INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

NPU Speaker Verification System for INTERSPEECH 2020 Far-Field Speaker Verification Challenge

*Li Zhang, Jian Wu, Lei Xie**

Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, Xi'an, China

lizhang.aslp.npu@gmail.com, lxie@nwpu.edu.cn

Abstract

This paper describes the NPU system submitted to Interspeech 2020 Far-Field Speaker Verification Challenge (FFSVC). We particularly focus on far-field text-dependent SV from single (task1) and multiple microphone arrays (task3). The major challenges in such scenarios are *short utterance* and *cross-channel and distance mismatch* for enrollment and test. With the belief that better speaker embedding can alleviate the effects from short utterance, we introduce a new speaker embedding architecture - ResNet-BAM, which integrates a bottleneck attention module with ResNet as a simple and efficient way to further improve representation power of ResNet. This contribution brings up to 1% EER reduction. We further address the mismatch problem in three directions. First, *domain adversarial training*, which aims to learn domain-invariant features, can yield to 0.8% EER reduction. Second, *front-end signal processing*, including WPE and beamforming, has no obvious contribution, but together with data selection and domain adversarial training, can further contribute to 0.5% EER reduction. Finally, data augmentation, which works with a specifically-designed data selection strategy, can lead to 2% EER reduction. Together with the above contributions, in the middle challenge results, our single submission system (without multi-system fusion) achieves the first and second place on task 1 and task 3, respectively.

Index Terms: speaker verification, far-field, domain adversarial training, data augmentation

1. Introduction

With the rise of deep neural networks (DNN) and easy availability of computing resources and massive data, speaker verification (SV) performance has been significantly improved in the past several years. However, such advances are mainly achieved on close-talk scenarios with less interference. With the fast proliferating of smart devices, such as smart speakers and various voice-enabled IoT gadgets, the need for far-field speech interaction will continue to grow. Recognizing who is speaking is essential to such smart devices to provide customized services. Far-field speech tasks including speaker recognition remain challenging yet due to attenuated speech signals, noise interference as well as room reverberations. Particularly for smart devices, it is more convenient for users to enroll on a short utterance, e.g., a trigger word, from a close-talk portable device such as a cellphone but talk to a smart device from distance to obtain authentication. It apparently raises other problems – short utterance verification, data mismatch between enroll and test in terms of channels and distances. Interspeech 2020 Far-Field Speaker Verification Challenge (FFSVC) [1] provides a common testbed for researchers to address the above mentioned difficult problems – *deteriorated signal*, *short utterance* and *data mismatch*.

In this paper, we present our efforts to deal with the above mentioned problems with our submitted system to FFSVC. We particularly introduce our approaches in the two text-dependent tasks, i.e., far-field text-dependent SV from single (task1) and multiple arrays (task3). We introduce a new speaker embedding architecture with more powerful speaker representation, which is built on ResNet with attention module. The new architecture ResNet-BAM achieves 1% abosolute equal error rate (EER) reduction compared to the baseline ResNet model. We further address deteriorated signal and mismatch problem with front-end processing and domain adversarial training (DAT). The two methods can bring 0.5% and 0.8% absolute EER reduction, respectively. Finally, data augmentation, which works with a specifically-designed data selection strategy, can lead to 2% absolute EER reduction. With the above contributions, our single submission system (without multi-system fusion) achieves the first and second place on task 1 and task 3, respectively.

The rest is organized as follows. Section 2 introduces the related works and Section 3 describes the system overview. Section 4 details the proposed Resnet-BAM model for better speaker embedding, followed by experiments to validate its efficacy. Section 5 focuses on domain adversarial training and its evaluation. Section 6 analyzes the effects from front-end processing and data augmentation and selection. Section 7 summarizes the official evaluation results and concludes this paper.

2. Related Works

Most of approaches to deal with short utterance SV task focus on improving the speaker embedding network with stronger extracting capabilities. Some improve x-vector-based models [2] while others work on start-of-the-art convolution neural networks (CNN) on various datasets [3, 4]. Li *et al.* [5] have reported that CNN-based network even can recognize speaker by cough or laugh recordings [6], which are extremely short 'utternaces'. Wang *et al.* [7] integrated deep discriminant analysis into CNN-based structure to achieve good performance on short utterance SV. Study from [8] has confirmed that ResNet architectures outperform the standard x-vector approach in terms of SV quality for both long-duration and short-duration utterances. Since then, ResNet has become the most popular network structure for speaker embedding extraction. As loss function is also essential to the network's learning ability, there have been various studies exploring in this direction [9].

Training-testing or enroll-testing mismatch is a research problem explored for many years. Model adaptation or more formally domain adaptation, which aims to transfer the source model to the target domain, has been studied extensively [10, 11]. Domain adversarial training (DAT) is the most recent approaches developed with DNN's strong and flexible modeling

* Corresponding author.

http://dx.doi.org/10.21437/Interspeech.2020-2688

ability. As a specifically designed multi-task learning framework, domain adversarial neural network (DANN) injects a domain classifier as an auxiliary task with a gradient reversal layer to learn domain-invariant features [12]. Wang *et al.* [13] have recently applied DANN to remove cross-dataset variation and project data from difference datasets into the same subspace and superior SV performance has been reported on 2013 domain adaptation challenge (DAC) data. Many followers along this direction have investigated on multi-language [14, 15] and multi-channel adaptation [16, 17].

The mismatch from cross-channel (and distance) enrollment and test usually can be compensated by so-called *front-end processing* which utilizes traditional signal processing technologies. A dereverberation module, such as the one adopts weighted prediction error (WPE) [18] algorithm, is usually employed to remove reverberations from far-field collected speech signal thus to match the cross-talk signal. Moreover, beamformers [19] are adopted to process multi-channel signals collected from microphone array(s), resulting in a single-channel enhanced speech signal. There are also some new development on neural front-ends, such as neural dereverberation [20] and neural beamforming [21, 22]. Yang *et al.* [23] have proposed joint optimization of neural beamforming and dereverberation with x-vectors for robust speaker verification.

Data augmentation [24] is another commonly used simple-but-effective trick to alleviate data mismatch. For instance, many approaches on VOiCEs challenge [18, 25, 26] have considered this trick with improved performances. Augmentation to the training data makes the model 'see' more acoustic environments with diversity, leading to more robust speaker embedding [27]. Meanwhile, augmentation in enrollment and test not only can compensate the mismatch between enrollment and test, but also can make up for the negative effects from short utterance duration during evaluation. So far, data augmentation has been an effective and intuitive way for robust modeling by improving the diversity of the data. But there is still a fundamental problem: does all the augmented data work well and how to select more effective data? This paper tries to answer the question with a simple data selection strategy.

3. System Overview

Figure 1 illustrates the basic diagram of our system in FFSVC. It mainly consists of two sub-modules. Aiming to alleviate the data mismatch problem, the data processing module is composed of front-end processing and data augmentation, while both go through data selection to result in the final 'high-quality' augmented data. Another core module is speaker embedding extractor, which is composed of the proposed ResNet-BAM network for extracting better speaker embedding and the deep adversarial training built upon the embedding network to learn domain-invariant and speaker-discriminative features.

4. ResNet-BAM Model

Our speaker embedding extractor is a CNN-based ResNet-50 model. We replace the average pooling layer of ResNet-50 [28] with a statistic pooling layer the same as that in x-vector [27], and then add a fully connected layer followed with the statistic pooling layer as our baseline model. We propose to use an attention-enhanced ResNet structure to further improve the ability of the baseline embedding extractor, which originally shows superior performance in several image classification and detection tasks [29]. Specifically, we add bottleneck attention modules (BAM) [29] followed with bottleneck layers (ResNet-

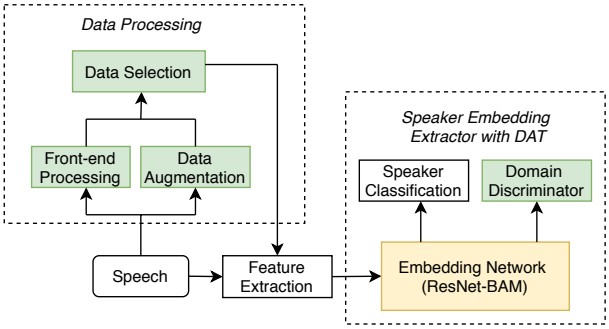

Figure 1: *The overview of our speaker verification system.*

BAM) in ResNet-50 to extract better speaker embedding. BAM is able to emphasize important elements in 3D feature map generated from convolution. In speech, 3D feature map has channel dimension (filter number of convolution), time dimension and frequency dimension. There are two branches to calculate attention masks: channel attention is to learn which channels are more important for the final classification task, while time-frequency attention aims to learn which points in time-frequency domain are more effective for the classification task. The two branches (channels and time-frequency) explicitly learn 'what' and 'where' in the spectral graph to focus on. The structure of ResNet-BAM is shown in Figure 2.

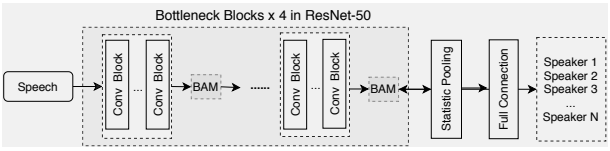

Figure 2: *The structrue of ResNet-BAM*

4.1. Attention module

The detail of bottleneck attention module in ResNet-BAM is illustrated in Figure 3. After several layers of convolution on input x, we obtain the 3D feature map $F' \in R^{C\times T\times F}$ after the bottleneck layer. Then an attention module infers a 3D attention map $M(F) \in R^{H\times T\times F}$. The refined feature map after the attention module F'' is computed as

$$F'' = F' + F' \cdot M(F') \tag{1}$$

where $\cdot$ denotes element-wise multiplication. $M(F')$ is combined with two attentions masks – channel attention mask $M_C(F') \in R^C$ and time-frequency attention mask $M_{tf}(F') \in R^{T\times F}$. Two branches of attention are computed in parallel. In the original ResNet-BAM [29], $M(F')$ is the direct addition between $M_{tf}(F')$ and $M_{tf}(F')$, and then normalized by sigmoid function into a 3D attention mask in range of (0,1). But according to our empirical experiments, there exists an offset problem if corresponding elements of the two attention masks have different sign, and we cannot judge the direct relationship between the positive or negative values of the attention masks and the final recognition result. Hence we switch the two steps: first do sigmoid on the two masks and then add together. Finally the attention mask $M(F')$ is calculated as

$$M(F') = (\text{Sigmoid}(M_C(F')) + \text{Sigmoid}(M_{tf}(F'))/2. \tag{2}$$

Although each channel (filters in convolution) contains a specific feature representation, different channel elements cannot have the same effect on the final recognition task. Learning the importance mask of each channel elements thus not only can guide the model to focus on the more effective points but also speed up model convergence. Given a 3D feature map F', we use global average pooling to get a vector in channel dimension.

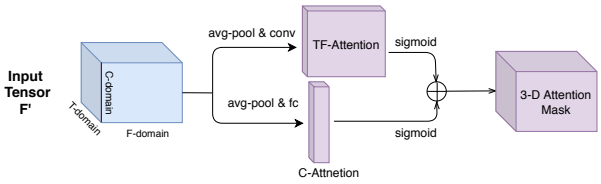

Figure 3: *Details of bottleneck attention module (BAM) in ResNet-BAM. F' is bottleneck output of several convolution.*

Then multi-layer perceptron (MLP) is utilized to estimate attention across channels and after batch normalization, the output is produced. In brief, the channel attention is computed as

$$M_C(F') = \text{BN}(\text{MLP}(\text{AvgPool}(F')_{C\times 1\times 1})) \quad (3)$$

The time-frequency (TF) branch aims to learn an attention map to emphasize or suppress different points on the spectral graph. $M_{tf}(F')$ is calculated in time-frequency domain. Firstly, average pooling is conducted on channel dimension, then we use convolutions to learn a 2D mask in TF domain. After we normalize the time-frequency mask. The TF mask is calculated as

$$M_{tf}(F') = \text{BN}(f'(\text{AvgPool}(F')_{1\times T\times F})) \quad (4)$$

where f' represents the convolution operations after average pooling on 3D input feature.

Finally, channel-attention and TF-attention masks are combined through Eq.(2).

4.2. Experiments on ResNet-BAM

The challenge only allows to use the datasets shared on OpenSLR for model training. Our team choose five datasets (SLR 33, 38, 62, 82 and 85) together with the FFSVC official data (FFSVC20) as the basic training data, in which total speaker number is 3,211 with about 2,100 hours of Mandarin speech. The official development set includes 35 speakers. Trials are 53,996 pairs in both task 1 and 3. Enrollment data is recorded from iPhone. Test data in task 1 is from one random selected microphone array with four channels while test data in task 3 is from 2-4 random selected microphone arrays, each array with four channels [1].

We conduct all experiments on Pytorch. Acoustic feature is 30-dim MFCCs with kaldi [30] energy VAD to remove silence frames beforehand. Batch size is set to 64 and input tensor size is [1,256,30]. We trunk every utterance randomly. If the audio not reaches to 256 frames, we repeat the original audio and random trunk again. Initial learning rate is 0.1 and it decays to the original 10% every 5 epochs. Optimizer is Stochastic gradient descent in Pytorch. The core metrics of the challenge are equal error rate (EER) and minimum detection cost function (minDCF). All scores of this paper we calculated are based on cosine distance.

Results on development data for task 1 and 3 are summarized in Table 1. We can see that the proposed ResNet-BAM model can bring roughly 1% reduction in EER for both task 1 and task 3. We believe that the performance gain mainly comes from the improved speaker representation power by using the specifically-designed attention module which can guide the model to learn discriminative embedding more effectively for the speaker verification task.

5. Domain Adversarial Training

To alleviate domain mismatch, an intuitive idea is to project two different domains into a common space for speaker recognition. This can be achieved by domain adversarial training (DAT) with a gradient reversal layer (GRL) which aims to learn domain-invariant and discriminative speaker embedding [13].

Table 1: *Results on development set for two embedding networks (ResNet and ResNet-BAM) w/o and w/ DAT.*

Model Name	Task1		Task3	
	EER (%)	minDCF	EER (%)	minDCF
ResNet	8.34	0.8539	7.98	0.8231
ResNet-BAM	**7.43**	**0.7707**	**6.89**	**0.7312**
ResNet-DAT	7.59	0.7921	7.05	0.7230
ResNet-BAM-DAT	**6.71**	**0.7507**	**6.19**	**0.7023**

5.1. Invariant feature learning via DAT

As shown in Figure 4, the DAT module is built on a pre-trained ResNet-BAM model, formed as a multi-task learning (MTL) problem, where the main task is speaker recognition and the auxiliary task is domain discrimination. Specifically, the domain discriminator sub-network in our task is designed to distinguish close-talk speech from far-field speech. The two branches take input from a shared feature extractor sub-network that aims to learn representations that capture the underlying speaker discriminative information and are independent of speech domain. Different from conventional MTL, an inserted gradient reversal layer is essential to learn domain-independent features.

Given an input x and its speaker label y as well domain label d, the predicted speaker and domain label are y' and d', respectively. The loss of the two tasks are combined as

$$L(\theta_f, \theta_y, \theta_d) = \frac{1}{n}\sum_{i=1}^{n} L_y^i(\theta_f, \theta_y) - \lambda\frac{1}{n}\sum_{i=1}^{n} L_y^i(\theta_f, \theta_d) \quad (5)$$

where $\theta_f, \theta_y, \theta_d$ are parameters of the shared feature extractor and the two classifiers, and L_y and L_d are the speaker prediction loss and the domain classification loss, respectively. n is the number of training samples. The joint loss is to minimize the speaker classification loss and maximize the domain discriminator loss at the same time, which is achieved by the GRL to reverse the sign of gradient before the domain discriminator. By this way, the feature extractor is able to learn domain-invariant and speaker-discriminative features.

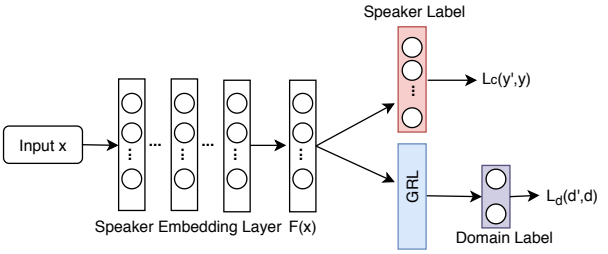

Figure 4: *DAT in speaker embedding network. GRL is a gradient reverse layer.*

5.2. Experiments on DAT

Our DAT approach is based on the pre-trained ReseNet and ReseNet-BAM models in Section 4. Recall that the two models are trained using the datasets introduced in Section 4.2. Then the two models are equipped with the DAT structure and fine-tuned using the official FFSVC20 data. Here, data recorded on iPhone is regard as source while data recorded by microphone array(s) is considered as target. Results in Table 1 shows that with the help of DAT, EER for both tasks has been reduced about 0.8%. This performance gain can be observed for both ReseNet and ReseNet-BAM models. Using this DAT approach, we are able to minimize the gap between the source and target

feature distributions. Therefore, the learned embedding is less dependent on the domain shift.

6. Front-end and Data Augmentation

6.1. Front-end processing

In evaluation stage, we use the WPE algorithm [31] to handle reverberation issues. As the test data provide multi-channel speech signal, we adopt minimum variance distortionless response (MVDR) beamformer to suppress interfering noise, which has been proved to benifit other speech tasks, e.g., speech recognition in previous studies. The covariance matrices in MVDR are estimated using time-frequency masks generated by two components complex Gaussian mixture models [32]. Given test audio samples $a_j \in A = \{a_1, a_2, ...a_M\}$, after performing WPE and MVDR on utterance a_j, we get enhanced audio set $A' = \{a'_1, a'_2, ...a'_M\}$. Then 30-dimensional MFCCs feature is extracted from A' and fed into the well-trained embedding extractor $F(x)$ to obtain the new embedding for verification. The experiments results are in Table 2.

Table 2: *Experiments results on ResNet-BAM-DAT with front-end methods on development dataset*

WPE	Beamformer	DAT	Data Selection	Task1		Task3	
				EER (%)	minDCF	EER (%)	minDCF
✓	×	×	×	7.12	0.7820	6.83	0.7216
×	✓	×	×	7.21	0.8177	6.92	0.7541
✓	×	×	✓	6.46	0.6901	5.97	0.6728
×	✓	✓	×	6.39	0.6873	5.89	0.6538
✓	✓	✓	✓	**6.22**	**0.6518**	**5.86**	**0.6006**

We find that there is no improvement but worse results on the use of the front-end processing pipeline. Hence we double-check the experimental settings and find that after the enhancement processing some audios are contaminated with stronger noise due to the failure of the original microphone channels, which accounts for 10% of the total development set. To deal with the problem, we propose a data selection strategy to exclude the failure utterances from the processed data. We compare the cosine distance between original embedding and enhanced embedding and discard the enhanced data which score is lower than a threshold θ (empirically set to 0.7 in our experiments). This way can ensure the processed data not too outrageous and avoid performance degradation.

Data processed by beamformer, paired with iPhone-recorded data are used to fine-tune the ResNet-BAM-DAT model again. Note that the two domains for adversarial training are close-talk speech and far-field speech. The evaluation flow is shown in Figure 5. After the data selection and another round of model fine-tuning, we achieve reasonable results on the development set, as shown in Table 2. With WPE and MVDR beamforming, we can achieve 0.5% EER absolute reduction.

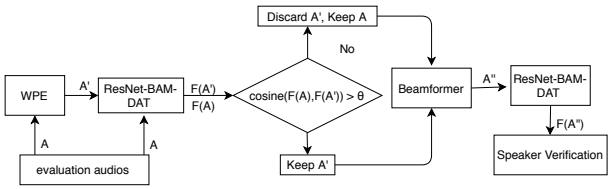

Figure 5: *Evaluation flow with front-end technologies.*

6.2. Data augmentation and selection

Data augmentation is a commonly strategy to improve the data coverage. We use open-sourced MUSAN [33] noise and room impulse response (RIR) databases from [34] to perform training data augmentation, using the official scripts provided by Kaldi. At the same time, we do voice variable speed augmentation on enrollment and test of development and evaluation data.

Besides, to match the far-field conditions, we simulate multi-channel version of the enrollment data using artificially generated RIRs. We generate totally 40,000 RIRs from 200 different room configurations and the same configurations of microphone arrays as described in [1], which aims to cover the recording environment of the challenge data. After the simulation procedure, we adopt a different selection strategy from Section 6.1 to pick up well quality simulated far-field utterances. We has a clever strategy to determine hyperparameter θ which is not a fixed empirical data but dependent on the development data. Development test trials with labels and evaluation test trials without labels [1] have recorded in the same room and devices. Acoustic properties caused by far-field scenes are the same. We use simulated development data doing cosine distance with corresponding speaker's test data (far-field) to obtain a series suitable parameters about generating RIR and a good hyperparameter θ. We use the fit RIR parameters to simulate enrollment data of evaluation data and then use the appropriate θ to select high quality simulated data to do test. This way can find more benefit RIR parameters to fit the real recorded environment of test data to reduce the mismatch between simulated enrollment and test data.

The effects of data augmentation are shown in Table 3. Augmentation on training data brings 1% reduction on EER. Meanwhile, doing data augmentation on enrollment and test set, EER is reduced the most, even up to 2%.

Table 3: *Experiment results with data augmentation on ResNet-BAM on development dataset*

Aug. Training	Aug. Enrollment	Aug. Test	Task1		Task3	
			EER (%)	minDCF	EER (%)	minDCF
✓	×	×	6.37	0.6911	5.45	0.5622
✓	✓	×	4.81	0.5009	3.55	0.3977
✓	✓	✓	**4.22**	**0.4213**	**3.39**	**0.3728**

7. Evaluation Results and Conclusions

At the middle deadline, our submitted system is ResNet-BAM trained with augmented (and selected) data. The evaluation results of task 1 and task 3 on the leadboard of FFSCV Official website are shown in Table 4. Our submission achieves the first and second place on task 1 and 3 respectively. The scores of our submitted systems are bold and red-colored.

Table 4: *Evaluation dataset results of the top 3 teams by middle deadline on leadboard.*

Rank	Task1		Task3	
	EER (%)	minDCF	EER (%)	minDCF
1	**5.39**	**0.4636**	5.53	0.4584
2	5.08	0.5002	**6.44**	**0.4585**
3	4.72	0.5200	5.14	0.4708

This paper introduces the main approaches used in our submitted system to FFSVC, specially designed speaker embedding network ResNet-BAM, domain adversarial training, front-end processing and data augmentation as well as selection. The most profitable method is data augmentation with data selection. There is still potential space to improve in far-field SV. For instance, we expect front-end processing should play a vital role in dealing with the mismatch problem. Especially, we plan to explore the recent neural front-end approaches, such as neural dereveberation [35] and neural beamformers [21, 36]. Moreover, we hope the performance of far-field SV can be further boosted through front-end and speaker embedding network joint training.

8. References

[1] X. Qin, M. Li, H. Bu, R. K. Das, W. Rao, S. Narayanan, and H. Li, "The ffsvc 2020 evaluation plan," *arXiv preprint arXiv:2002.00387*, 2020.

[2] A. Kanagasundaram, S. Sridharan, S. Ganapathy, P. Singh, and C. B. Fookes, "A study of x-vector based speaker recognition on short utterances," 2019.

[3] S. Yadav and A. Rai, "Frequency and temporal convolutional attention for text-independent speaker recognition," *arXiv preprint arXiv:1910.07364*, 2019.

[4] T. Zhou, Y. Zhao, J. Li, Y. Gong, and J. Wu, "Cnn with phonetic attention for text-independent speaker verification," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 718–725.

[5] L. Li, Y. Chen, Y. Shi, Z. Tang, and D. Wang, "Deep speaker feature learning for text-independent speaker verification," *arXiv preprint arXiv:1705.03670*, 2017.

[6] M. Zhang, Y. Chen, L. Li, and D. Wang, "Speaker recognition with cough, laugh and" wei"," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 497–501.

[7] S. Wang, Z. Huang, Y. Qian, and K. Yu, "Discriminative neural embedding learning for short-duration text-independent speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1686–1696, 2019.

[8] A. Gusev, V. Volokhov, T. Andzhukaev, S. Novoselov, G. Lavrentyeva, M. Volkova, A. Gazizullina, A. Shulipa, A. Gorlanov, A. Avdeeva *et al.*, "Deep speaker embeddings for far-field speaker recognition on short utterances," *arXiv preprint arXiv:2002.06033*, 2020.

[9] J. M. Coria, H. Bredin, S. Ghannay, and S. Rosset, "A comparison of metric learning loss functions for end-to-end speaker verification," *arXiv preprint arXiv:2003.14021*, 2020.

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

[11] Y. Tu, M.-W. Mak, and J.-T. Chien, "Variational domain adversarial learning for speaker verification," *Proc. Interspeech 2019*, pp. 4315–4319, 2019.

[12] S. Sun, B. Zhang, L. Xie, and Y. Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79–87, 2017.

[13] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4889–4893.

[14] J. Rohdin, T. Stafylakis, A. Silnova, H. Zeinali, L. Burget, and O. Plchot, "Speaker verification using end-to-end adversarial language adaptation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6006–6010.

[15] W. Xia, J. Huang, and J. H. Hansen, "Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5816–5820.

[16] S. Shon, S. Mun, W. Kim, and H. Ko, "Autoencoder based domain adaptation for speaker recognition under insufficient channel information," *arXiv preprint arXiv:1708.01227*, 2017.

[17] C. Luu, P. Bell, and S. Renals, "Channel adversarial training for speaker verification and diarization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7094–7098.

[18] D. Cai, X. Qin, W. Cai, and M. Li, "The dku system for the speaker recognition task of the 2019 voices from a distance challenge," *arXiv preprint arXiv:1907.02194*, 2019.

[19] H. Taherian, Z.-Q. Wang, and D. Wang, "Deep learning based multi-channel speaker recognition in noisy and reverberant environments," *Proc. Interspeech 2019*, pp. 4070–4074, 2019.

[20] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online wpe dereverberation." in *Interspeech*, 2017, pp. 384–388.

[21] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," 2016.

[22] L. Drude, C. Boeddeker, J. Heymann, R. Haeb-Umbach, K. Kinoshita, M. Delcroix, and T. Nakatani, "Integrating neural network based beamforming and weighted prediction error dereverberation." in *Interspeech*, 2018, pp. 3043–3047.

[23] J.-Y. Yang and J.-H. Chang, "Joint optimization of neural acoustic beamforming and dereverberation with x-vectors for robust speaker verification," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019, 2019, pp. 4075–4079.

[24] D. D. Dummy, "The interspeech 2020 far-field speaker verification challenge," Qin, Xiaoyi and Li, Ming and Bu, Hui and Rao, Wei and Das, Rohan Kumar and Narayanan, Shrikanth and Li, Haizhou, submitted to Interspeech 2020.

[25] S. Novoselov, A. Gusev, A. Ivanov, T. Pekhovsky, A. Shulipa, G. Lavrentyeva, V. Volokhov, and A. Kozlov, "Stc speaker recognition systems for the voices from a distance challenge," *arXiv preprint arXiv:1904.06093*, 2019.

[26] L. Burget, O. Novotnỳ, and O. Glembek, "Analysis of but submission in far-field scenarios of voices 2019 challenge," 2019.

[27] 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.

[28] 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.

[29] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.

[30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[31] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[32] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5210–5214.

[33] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[34] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[35] K. Wang, J. Zhang, S. Sun, Y. Wang, F. Xiang, and L. Xie, "Investigating generative adversarial networks based speech dereverberation for robust speech recognition," *arXiv preprint arXiv:1803.10132*, 2018.

[36] M. Wu, K. Kumatani, S. Sundaram, N. Ström, and B. Hoffmeister, "Frequency domain multi-channel acoustic modeling for distant speech recognition," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.