Multi-level Adversarial Training with Data Augmentation for Robust Speaker Verification

Xiaolei Zhang[1,2] Zhihua Fang[2,3] Liang He[2,3,4,⋆]
[1]School of Software, Xinjiang University, Urumqi 830091, China
[2]Xinjiang Multimodal Information Technology Engineering Research Center, Urumqi 830017, China
[3]School of Computer Science and Technology, Xinjiang University, Urumqi 830017, China
[4]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
⋆Corresponding author (heliang@mail.tsinghua.edu.cn)

***Abstract*—Data augmentation (DA) is extensively utilized in deep speaker models and significantly enhances their performance. This process not only enriches the training dataset but also enables deep neural networks to acquire more robust and generalizable features. However, the diverse augmentations from conventional data augmentation methods can lead to unwanted distortions. To tackle this problem, this study introduces a new strategy named multi-level adversarial training with data augmentation (MAT-DA). In particular, both an augmentation discriminator and an augmentation-type discriminator are incorporated at the frame level, while an augmentation discriminator is incorporated at the embedding level to classify various augmentation types. Additionally, we impose an extra constraint on speaker embedding to help the network achieve more robust and generalizable speaker embeddings when encountering different acoustic variations. Experiments conducted on VoxCeleb dataset indicate that the proposed method outperforms standard DA under both augmentation-matched and mismatched test conditions and the results show a performance improvement of up to** 13.7%**.**

I. Introduction

Automatic speaker verification (ASV), which aims to verify a speaker's claimed identity based on their vocal characteristics, has witnessed significant advancements, largely due to the emergence and development of deep neural network (DNN)-based speaker embeddings [1]. Over the years, the accumulation of speech data and the prevalence of DNN-based speaker embedding models have propelled current ASV systems to remarkable progress [2]–[4]. With continuous efforts to improve architectures to extract superior speaker embeddings and optimize loss functions to improve discrimination, deep embedding models have achieved state-of-the-art performance in numerous ASV evaluation tasks.

Despite significant advancements, current ASV systems still face substantial robustness challenges in real-world applications. A key challenge is the complex interplay between speaker characteristics and diverse acoustic variations such as background noise, music, and multi-speaker conversations. These variations can cause unpredictable shifts in speaker embedding models and performance degradation [5]–[8]. To address this issue, data augmentation strategies have been proposed. Due to their ease of implementation and remarkable effectiveness, they have become one of the most widely used techniques in recent years. DA aims to expand the quantity and diversity of training data by simulating complex acoustic variations. Current DA methods can be broadly categorized into two types. One category involves manipulating the original speech signal, such as adding additive noise and reverberation [9], [10], applying speed perturbation [11], [12], and inducing volume perturbation [13]. The other category focuses on augmenting spectrograms by applying random masking in the temporal and frequency domains [14], [15]. Training speaker recognition systems with both raw and augmented data typically produces good results. However, excessive augmented data can cause augmentation residual issues. Specifically, speaker embeddings derived from data augmented by certain methods may suffer from systematic distortions, which degrade the model's generalization ability and lead to a sharp decline in the SV system's performance when facing unseen acoustic variations.

Over the past few years, domain adversarial learning (DAT) [16], which implicitly reduces distinctions among diverse domain data through a min-max two-player game, has shown great success in speech-related tasks. Chen *et al.* [17] learned channel-invariant and speaker-discriminative representations through adversarial training. Qin *et al.* [18] proposed age-agnostic adversarial learning to obtain age-invariant and speaker-discriminative representations. Xing *et al.* [19] proposed a joint approach introducing DAT on the basis of noise disentanglement, helping the speaker verification system establish a noise-agnostic, speaker-invariant embedding space and thereby enhancing the robustness of speaker verification.

In this paper, we propose a novel training strategy called multi-level adversarial training with data augmentation (MAT-DA). We regard the raw data and augmented data as originating from different domains. Furthermore, for the augmented data, we subdivide them into different subdomains based on the types of augmentation. After feature extraction at both the frame level and the embedding level of the backbone network, we introduce gradient reversal layers, coupled with domain discriminators and subdomain discriminators of different architectures, to conduct multi-task learning. The domain discriminator is used to distinguish whether the data is raw, while the subdomain discriminator is used to identify the specific acoustic types of the augmented data. The two-stage multi-granularity adversarial training can effectively mitigate

979-8-3315-7206-8/25/$31.00

the distortions and alterations of voiceprint features caused by different data augmentation techniques. This approach enables the speaker embedding system to obtain more intrinsic and robust speaker embeddings, maintaining robustness when facing various acoustic variations. In addition, at the speaker embedding level, we introduce an additional constraint to reduce the negative impact of augmentation residuals further.

Our experiments initially employed the VoxCeleb dataset [20], incorporating noise, speech, and music from the MUSAN dataset [21] for data augmentation. The results indicate that under training-matched augmentation conditions, our proposed MAT-DA method outperforms the standard data augmentation method in terms of robustness. Furthermore, we used the singing and interview data from the CN-Celeb evaluation set [22] to verify the generalizability of the MAT-DA method under unseen augmented test conditions. Finally, we also employed the SpecAugment [14] method on the VoxCeleb dataset to validate the effectiveness of the MAT-DA method for different types of data augmentation. The experimental results consistently demonstrate that our proposed MAT-DA method has a performance advantage under these diverse augmented test conditions, highlighting its excellent generalization capability against acoustic variations.

II. Related Work

A. TDNN for Deep Speaker Embedding

The predominant deep neural networks employed for speaker embedding extraction encompass Residual Neural Networks (ResNet), Time-Delay Neural Networks (TDNN), and Convolutional Neural Networks (CNN). In this study, we opted for ECAPA-TDNN [2] to extract speaker embeddings. ECAPA-TDNN enhances the traditional TDNN by incorporating several key components: 1.TDNN Blocks : Capture local contextual information through time-delayed filters. 2.SE-Res2Blocks: Combine Squeeze-and-Excitation (SE) mechanisms with Res2Net structures to capture multi-scale features. 3. Statistics Pooling Layer: Aggregates frame-level features into a fixed-length vector. 4. Channel Attention Module: Enhances feature discriminability by emphasizing important channels. 5. Skip Connections: Facilitate information propagation across layers. The speaker embeddings are then extracted via a fully connected layer. In this study, the frame-level features are the outputs from each SE-Res2Block module.

B. Data Augmentation Adversarial Training

Data Augmentation Adversarial Training (DA-AT) integrates data augmentation with adversarial training. It enriches training data and eases residual issues from diverse augmentations. DAT uses a Gradient Reverse Layer (GRL) in a multi-task framework to learn class-discriminative and domain-invariant features. Though initially proposed for visual domains, DAT was first applied to unsupervised domain adaptation in speaker verification. It addresses distribution mismatches between source and target domains. Later, it was extended to multi-domain speaker verification. The goal is to create a domain-invariant speaker representation for better cross-domain robustness.

Data Augmentation (DA) mimics diverse channel conditions by altering raw data with specific acoustic variations to create more diverse datasets. This helps the learned embeddings become more robust to channel variance, yet there is no explicit constraint on how data augmentation affects speaker embeddings. Kang *et al.* [23] applied data augmentation and adversarial training to self-supervised speaker recognition. They used data augmentation and negative sampling in contrastive learning to extract speaker embeddings. Zhou *et al.* [24] developed an adversarial data augmentation (A-DA) strategy for supervised speaker verification, which also achieved success. These studies explicitly guide the network via adversarial training to learn augmentation-insensitive speaker representations, making the learned speaker embeddings more robust. However, these studies only apply adversarial training directly to the speaker embedding layer. This study further explores adding GRLs to frame-level speaker features with temporal dimensions to determine whether this approach facilitates obtaining speaker-discriminative and augmentation-invariant speaker embeddings. Experimental results show that frame-level features positively contribute to embedding robustness in adversarial training and can be combined with the embedding layer to achieve better performance.

III. Proposed Methods

The proposed data augmentation method, based on a multi-layer adversarial training framework, comprises three modules: backbone network, a frame-level adversarial training module, and an embedding-level adversarial training module, as shown in Figure 1. The frame-level adversarial training module contains a binary domain classifier $g_{\mathrm{f}}^{1}(\cdot)$ to detect augmentations and a multi-domain classifier $g_{\mathrm{f}}^{2}(\cdot)$ to identify augmentation types. The embedding-level adversarial training module includes only a binary domain classifier $g_{\mathrm{e}}^{1}(\cdot)$ to detect augmentations. These discriminators are connected to the backbone network via GRLs, preventing the backbone from learning augmentation-induced acoustic variations and promoting robust and essential speaker embeddings.

A. Frame-level Adversarial Training Module

To address the issue that data augmentation may corrupt the speaker-specific traits in clean speech, we propose MAT-DA to help the speaker verification model extract pure speaker information from augmented speech. Unlike speaker embeddings, frame-level features carry more speaker-irrelevant information. Training a domain classifier with traditional fully connected networks requires a large number of parameters and offers poor classification performance. We use the standard ResNet-18 architecture as the frame-level domain classifiers. During the training of the frame-level sub-domain classifier, we only use augmented data, treat different data augmentation types as distinct sub-domains, and perform multi-class training with augmentation labels such as $Augment_1$, $Augment_2$, $Augment_3$, and so on. During the training of the frame-level binary domain

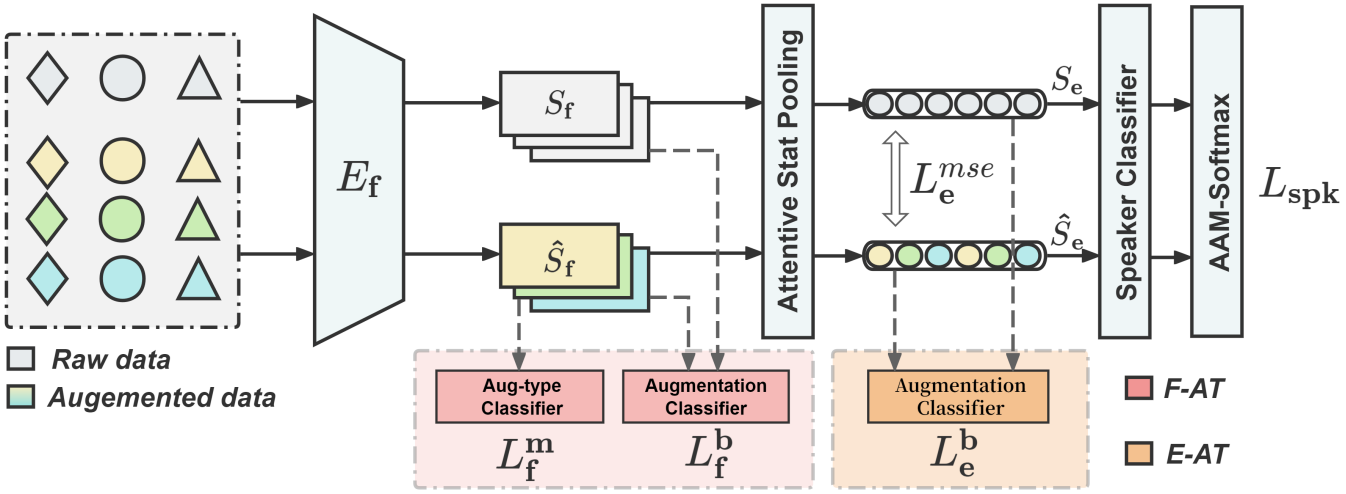

Fig. 1. Illustration on the training strategy of our proposed multi-level adversarial training with data augmentation method. The data represented in gray and color in the figure correspond to the raw and augmented data, and different colors indicating different types of augmentation.

classifier, we treat whether data augmentation is performed as different domains, and conduct binary classification training with $raw/augmented$ labels for original and augmented data.

During backpropagation, we employ a gradient reversal layer to force the backbone to produce frame-level features that the domain classifier cannot effectively distinguish whether data augmentation has been applied and what type of augmentation was used. This approach suppresses the augmentation residuals in the early stage of speaker embedding extraction, thereby rendering the frame-level features stable against various data augmentation methods. Consequently, the model can learn augmentation-invariant intermediate features.

Initially, both the clean frame-level speaker features S_{f} and the augmented frame-level speaker features $\hat{S}_{\mathrm{f}}$ are fed into the augmentation discriminator. By employing a gradient reversal layer, the speaker encoder E_{f} is supervised to generate an augmentation-invariant speaker intermediate feature without losing speaker information. The adversarial cost function L_{f}^{b} is defined as the cross-entropy:

$$L_{\mathrm{f}}^{b} = -\frac{1}{2N}\sum_{i=1}^{2N} p_i \cdot \log(\mathrm{Softmax}(g_{\mathrm{f}}^{1}(S_{\mathrm{a}}^{i}))) \tag{1}$$

where N is the batch size, p_i is the augmentation label of the i-th utternace, $g_{\mathrm{f}}^{1}(\cdot)$ is a frame-level multi-domain classifier and S_a is the set of S_{f} and $\hat{S}_{\mathrm{f}}$.

Furthermore, the augmented frame-level speaker features $\hat{S}_{\mathrm{f}}$ are also fed into the augmentation type discriminator. This aims to alleviate the residual issues in speaker intermediate features caused by various data augmentation methods. The specific loss function is defined as follows:

$$L_{\mathrm{f}}^{m} = -\frac{1}{N}\sum_{i=1}^{N} q_i \cdot \log(\mathrm{Softmax}(g_{\mathrm{f}}^{2}(\hat{S}_{\mathrm{f}}^{i}))) \tag{2}$$

where q_i is the augmentation type label of the i-th utterance, and $g_{\mathrm{f}}^{2}(\cdot)$ is a frame-level binary domain classifier.

B. Embedding-level Adversarial Training Module

However, frame-level adversarial learning cannot completely eliminate the degradation of speaker traits caused by data augmentation, as the final speaker embeddings are required for comparison in speaker verification tasks. Therefore, at the embedding layer, we employ a traditional fully connected domain discriminator and connect it to the backbone network via a gradient reversal layer to more directly constrain the domain invariance of speaker embeddings, obtaining class-discriminative and domain-invariant speaker embeddings. The loss function is defined as follows:

$$L_{\mathrm{e}}^{b} = -\frac{1}{2N}\sum_{p=1}^{2N} p_i \cdot \log(\mathrm{Softmax}(g_{\mathrm{e}}^{1}(\hat{S}_{\mathrm{a}}^{i}))) \tag{3}$$

where $\hat{S}_{\mathrm{a}}$ is the set of S_{e} and $\hat{S}_{\mathrm{e}}$, and $g_{\mathrm{e}}^{1}(\cdot)$ is a embedding-level binary domain classifier.

Additionally, the Mean Squared Error (MSE) loss is used as an additional constraint to minimize the distance between clean speaker embeddings S_{e} and augmented speaker embeddings $\hat{S}_{\mathrm{e}}$. This further ensures that the obtained speaker embeddings will not be compromised by the introduction of the domain adversarial learning strategy, thus preserving the clean speaker information. The loss function is defined as follows:

$$L_{\mathrm{mse}} = \frac{1}{N}\sum_{i=1}^{N}(S_{\mathrm{e}}^{i} - \hat{S}_{\mathrm{e}}^{i})^2 \tag{4}$$

Then, S_{e} and $\hat{S}_{\mathrm{e}}$ are fed into the speaker classifier simultaneously, to calculate the classification loss using Angular

TABLE I
EER(%) AND MINDCF(0.01) RESULTS ON VOXCELEB1 UNDER DIFFERENT TEST CONDITIONS.

Method	Aug	Mse	Vox1-O (raw)		Vox1-O (noise)		Vox1-O (speech)		Vox1-O (music)		Vox1-O (all)	
			EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
ECAPA-TDNN	×	–	3.950	0.2530	12.041	0.6204	8.610	0.5056	7.963	0.47312	8.976	0.4956
	✓	–	3.101	0.2022	5.567	0.3611	4.793	0.3272	4.395	0.2939	4.809	0.3209
+E-AT	✓	–	2.979	0.2154	5.562	0.3559	4.708	0.3282	4.310	0.2856	4.676	0.3158
+F-AT(S_1)	✓	–	2.969	0.2044	5.254	0.3408	4.711	0.3358	4.178	0.2799	4.570	0.2939
+F-AT(S_2)	✓	–	2.921	0.2148	5.238	0.3554	4.602	0.3057	4.252	0.3020	4.612	0.3034
+F-AT(S_3)	✓	–	2.903	0.2027	5.514	0.3375	4.671	0.3037	4.167	0.2731	4.533	0.2879
MAT-DA(S_1)	✓	×	3.006	0.1969	5.376	0.3197	4.533	0.3109	4.262	0.2734	4.429	0.3071
MAT-DA(S_2)	✓	×	2.900	0.2027	5.328	0.3262	4.411	0.3083	4.061	0.2779	4.319	0.3036
MAT-DA(S_3)	✓	×	2.852	0.1945	5.270	0.3193	4.390	0.3015	4.008	0.2741	4.471	0.2979
MAT-DA(S_1)	✓	✓	2.826	0.2011	5.259	**0.3189**	4.637	0.3496	4.162	**0.2659**	4.284	**0.2842**
MAT-DA(S_2)	✓	✓	2.783	0.1977	5.212	0.3363	4.443	0.3105	4.103	0.2882	4.321	0.3017
MAT-DA(S_3)	✓	✓	**2.757**	**0.1868**	**4.936**	0.3291	**4.384**	**0.3001**	**3.918**	0.2715	**4.151**	0.2884

Additive Margin Softmax(AAM-Softmax):

$$L_{\text{spk}} = -\frac{1}{2N}\sum_{i=1}^{2N}\log\frac{e^{s\cdot\cos(\theta_{y_i}+m)}}{e^{s\cdot\cos(\theta_{y_i}+m)}+\sum_{j=1,j\neq y_i}^{C}e^{s\cdot\cos(\theta_j)}} \tag{5}$$

where y_i represents the speaker label of the i-th utterance, s and m are two hyperparameters for AAM-Softmax.

In summary, our proposed method, through multi-level adversarial training combined with data augmentation strategies, can maximize the incentive for the backbone B to learn a stable speaker embedding space that is not distorted by data augmentation methods. The expression for the entire loss function L is shown as follows:

$$L = L_{\text{spk}} + L_{\text{mse}} - \lambda(L_{\text{f}}^{b} + L_{\text{f}}^{m} + L_{\text{e}}^{b}) \tag{6}$$

where λ is the weight coefficient controlling multi-objective optimization.

IV. EXPERIMENTAL SETTINGS

A. Datasets

1) VoxCeleb: This is a large-scale speaker dataset collected by the University of Oxford, UK. Following the common experiment settings, experiments are conducted on the VoxCeleb1 dataset. The development set contains 148642 utterances from 1211 speakers. The test set contains 4874 utterances from 40 speakers, which constitutes 37720 test trials. Although this dataset is not strictly in clean conditions, we assumed the original data to be a clean dataset and generated augmented data based on this raw data.

2) CN-Celeb: This is a multi-genre speaker dataset collected by Tsinghua University. We used its standard development set, which consists of 797 speakers from 11 diverse genres, for unseen data augmentation test. CN-Celeb1 can be used to validate the generalizability of the models, because of its acoustic characteristics.

3) MUSAN: The MUSAN database was used to sample interference signals for data augmentation, which contains 60 hours of speech, 42 hours of music, and 6 hours of assorted noise. The MUSAN dataset is divided into two non-overlapping subsets for generating augmented training and testing utterances, respectively.

B. Implementation Details

The input features are 80-dimensional log mel spectrogram features from a 25 ms window with a 10 ms frame shift, which are normalized through cepstral mean subtraction, and no voice activity detection is applied. During the training stage, 2s segments are randomly selected from each original utterance. One clean and one augmented utterance per 100 randomly selected speakers, totaling 200 utterances, are grouped as one batch and fed into the systems. All systems are trained using AAM-softmax with a margin of 0.2 and a scaling factor of 30, except that the loss function for the domain classifier is defined as cross-entropy. For optimization, the Adam optimizer with an initial learning rate of 0.001, a learning rate decay of 0.97, and the weight decay of $2e^{-5}$ is used to train the whole network.

The ECAPA-TDNN network serves as the speaker embedding extractor in this study, leveraging its straightforward architecture and incorporating 512 channels within the convolutional frame layers. Upon completion of the training phase, the network generates 192-dimensional speaker embeddings. During the testing phase, the entire utterance is utilized to derive the speaker embeddings. The cosine similarity metric is employed for scoring purposes. Furthermore, the performance of the system is evaluated using both the equal error rate (EER) and the minimum detection cost function (minDCF) as metrics.

V. RESULTS AND DISCUSSION

A. Main Results

In this section, the performance comparison between the baseline and the introduction of GRL at various stages of the model is presented in Table I. Experiments were conducted on the ECAPA-TDNN baseline using a variety of adversarial training methods. Four sets of systems were constructed: one without adversarial training, trained with raw data and standard data augmentation, one with adversarial training introduced

solely at the embedding layer, one with adversarial training introduced solely at the frame level, and one employing our proposed Multi-level Adversarial Training with Data Augmentation approach.

The initial set of experiments demonstrates that standard data augmentation methods significantly enhance performance. In the subsequent experiments, adversarial training was introduced to mitigate the adverse effects of data augmentation. Firstly, a GRL module was incorporated in the embedding layer to assess the efficacy of direct adversarial training for the domain classification of speaker embeddings. Although this led to some improvement, the impact was modest rather than substantial. In the following two sets of experiments, intermediate frame-level features(S_i is the output of the i-th SE-Res2Block in ECAPA-TDNN) were introduced for adversarial training. It was observed that even with the introduction of frame-level features alone, the model's performance was still moderately enhanced, indicating that controlling domain invariance of intermediate features through GRL is beneficial for obtaining stable speaker embeddings in speaker verification systems. In the fourth set of experiments, it was observed that the integration of multiple GRLs at different stages further improved system performance over the baseline in various scenarios. This observation underscores the inherent complementarity between Embedding-level Adversarial Training (E-AT) and Frame-level Adversarial Training (F-AT), highlighting their ability to remove nonspeaker-related information at different hierarchical levels.

Overall, it was found that the selection of different frame-level intermediate features also had a certain impact on model performance. Additionally, the intermediate features closer to the Statistics Pooling Layer appeared to be more advantageous, possibly because the gradients propagated through GRL also brought beneficial improvements to the lower-level modules.

TABLE II
EER(%) AND MINDCF(0.01) RESULTS ON VOXCELEB1 UNDER UNSEEN AUGMENTATIONS.

Method	Aug	Vox1-O (singing)		Vox1-O (interview)	
		EER	minDCF	EER	minDCF
ECAPA-TDNN	×	9.957	0.5755	8.356	0.5115
	✓	5.323	0.3581	4.936	0.3322
+E-AT	✓	5.169	0.3433	4.692	0.3152
+F-AT(S_3)	✓	5.084	**0.3293**	4.586	0.3077
MAT-DA(S_3)	✓	**5.068**	0.3372	**4.374**	**0.3034**

B. Further Analysis

To further validate the effectiveness of our proposed MAT-DA method, we conducted a series of performance evaluations under unseen augmentation types and more complex test conditions. The experimental results are presented in Table II. On the one hand, we utilized the singing and interview data from the multi-genre CN-Celeb evaluation set to introduce additive noise in the Vox Celeb test set. This was used to test the robustness of the model against unseen additive augmentation variations, with EER and minDCF serving as performance metrics. Obviously, it can be seen that both E-AT and F-AT outperform the baseline under these more complex test conditions, providing further evidence for the effectiveness of the data augmentation method.

TABLE III
EER(%) AND MINDCF(0.01) RESULTS ON VOXCELEB1 UNDER SPECAUGMENT.

Method	Spec	Vox1-O (raw)		Vox1-O (Spec)	
		EER	minDCF	EER	minDCF
ECAPA-TDNN	×	3.950	0.2530	8.130	0.5143
	✓	3.416	0.2339	4.373	0.2936
+E-AT	✓	3.294	**0.2012**	4.423	0.2980
+F-AT(S_3)	✓	3.337	0.2509	4.439	0.3182
MAT-DA(S_3)	✓	**3.176**	0.2202	**4.350**	**0.2865**

However, SpecAugment is applied on the log mel spectrogram of the samples, where 10 to 20 channels in the frequency domain and 10 to 20 frames in the time domain are randomly masked. The experimental results are presented in Table III. This was designed to assess the generalizability of the model to different data augmentation methods. We can observe that for both unseen additive augmentation variations and the test condition of the total different augmentation method, MAT-DA still achieves a consistent performance advantage compared to the traditional data augmentation method. This indicates that the MAT-DA (Multi-level Adversarial Training with Data Augmentation) approach possesses superior robustness and generalization capabilities compared to the direct application of standard data augmentation techniques.

VI. CONCLUSIONS

In this paper, we propose a novel method that combines multi-level adversarial training with data augmentation (MAT-DA), aiming to enhance the robustness and generalization of speaker verification systems. By introducing adversarial training at both the frame level and the embedding level, we were able to effectively mitigate the negative impacts of data augmentation and obtain more stable speaker embeddings under various acoustic variations. Experimental results demonstrate that our approach outperforms traditional data augmentation methods on the VoxCeleb dataset. Furthermore, we found that applying adversarial training at different hierarchical levels can further enhance the model's adaptability to unseen acoustic variations. This indicates that our MAT-DA method not only improves the performance of speaker verification systems but also enhances their applicability in real-world applications. Future work will focus on further optimizing our multi-level adversarial training strategy and exploring a wider variety of data augmentation methods to enhance the performance of speaker verification systems further.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. 62366051).

REFERENCES

[1] H. Zhang, L. Wang, K. A. Lee, M. Liu, J. Dang, and H. Meng, "Meta-Generalization for domain-invariant speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1024–1036, 2023.

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

[3] Y. Zhang *et al.*, "MFA-Conformer: Multi-scale feature aggregation conformer for automatic speaker verification," in *Interspeech 2022*, 2022, pp. 306–310.

[4] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "CAM++: A fast and efficient network for speaker verification using context-aware masking," in *Interspeech 2023*, 2023, pp. 5301–5305.

[5] R. Peri, M. Pal, A. Jati, K. Somandepalli, and S. Narayanan, "Robust Speaker Recognition Using Unsupervised Adversarial Invariance," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6614–6618.

[6] J. Li, J. Han, and H. Song, "Gradient regularization for noise-robust speaker verification," in *Interspeech 2021*, 2021, pp. 1074–1078.

[7] Y. Wu, L. Wang, K. A. Lee, M. Liu, and J. Dang, "Joint feature enhancement and speaker recognition with multi-objective task-oriented network," in *Interspeech 2021*, 2021, pp. 1089–1093.

[8] Z. Wang, Z. Fang, and L. He, "Stable extended u-net for noise-robust speaker verification," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[9] M. Mohammad Amini and D. Matrouf, "Data augmentation versus noise compensation for x-vector speaker recognition systems in noisy environments," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1–5.

[10] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[11] H. Yamamoto, K. A. Lee, K. Okabe, and T. Koshinaka, "Speaker augmentation and bandwidth extension for deep speaker embedding," in *Interspeech 2019*, 2019, pp. 406–410.

[12] Z. Chen, B. Han, X. Xiang, H. Huang, B. Liu, and Y. Qian, "Build a SRE challenge system: Lessons from VoxSRC 2022 and CNSRC 2022," in *Interspeech 2023*, 2023, pp. 3202–3206.

[13] C.-L. Huang, "Exploring effective data augmentation with tdnn-lstm neural network embedding for speaker recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2020, pp. 291–295.

[14] D. S. Park *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*, 2019, pp. 2613–2617.

[15] S. Wang, J. Rohdin, O. Plchot, L. Burget, K. Yu, and J. Černocký, "Investigation of specaugment for deep speaker embedding learning," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7139–7143.

[16] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[17] Z. Chen, S. Wang, Y. Qian, and K. Yu, "Channel invariant speaker embedding learning with joint multi-task and adversarial training," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6574–6578.

[18] X. Qin, N. Li, W. Chao, D. Su, and M. Li, "Cross-age speaker verification: Learning age-invariant speaker embeddings," in *Interspeech 2022*, 2022, pp. 1436–1440.

[19] X. Xing, M. Xu, and T. F. Zheng, "A joint noise disentanglement and adversarial training framework for robust speaker verification," in *Interspeech 2024*, 2024, pp. 707–711.

[20] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech 2017*, 2017, pp. 2616–2620.

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

[22] Y. Fan *et al.*, "CN-Celeb: A challenging chinese speaker recognition dataset," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7604–7608.

[23] J. Kang, J. Huh, H. S. Heo, and J. S. Chung, "Augmentation adversarial training for self-supervised speaker representation learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1253–1262, 2022.

[24] Z. Zhou, J. Chen, N. Wang, L. Li, and D. Wang, "Adversarial data augmentation for robust speaker verification," in *Proceedings of the 2023 9th International Conference on Communication and Information Processing*, 2024, pp. 226–230.