types) and continuous (SNR values) With ASV, speaker-discriminative and condition-invariant deep embeddings can be extracted for both enrollment and test speech.

Adversarial learning has been applied to speaker modeling in [22]. The proposed ASV is significantly different from it in that: (1) ASV suppresses two kinds of condition variability in speaker modeling using different methods, whereas [22] aims at adapting a well trained speaker model to the unlabeled target domain data, other than condition robustness of the model. (2) The proposed system train a network directly takes acoustic features as the input while the input of the system in [22] are utterance-level i-vectors which requires additional computational time and resources.

We perform ASV experiments to normalize environment and SNR variabilities on a Microsoft Cortana TD-SV task. ASV 8.8% and 14.5% relative improvements over the baseline for known and unknown conditions, respectively. The twice as much relative gain on *unknown* conditions as on known conditions shows the significant advantage of ASV in normalizing out the unexpected condition factors in speaker modeling and the strong capability of generalizing to unknown conditions.

2. DEEP EMBEDDING FOR SPEAKER VERIFICATION

Deep embedding has been widely used for speaker verification. In the training stage, we train a background DNN to disriminate between speakers of the training set. We view the hidden layers of the background DNN as a feature extractor network M_f with parameters θ_f that maps input speech frames $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}, \mathbf{x}_t \in \mathbb{R}^{r_x}, t = 1, \ldots, T$ from training set to intermediate deep hidden features $\mathbf{F} = \{\mathbf{f}_1, \ldots, \mathbf{f}_T\}, \mathbf{f}_t \in \mathbb{R}^{r_f}$ and the upper layers of the background DNN as a speaker classifier network M_y with parameters θ_y that maps the deep features $\mathbf{F}$ to the speaker posteriors

$p(a|\mathbf{f}_t;\theta_y), a \in \mathbb{A}$ as follows:

$$\mathbf{f}_t = M_f(\mathbf{x}_t) \quad (1)$$

$$p(a|\mathbf{f}_t;\theta_y) = M_y(\mathbf{f}_t), \quad (2)$$

where $\mathbb{A}$ is the set of all speakers in the training set. θ_f and θ_y are optimized by minimizing the cross-entropy loss of speaker classification as below

$$\begin{aligned}\mathcal{L}_{\text{speaker}}(\theta_f,\theta_y) &= -\frac{1}{T}\sum_{t=1}^{T}\log p(y_t|\mathbf{x}_t;\theta_f,\theta_y)\\ &= -\frac{1}{T}\sum_{t=1}^{T}\sum_{a\in\mathbb{A}}\mathbb{1}[a=y_t]\log M_y(M_f(\mathbf{x}_t)), \quad (3)\end{aligned}$$

where $\mathbf{Y} = \{y_1,\ldots,y_T\}, y_t \in \mathbb{A}$ is a sequence of speaker labels aligned with $\mathbf{X}$ and $\mathbb{1}[\cdot]$ is the indicator function which equals to 1 if the condition in the squared bracket is satisfied and 0 otherwise. Once the speaker classifier is trained, we compute deep hidden features for given utterances and average them to form a compact deep embedding for that speaker. Note that the speakers in the enrollment and training sets do not overlap. During evaluation, the cosine distance between the embeddings of the test utterance and the claimed speaker is computed and compared with a threshold to make the verification decision.

3. ADVERSARIAL SPEAKER VERIFICATION

With the background DNN for speaker classification, we are able to learn speaker-discriminative embeddings. However, in some scenarios, the speakers are enrolled in different conditions (i.e., environments, SNR values, etc.) from those in the training set and the test utterances are recorded in different conditions from the training and enrollment sets. Under these scenarios, the embeddings of the enrolled speakers and the test utterances are mismatched and may lead to degraded speaker verification performance because the new conditions for enrollment and testing are *unknown* to the background DNN trained on the training set. We propose ASV to reduce the effects of condition variability on the background DNN, i.e., the high variances of hidden and output unit distributions of the network caused by the inherent inter-condition variability in the speech signal. With such a background DNN, we can extract *condition-invariant* deep embeddings for enrolled speakers and test utterances to perform robust SV.

3.1. Adversarial Training of Background DNN

With ASV, our goal is to learn a *condition-invariant* and *speaker-discriminative* deep hidden feature in the background DNN through adversarial multi-task learning such that a noise-robust deep embedding can be obtained from these deep features for a enrolled speaker or a test utterance. In real application, conditions that affect the speaker modeling can be represented by either a categorical or a continuous variable. For example, different kinds of environments can be characterized by a categorical variable while the noise levels, i.e., the SNR values, of the input speech frames in the same or different environments are represented by a continuous variable. We propose different methods to deal with these two types of condition variability as follows.

3.1.1. Categorical Condition Classification Loss

As shown in Fig. 1, to address the conditions that are characterized as a categorical variable, we introduce an additional condition classification network M_c which predicts the condition posteriors $p(b|\mathbf{f}_t;\theta_f), b \in \mathbb{B}$ given the deep features $\mathbf{F}$ from the training set as follows:

$$M_c(\mathbf{f}_t) = p(b|\mathbf{f}_t;\theta_c) = p(b|\mathbf{x}_t;\theta_f,\theta_c), \quad (4)$$

where $\mathbb{B}$ is the set of all conditions in the training set. With a sequence of condition labels $\mathbf{C} = \{c_1,\ldots,c_T\}$ that is aligned with $\mathbf{X}$, we are able to compute the condition classification loss through cross-entropy as follows

$$\begin{aligned}\mathcal{L}_{\text{condition}}(\theta_f,\theta_c) &= -\frac{1}{T}\sum_{t=1}^{T}\log p(c_t|\mathbf{f}_t;\theta_c)\\ &= -\frac{1}{T}\sum_{t=1}^{T}\sum_{b\in\mathbb{B}}\mathbb{1}[b=c_t]\log M_c(M_f(\mathbf{x}_t)). \quad (5)\end{aligned}$$

3.1.2. Continuous Condition Regression Loss

The speaker modeling is also affected by conditions that are represented by a continuous variable. Different from a categorical variable such as the environment type, continuous conditions are real numbers or real vectors that can hardly be evaluated by a classification network of which the outputs represent the posteriors of discrete classes. Therefore, as shown in Fig. 1, we introduce an additional condition regression network M_c instead to predict the frame-level condition value (e.g., SNR value) $\hat{\mathbf{c}}_t \in \mathbb{R}^{r_c}$ given the deep features $\mathbf{F}$ from the training set as follows:

$$M_c(\mathbf{f}_t) = \hat{\mathbf{c}}_t. \quad (6)$$

With a sequence of ground truth condition values $\mathbf{C} = \{\mathbf{c}_1,\ldots,\mathbf{c}_T\}$ that is aligned with $\mathbf{X}$, we are able to compute the condition regression loss through mean-square error as follows

$$\begin{aligned}\mathcal{L}_{\text{condition}}(\theta_f,\theta_c) &= -\frac{1}{T}\sum_{t=1}^{T}(\hat{\mathbf{c}}_t-\mathbf{c}_t)^2\\ &= -\frac{1}{T}\sum_{t=1}^{T}\left[M_c(M_f(\mathbf{x}_t))-\mathbf{c}_t\right]^2. \quad (7)\end{aligned}$$

3.1.3. Optimization

For brevity, we name both condition classification and regression loss as condition loss $\mathcal{L}_{\text{condition}}$ and we call both condition classification and regression networks as condition network M_c. To make the deep features $\mathbf{F}$ condition-invariant, the distributions of the deep features from different conditions should be as close to each other as possible. Therefore, the M_f and M_c are jointly trained with an adversarial objective, in which θ_f is adjusted to *maximize* the frame-level condition loss $\mathcal{L}_{\text{condition}}$ while θ_c is adjusted to *minimize* $\mathcal{L}_{\text{condition}}$. This minimax competition will first increase the discriminativity of M_c and the speaker-invariance of the deep features generated by M_f, and will eventually converge to the point where M_f generates extremely confusing deep features that M_c is unable to distinguish.

At the same time, we want to make the deep features speaker-discriminative by minimizing the speaker classification loss $\mathcal{L}_{\text{speaker}}$ as in Eq. (3).

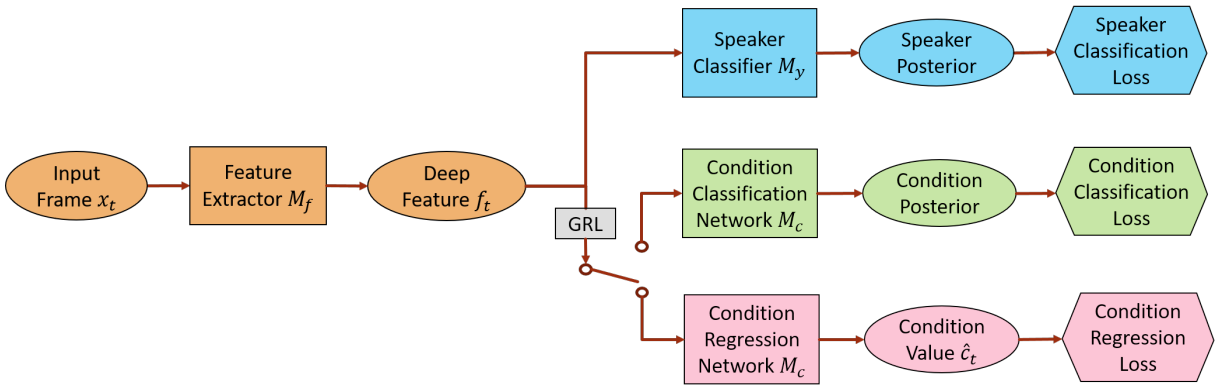

Fig. 1. Adversarial training of background DNN (consisting of M_f followed by M_y) for condition-robust speaker verification. The condition *classification* network is selected as the auxiliary network if the conditions can be represented as a categorical variable and the condition *regression* network is used when the conditions are expressed as a continuous variable.

Overall, we find the optimal parameters $\hat{\theta}_y, \hat{\theta}_f, \hat{\theta}_c$ through adversarial multi-task learning as follows

$$(\hat{\theta}_f, \hat{\theta}_y) = \arg\min_{\theta_f, \theta_y} \mathcal{L}_{\text{speaker}}(\theta_f, \theta_y) - \lambda \mathcal{L}_{\text{condition}}(\theta_f, \hat{\theta}_c) \quad (8)$$

$$(\hat{\theta}_c) = \arg\min_{\theta_c} \mathcal{L}_{\text{condition}}(\hat{\theta}_f, \theta_c), \quad (9)$$

where λ controls the trade-off between the speaker classification loss and the condition loss in Eq.(3) or Eq.(5) respectively.

The optimization can be implemented through standard stochastic gradient descent by inserting a GRL in between the feature extractor and the condition network [25]. This GRL serves as an identity transform in the forward propagation and multiplies the gradient by $-\lambda$ during the backward propagation.

3.2. Enrollment and Evaluation

The optimized feature extractor M_f is used to generate *condition-invariant* deep embeddings for enrolled speakers and test utterances. We use $\mathbf{h}^s$, the mean of deep features given speech frames of an enrolled speaker s at the input of M_f, as the deep embedding for s and use $\mathbf{h}^u$, the mean of deep features generated by feeding a test utterance u into M_f, as the deep embedding for u. The cosine distance between each pair of $\mathbf{h}^u$ and $\mathbf{h}^s$ is computed as the score and is compared with a threshold to make the speaker verification decision.

Note that we do not perform probabilistic linear discriminant analysis (PLDA) scoring which is widely adopted in most i-vector and deep embedding systems because: (1) The parameters of PLDA classifiers are learned from the training data and may not generalize well to the test data, especially that with *unknown* noise. Cosine distance is a non-parametric metric and is more robust to mismatched noisy conditions during testing. (2) In this work, the extracted deep embeddings are already low-dimensional (e.g., 200) and PLDA does not further improve the SV performance from our experience. In other words, the dimension reduction has been implicitly performed by the background DNN.

4. EXPERIMENTS

In this work, we perform ASV by suppressing the environment and SNR variabilities in speaker modeling and evaluate its performance on Microsoft Cortana TD-SV task.

4.1. Dataset Description

We collect a set of utterances that start with the voice activation phrase "Hey Cortana" from the Windows 10 desktop Cortana service logs. "Hey Cortana" segments are cut out using a keyword detector. The duration of each keyword segment is around 65 to 110 frames. We select 6.8M utterances from 8k different speakers, each with 100 to 1000 utterances as the training set. We then simulate 20.4M noisy utterances by adding 4 types of real noise (on buses (BUS), in cafes (CAF), in pedestrian areas (PED), at street junctions (STR)) from CHiME-3 [26] dataset to the 6.8M clean Cortana data to form the noisy training data. The noise type is randomly selected and the noise level is randomly scaled before simulating each noisy utterance to make sure the amount of noisy data of each type is almost the same and the utterance-level SNR values of simulated data are within the range of 0dB to 20dB. The final training set consists of both the clean utterances and the simulated noisy utterances.

From the clean Cortana data, we select 6 utterances from each of the 3k speakers as the enrollment data (called "Enroll A"). We select 60k utterances from 3k target speakers and 3k impostors in Cortana dataset and mix them with CHiME-3 real noise to generate the noisy evaluation set (called "Test A") in the same way as training data simulation. Enroll A is always used for enrollment when Test A is used for evaluation. Speakers in the training set and Enroll A (or Test A) do not overlap. Since Test A share the same types of noise with the training set, it is used to evaluate the ASV performance with known conditions.

Further, we record 4 close-talk clean utterances from each of the 183 speakers using far-field devices to create a new enrollment set (called "Enroll B"). We use devices of the same type to collect 5546 real noisy test utterances from 183 target speakers and 183 impostors under various environments (e.g., background music, TV, etc.) and with different recording distances (e.g., at 1m and 5m). We call this test set "Test B". Enroll B is always used for enrollment when Test B is used for evaluation. There is no overlap between the speakers in the training set and Enroll B (or Test B). With completely different recording conditions, Test B is used to evaluate the generalization capability of ASV to unknown conditions.

4.2. Baseline System

As the baseline system, we first train a feed-forward background DNN for speaker classification using 6.8M utterances from the train-

ing set with cross-entropy criterion and extract the deep embeddings of enrolled speakers and test utterances for speaker verification as described in Section 2. Our baseline is similar to the x-vector system [8] in that data augmentation is applied by adding different types of noise to improve the robustness of deep embeddings.

The 29-dimensional log Mel filterbank features together with 1st and 2nd order delta features (totally 87-dimensional) are extracted. Each frame is spliced together with 25 left and 25 right context frames to form a 4437-dimensional input feature. The spliced features are fed as the input of the feed-forward DNN after global mean and variance normalization. The DNN has 5 hidden layers with 2048, 1024, 1024, 512, 200 hidden units for the bottom to the top hidden layers, respectively. The non-linear activation function for each hidden layer is relu. The output layer of the DNN has 8398 output units corresponding to 8398 speakers in the training set with softmax non-linearity. The 200-dimensional deep embeddings for enrolled speakers and test utterances are computed by taking the average of the last hidden layer outputs.

As shown in Tables 1 and 2, the EERs for the deep embedding are 4.22% and 13.02% on Test A and Test B, respectively. In Table 2, Test B is first categorized into Quiet, TV and Music based on the type of noise/interference and is then re-classified according to the recording distance into 1m, 3m and 5m categories. For example, a test utterance can be recorded under background Music with a distance of 3m from the speaker.

System	DE	EI DE	SI DE	EI+SI DE
EER (%)	4.22	3.95	3.98	**3.85**

Table 1. The speaker verification EER (%) of baseline deep embedding (DE), ASV with environment-invariant (EI) DE only, ASV with SNR-invariant (SI) DE only and ASV with both EI and SI DE on Test A with *known* conditions.

System	Quiet	TV	Music	1m	3m	5m	Total
DE	10.52	15.53	11.71	10.06	11.98	13.52	13.02
EI DE	9.75	14.18	10.08	9.00	10.74	13.00	11.71
SI DE	9.95	14.18	10.40	**8.76**	10.95	12.93	11.80
EI+SI DE	**9.32**	**14.00**	**9.46**	8.87	**10.09**	**12.72**	**11.13**

Table 2. The speaker verification EER (%) of baseline deep embedding (DE), ASV with environment-invariant (EI) DE only, ASV with SNR-invariant (SI) DE only and ASV with both EI and SI DE on Test B with *unknown* conditions.

4.3. Adversarial Speaker Verification

We further perform adversarial training of the baseline background DNN with 6.8M utterances in the training set to learn condition-invariant deep embeddings for ASV. The feature extractor M_f is initialized with the input layer and 5 hidden layers of the background DNN and the speaker classifier network M_y is initialized with the output layer. The deep hidden feature is the 200-dimensional output of the last hidden layer of the background DNN. We first address the effect of environment variability which is a factor of categorical conditions. The environment classification network M_c is a feed-forward DNN with 2 hidden layers and 512 hidden units for each layer. The output layer of M_c has 5 units predicting the posteriors of 4 noisy and 1 clean environments in the training set. As shown in Tables 1 and 2, with environment-invariant deep embeddings, the ASV achieves 3.95% and 11.71% EERs, which are 6.4% and 10.1% relatively improved over the baseline deep embedding on Test A and Test B, respectively, when $\lambda = 1.5$.

Then we explore the reduction of SNR variability, a factor of continuous conditions. We introduce an SNR regression network M_c designed as a feedforward DNN with 2 hidden layers and 512 hidden units for each layer. The output layer of M_c has 1 unit predicting the SNR value of each input speech frame in the training set. The frames in the same utterance share the same utterance-averaged SNR. With SNR-invariant deep embeddings, the ASV achieves 3.98% and 11.80% EERs, which are 5.7% and 9.4% relatively improved over the baseline deep embedding on Test A and Test B, respectively, when $\lambda = 0.002$. Note that the initial SNR regression loss is at the order of magnitude of 2, while the initial speaker classification loss is at -1. λ needs to be small enough to match the dynamic ranges of the two losses. We see that environment-invariant deep embeddings achieve slightly better ASV EER than SNR-invariant ones.

Multi-factor (MF) adversarial learning was proposed in [27] to simultaneously suppress multiple factors that cause the condition variability. In this work, we perform MFA speaker verification to learn deep embeddings that are both environment-invariant and SNR-invariant. The EER further reduces to 3.85% and 11.13% with 8.8% and 14.5% relative improvements over the baseline deep embedding on Test A and Test B, respectively with $\lambda_1 = 1.5$ for enivironment factor and $\lambda_2 = 0.002$ for SNR factor. MF ASV achieves about 5.3% additional relative gain over the best single-factor ASV. In all cases, the SV performance is quite stable with the variation of λ values.

From Table 2, we see that MF ASV achieves smaller relative EER improvements under TV and 5m far-field conditions. The reason is that under TV environment, there exists background speech from another speaker in the TV program as the interference, which makes ASV much harder to verify the identity of the claimed speaker. The reverberance effect exists in addition to the background noise when the recording devices are placed 5m from the speakers. MF ASV only addresses noise type and noise level variabilities and does not explicitly tackle the variability of room impulse response, so its gain is relatively smaller under the far-field condition.

The significantly larger relative EER gains on Test B with *unknown* conditions for all three ASV systems show that ASV can effectively learn a canonical condition-invariant speaker model with remarkably increased generalization ability on unknown conditions.

4.4. Conclusions

We propose an adversarial speaker verification method in which a background DNN for speaker classification is jointly trained with a condition network to learn speaker-discriminative and condition-invariant deep embeddings for condition-robust SV. A regression network is used to reconstruct the continuous condition variable.

ASV achieves 8.8% and 14.5% relative improvements over the deep embedding baseline on Test A with known conditions and Test B with unknown conditions. Environment-invariant deep embeddings work better than SNR-invariant ones for ASV. The joint suppression of multiple factors of condition variability further improves the ASV performance. The significantly larger gains of ASV for the *unknown* conditions shows its strong generalization capability.

5. REFERENCES

[1] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez-Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification.," in *ICASSP*. Citeseer, 2014, vol. 14, pp. 4052–4056.

[3] Fred Richardson, Douglas Reynolds, and Najim Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[4] David Snyder, Pegah Ghahremani, Daniel Povey, Daniel Garcia-Romero, Yishay Carmiel, and Sanjeev Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 165–170.

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

[6] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[7] 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), 2016 IEEE*. IEEE, 2016, pp. 171–178.

[8] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *ICASSP 2018*, 2018.

[9] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 4, pp. 745–777, April 2014.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, pp. 2672–2680. 2014.

[11] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, "Segan: Speech enhancement generative adversarial network," in *INTERSPEECH*, 2017.

[12] Chris Donahue, Bo Li, and Rohit Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," *arXiv preprint arXiv:1711.05747*, 2017.

[13] Zhong Meng, Jinyu Li, Yifan Gong, and Biing-Hwang (Fred) Juang, "Adversarial feature-mapping for speech enhancement," in *Proc. Interspeech*, 2018.

[14] Takuhiro Kaneko and Hirokazu Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv preprint arXiv:1711.11293*, 2017.

[15] Zhong Meng, Jinyu Li, and Yifan Gong, "Adversarial speaker adaptation," in *Proc. ICASSP*, 2019.

[16] Sining Sun, Binbin Zhang, Lei Xie, and Yanning Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79 – 87, 2017, Machine Learning and Signal Processing for Big Multimedia Analysis.

[17] Z. Meng, Z. Chen, V. Mazalov, J. Li, and Y. Gong, "Unsupervised adaptation with domain separation networks for robust speech recognition," in *Proceeding of ASRU*, 2017.

[18] Yusuke Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition.," in *INTERSPEECH*, 2016, pp. 2369–2372.

[19] D. Serdyuk, K. Audhkhasi, P. Brakel, B. Ramabhadran, et al., "Invariant representations for noisy speech recognition," in *NIPS Workshop*, 2016.

[20] George Saon, Gakuto Kurata, Tom Sercu, et al., "English conversational telephone speech recognition by humans and machines," *Proc. Interspeech*, 2017.

[21] Z. Meng, J. Li, Z. Chen, et al., "Speaker-invariant training via adversarial learning," in *Proc. ICASSP*, 2018.

[22] Qing Wang, Wei Rao, Sining Sun, Lei Xie, Eng Siong Chng, and Haizhou Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," *ICASSP 2018*, 2018.

[23] Zhong Meng, Jinyu Li, and Yifan Gong, "Cycle-consistent speech enhancement," *Interspeech*, 2018.

[24] Zhong Meng, Jinyu Li, and Yifan Gong, "Adversarial feature-mapping for speech enhancement," *Interspeech*, 2018.

[25] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, 2015, vol. 37 of *Proceedings of Machine Learning Research*, pp. 1180–1189, PMLR.

[26] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. ASRU*, 2015, pp. 504–511.

[27] Zhong Meng, Jinyu Li, Yifan Gong, and Biing-Hwang Juang, "Adversarial teacher-student learning for unsupervised domain adaptation," in *Proc. ICASSP*. IEEE, 2018, pp. 5949–5953.