ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094698

STARGAN-VC BASED CROSS-DOMAIN DATA AUGMENTATION FOR SPEAKER VERIFICATION

Hang-Rui Hu[1,2]*, Yan Song*[1]*, Jian-Tao Zhang*[1]*, Li-Rong Dai*[1]*, Ian McLoughlin*[1]
Zhu Zhuo[2]*, Yu Zhou*[2]*, Yu-Hong Li*[2]*, Hui Xue*[2]

[1]National Engineering Research Center for Speech and Language Information Processing, University of Science and Technology of China, Hefei, China.
[2]Alibaba Group, China.

ABSTRACT

Automatic speaker verification (ASV) faces domain shift caused by the mismatch of intrinsic and extrinsic factors, such as recording device and speaking style, in real-world applications, which leads to severe performance degradation. Since single-speaker multi-condition (SSMC) data is difficult to collect in practice, existing domain adaptation methods are hard to ensure the feature consistency of the same class but different domains. To this end, we propose a cross-domain data generation method to obtain a domain-invariant ASV system. Inspired by voice conversion (VC) task, a StarGAN based generative model first learns cross-domain mappings from SSMC data, and then generates missing domain data for all speakers, thus increasing the intra-class diversity of the training set. Considering the difference between ASV and VC task, we renovate the corresponding training objectives and network structure to make the adaptation task-specific. Evaluations on achieve a relative performance improvement of about 5-8% over the baseline in terms of minDCF and EER, outperforming the CNSRC winner's system of the equivalent scale.

Index Terms— StarGAN, Domain Adaptation, Data Augmentation, Speaker Verification

1. INTRODUCTION

Automatic speaker verification (ASV) has achieved remarkable performance after decades of research. In recent years, numerous deep neural network (DNN) techniques have attained great success in ASV tasks. To improve the compactness and discriminative capability of speaker embeddings, previous works have primarily concentrated on devising various network architectures, pooling strategies, and optimizing objectives [1, 2, 3, 4, 5, 6, 7].

In spite of the high performance on existing benchmark datasets, DNN systems often often suffer significant performance degradation in real-world applications, this can be attributed to the severe domain mismatch resulting from the complex acoustic environments and speaking styles present in real-life scenarios. Unfortunately, this mismatch cannot be easily remedied by simply collecting more data[8]. To accurately distinguish the speaker's properties from other factors in the speech signal, the training data must contain speech from the same speaker but in different acoustic environments and speaking styles, i.e., single-speaker and multi-condition (SSMC) data. In contrast, ASR training requires single-word and multi-condition (SWMC) data. It is obvious that SSMC data is much more difficult to collect than SWMC data, and existing data augmentation strategies are not enough to cover real cross domain disturbances.

For this purpose, voice conversion (VC) techniques are widely employed in ASV tasks to adapt non-linguistic variations, thereby enhancing the diversity of the data. Owing to the dearth of parallel data, several modifications have been proposed for CycleGAN [9] to augment cross-domain data, such as integrating multiple discriminators and preserving class labels in both the forward and backward cycle [10, 11, 12]. Furthermore, some researchers found that the cycle-consistency constraint is overly rigid and constrains the predictive flexibility [13, 14], and instead propose the matching of high-level depth features to compensate for the loss of semantic information [15, 9].

However, the existing techniques are inadequate in ensuring cross-domain consistency within the same category, and are also not proficient in handling multi-domain settings. In this paper, we propose a novel StarGAN[16, 17, 18] based non-parallel domain adaptation method, termed as StarGAN-Aug. The main idea is to learn cross-domain mappings from multi-domain speakers and then generate missing domain data for all speakers, thereby augmenting the intra-class variety of the training set. Unlike VC task, which emphasizes the quality of the generated samples, StarGAN-Aug pay more attention to the cross-domain perturbations. Hence, we modify the relevant training objectives and network structure to make the adaptation task-specific.

The work was conducted during Hang-Rui Hu's internship at Alibaba Group with Yan Song as the corresponding author, and was supported by the Leading Plan of CAS (XDC0830200)

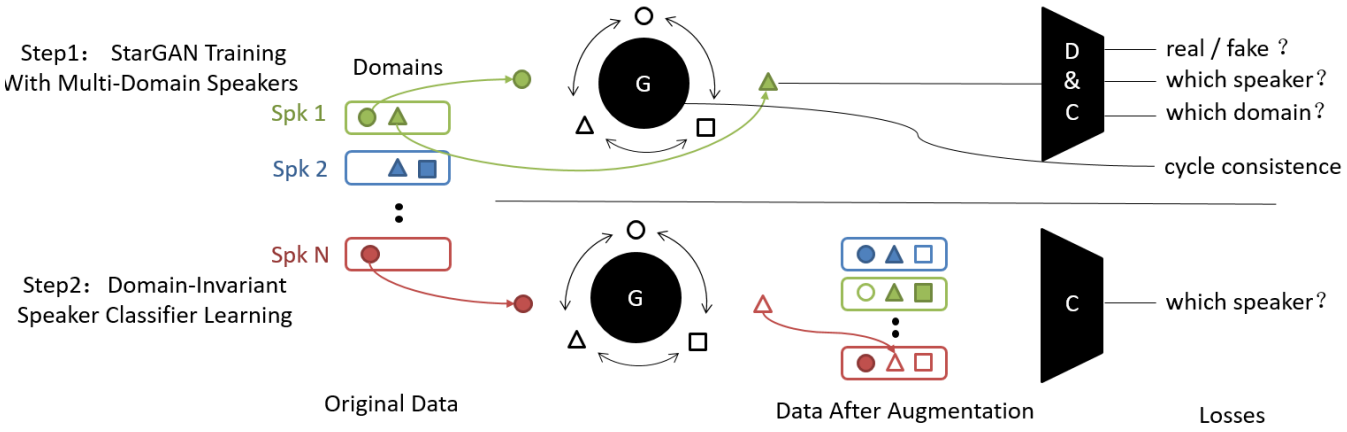

Fig. 1. Framework of the proposed StarGAN based cross-domain data augmentation (StarGAN-Aug) method. We first train a single generator G from single-speaker multi-condition data to learn how to convert an input acoustic feature into the target domain while retaining the speaker and semantic information. Then we can generate absent domain data for all speakers, thereby enhancing the cross-domain consistency of the speaker verification systems.

We evaluated the proposed approach on the CNCeleb dataset[19, 8], a new large-scale multi-domain speaker corpus. our methods achieved a relative enhancement of 5-8% in terms of Equal Error Rate (EER) compared to the baseline, surpassing the system of the CNSRC winner[20] of the equivalent scale.

2. METHODS

2.1. Overview of the proposed framework

In this paper, we propose a cross-domain data augmentation framework, termed as StarGAN-Aug, which aim to acquire knowledge on cross-domain mappings from multi-domain speakers, and then generate missing domain data for all speakers, thus boosting the intra-class diversity of the training set.

Generator: Inspired by StarGAN-VCs[17, 18], which was originally designed for multi-domain voice conversion, StarGAN-Aug employs a single generator, G, to convert an input acoustic feature into an target domain while preserving the speaker information. To achieve this, additional discriminator D and classifiers C are required to guarantee the generator's output are indistinguishable from the real target samples, and belong to the right speaker and domain. However, even with these loss functions, the generator's mapping is still is under-constrained and unstable and may discard the linguistic information of voice. Therefore, an inverse mapping is employed to reconstruct the source features from the generated ones. The full optimizing objective can be written as:

$$L_G = \lambda_{adv} L_D^{adv} + \lambda_{spk} L_C^{spk} + \lambda_{dom} L_C^{dom} + \lambda_{cycle} L^{cycle}$$

Each objective will be detailed in the Sec2.2, and λs are the corresponding loss weight.

Discriminator We employ the Patch-GAN approach to create a discriminator D, which classifies whether local segments of an input feature sequence are real or fake. In a minimax game between G and D, the discriminator tries to distinguish between real and fake samples (synthesized by G), while the generator aims to deceive the discriminator into believing that the synthesized sample is real.

Classifier: We do not directly use the original speaker classification network C_{spk} for domain classification, as it proves detrimental to procure a domain-invariant speaker network. Instead, we devise an auxiliary domain classifier C_{dom}, which employs a gated CNN to process an acoustic feature sequence x and produce a sequence of class probability distributions that measure the likelihood of each segment of x belonging to domain c. In particular, the domain classifier is speaker-specific, which leads to the calculation of class centers for each domain within the same speaker, rather than globally. This approach provides the generator with a more distinct and precise objective to follow.

2.2. Detailed Training objectives

This section details the optimization objectives of each component. Now let us denote the training set as $\{(x_i, y_i, c_i)\}$, where each training sample is a triplet with the acoustic feature x_i, its speaker label $y_i \in \{0, ..., N-1\}$ and the domain label $c_i \in \{0, ..., K-1\}$. N and K denote the total numbers of different speakers and speech conditions for training respectively. In addition, we can also count the set of known domains of each speaker $S_y = \{c \,|\, \exists\, i,\ s.t.\ y_i = y \,\&\, c_i = c\}$.

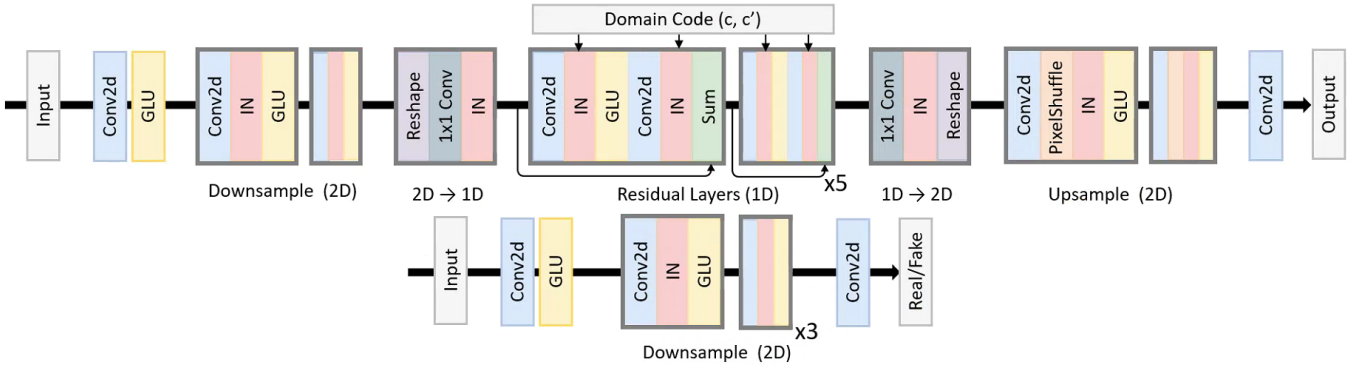

Fig. 2. Network architectures of generator G (top) and Discriminator D (bottom). The generator is fully convolutional, this allows an arbitrary length T to be input in inference. The discriminator is a fully convolutional PatchGAN architecture [14] that provides a real/fake prediction for each patch of the input FBank.

2.2.1. StarGAN-Aug Training with SSMC data

As described in 2.1, G is optimized via an adversarial loss, classification loss, and cycle-consistency loss, aiming to convert an input acoustic feature x from domain c and speaker y into an output feature x' of domain $c' \in S_y$ conditioned on the domain code (c, c') while preserving the speaker information, i.e., $x' = G(x, c, c')$. In this training step, the chosen target domain c' must appear in the current category y,

Adversarial loss: The adversarial loss endeavors to make the converted feature indistinguishable from the authentic target feature:

$$L_D^{adv} = -\mathbb{E}_{x,c}\left[log(D(x,c))\right] - \mathbb{E}_{x',c'}\left[log(1 - D(x',c'))\right]$$

$$L_G^{adv} = -\mathbb{E}_{x',c'}\left[log(D(x',c'))\right]$$

Classification loss: The speaker classification loss guarantees that G applies rational cross-domain perturbations to the sample while upholding the speaker information.

$$L_C^{cls} = L_C^{spk} + L_C^{dom} = -\mathbb{E}\left[logC(y|x)\right] - \mathbb{E}\left[logC(c|x,y)\right]$$
$$L_G^{cls} = L_G^{spk} + L_G^{dom} = -\mathbb{E}\left[logC(y|x')\right] - \mathbb{E}\left[logC(c'|x',y)\right]$$

The classification loss can adopt softmax loss and its variations, with the aim of bringing the generated sample closer to the correct class center than the others.

It is worth noting that the domain classifier is speaker-specific, resulting in more efficient and challenging non-target class centers in softmax losses.

Cycle-Consistence loss: Despite above training objects, the generator mapping remains susceptible to under-constraints and instability, and it may result in the abandonment of linguistic information from the input voice. Hence, an inverse mapping is utilized to reestablish the source features from the generated features, assuring that other information in speech, such as speech content, is not lost during the cross-domain conversion process.

$$L^{cycle} = \mathbb{E}_{x,c,c'}||G(G(x,c,c'),c',c) - x||^2$$

2.2.2. Domain-invariant SV system Learning

After the previous training step, we stochastically convert training samples into any desired domain c'', including the missing ones. Both the original and generated samples are utilized to enhance the robustness of the SV system.

$$L_C^{spk} = -\mathbb{E}_{x,y}\left[logC(y|x)\right] - \mathbb{E}_{x,y,c,c''}\left[logC(y|G(x,c,c''))\right]$$

Note that the augmented samples may contain label noise, which suggests that their impact on the training process should be less significant than that of real samples. To address this concern, we can employ lower loss weights, margins, or implement label smoothing for the augmented samples.

2.3. Model Structure

The generator model, shown in Figure 2, is comprised of downsampling, residual, and upsampling layers, following the architectural design proposed by StarGAN-VCs. The utilization of a downsampling-upsampling architecture effectively reduces computational complexity, while the incorporation of residual layers is important for addressing the issue of vanishing gradients. Additionally, it employs a 2-1-1D CNN architecture, where downsampling and upsampling blocks employ 2D convolutions, and residual blocks utilize 1D convolutions. Prior work has shown that this architecture allows the model to effectively capture a wide-range of structures and features without compromising performance [17, 18].

3. EXPERIMENTS

3.1. Experimental Settings

Experimental setup Experiments are conducted on the CNCeleb 1 and 2, which comprise over 130k utterances from 3000 Chinese celebrities across 11 diverse domains. Each training speaker may incorporate samples from multiple domains. Specifically, about 47% speakers are multi-domains.

Table 1. Cosine minDCF and EER results of the comparison systems on CNCeleb evaluation.

Backbone	System	minDCF$_{0.01}$	EER(%)
ResNet-18	Baseline	0.488	8.93
	StarGAN-Aug	**0.451**	**8.35**
ResNet-34	Baseline	0.415	7.92
	StarGAN-Aug	**0.393**	**7.49**
	Top1 [20]	0.395	7.98
ResNet-293	Top1 Fusion [20]	**0.297**	**4.91**

The feature extraction process uses the Kaldi toolkit [21]. We introduced noise and reverberation from MUSAN and RIR corpus to each utterance, followed by 40 or 80 dimensional FBank extraction using 25ms windows and 10ms frame shifts. A sliding window of 3s applied mean-normalization and a voice activity detection (VAD) technique removed silent segments. Random truncation into 4-8s short slices was applied to the training set features. Note that we did not apply speed perturbation to generate additional speakers.

Implementation configuration: C_{spk} and C_{dom} for speaker and domain classification use the ResNet backbone as in [5], and the generator G and discriminator D use StarGAN-VCs model detailed in Sec2.3. C_{spk} optimizes via AAM-Softmax [6] loss, with a scale of 30 and a margin of 0.2 for original data, and 0.1 for augmentation data. C_{dom} optimizes via standard Softmax loss. We gradually increased the frame size from 400 to 800, and also increased the margin from 0.2 to 0.4. All scores are normalized with AS-Norm [22], using the top 400 imposter scores.

We jointly trained the four models using an Adam optimizer until convergence, with a batch size of 128. For each sample, we randomly select a target domain code with equal probability. The learning rates for G and D were set to 0.0002 and 0.0001, respectively, with momentum terms of (0.05, 0.999), and the learning rate for C varied between 1e-8 and 1e-3 with a cyclical learning rate [23]. Additionally, we used λ_{cycle}=5, λ_{spk}=λ_{dom}=3, and λ_{adv}=1 to adjust the balance between different optimization objectives. Note that these loss weights differ slightly from the original VC settings, where λ_{cycle}=10 and λ_{dom}=1, as we prioritize cross-domain perturbations over the quality of generated samples.

3.2. Results

The performance was evaluated by the minimum Decision Cost Function (MinDCF) and Equal Error Rate (EER) and required by CN-Celeb Speaker Recognition Challenge(CNSRC) 2022.

Main Results: The main results are reported in Table 1, which compares against the our baseline system and the winner of the CNSRC. From the results, it can be seen that the proposed data augmentation methods achieves a relative minDCF reduction of about 8% on ResNet-18 backbone. Furthermore, on the deeper and wider ResNet-34, our approach also achieved a relative improvement of approximately 5%, outperforming the winner's system of the CNSRC 2022 of the equivalent scale. Interestingly, the improvement achieved on ResNet-34 is slightly lower than that on ResNet-18. This may be attributed to the fact that larger models already possess a stronger ability to integrate cross-domain information in the original data. It is worth noting that the CNSRC winner achieved significantly better results through the fusion of large-scale models such as ResNet-292. We posit that our method can help to prevent overfitting of the model on such a large scale. Due to the computational complexity, we leave this to the future work.

Table 2. MinDCF$_{0.01}$ results of different data augmentation methods on ResNet-34 backbone

Noise	Reverb	StarGAN-Aug	minDCF$_{0.01}$
✗	✗	✗	0.441
✓	✓	✗	0.415
✗	✗	✓	0.421
✓	✓	✓	0.395

Ablation Results: The results of different data augmentation methods are shown in Figure 2. We first verified the traditional augmentation methods,where the MUSAN corpus's additive noise and room impulse response (RIR) simulation [22] were leveraged as data augmentation for each utterance with a likelihood of 0.6. We can observe from the first two rows that these methods significantly enhance the ASV system's performance by emulating diverse sampling environments' impact on sound perception. However, traditional methods fail to effectively cover speech distortion resulting from distinct speaking styles and occasions within a single speaker.

Thanks to the unified modeling of multi domain mappings from the generation viewpoint, StarGAN-Aug yields additional gains that complement traditional methods. However, due to the inadequacy of single-speaker multi-domain data in CNCeleb and the suboptimal quality of domain labels, the proposed approach did not yield satisfactory results on CNCeleb data. We may look for more high-quality SSMC data for general cross-domain mapping training in the future, and we believe that the performance of the methods will also be further improved.

4. CONCLUSION

This paper has proposed a novel cross-domain data augmentation method, called StarGAN-Aug, for unsupervised non-parallel domain adaptation. The main idea is to learn cross-domain mappings from multi-domain speakers and generate absent domain data for all speakers, thereby enhancing the intra-class diversity of the training set. The corresponding training objectives and network structure are modified to make the adaptation task-specific. Experimental results have demonstrated the superiority of the proposed method.

5. REFERENCES

[1] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, 2015, pp. 3214–3218.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey *et al.*, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[3] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *Proc. Interspeech 2020*, pp. 3830–3834, 2020.

[4] Y. Liu, Y. Song, I. McLoughlin, L. Liu, and L.-r. Dai, "An effective deep embedding learning method based on dense-residual networks for speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6683–6687.

[5] Y. Liu, Y. Song, Y. Jiang, I. McLoughlin, L. Liu, and L. Dai, "An effective speaker recognition method based on joint identification and verification supervisions," in *INTERSPEECH*, 2020, pp. 3007–3011.

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[7] Z. Huang, S. Wang, and K. Yu, "Angular softmax for short-duration text-independent speaker verification." in *Interspeech*, 2018, pp. 3623–3627.

[8] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T. F. Zheng, and D. Wang, "Cn-celeb: multi-genre speaker recognition," *Speech Communication*, vol. 137, pp. 77–91, 2022.

[9] S. Kataria, J. Villalba *et al.*, "Deep feature cyclegans: Speaker identity preserving non-parallel microphone-telephone domain adaptation for speaker verification," in *INTERSPEECH*, 2021, pp. 1079–1083.

[10] P. S. Nidadavolu, J. Villalba, and N. Dehak, "Cycle-gans for domain adaptation of acoustic features for speaker recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6206–6210.

[11] P. S. Nidadavolu, S. Kataria, J. Villalba, and N. Dehak, "Low-resource domain adaptation for speaker recognition using cycle-gans," in *ASRU*. IEEE, 2019, pp. 710–717.

[12] E. Hosseini-Asl, Y. Zhou, C. Xiong, and R. Socher, "A multi-discriminator cyclegan for unsupervised non-parallel speech domain adaptation," *Proc. Interspeech 2018*, pp. 3758–3762, 2018.

[13] Y. Zhao, R. Wu, and H. Dong, "Unpaired image-to-image translation using adversarial consistency loss," in *ECCV*. Springer, 2020, pp. 800–815.

[14] E. Hosseini-Asl, Y. Zhou, C. Xiong, and R. Socher, "Augmented cyclic adversarial learning for low resource domain adaptation," *arXiv preprint arXiv:1807.00374*, 2018.

[15] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[16] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.

[17] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[18] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion," *Proc. Interspeech 2019*, pp. 679–683, 2019.

[19] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "Cn-celeb: a challenging chinese speaker recognition dataset," in *ICASSP*. IEEE, 2020, pp. 7604–7608.

[20] Z. Chen, B. Liu, B. Han, L. Zhang, and Y. Qian, "The sjtu x-lance lab system for cnsrc 2022," *arXiv preprint arXiv:2206.11699*, 2022.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, and N. G. et al, "The kaldi speech recognition toolkit," in *ASRU*, 2011.

[22] S. Cumani, P. D. Batzu, D. Colibro, C. Vair, P. Laface, and V. Vasilakakis, "Comparison of speaker recognition approaches for real applications." in *INTERSPEECH*, 2011, pp. 2365–2368.

[23] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2017, pp. 464–472.