INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Improving Generalization Ability of Countermeasures for New Mismatch Scenario by Combining Multiple Advanced Regularization Terms

Chang Zeng[1,2], *Xin Wang*[1], *Xiaoxiao Miao*[1], *Erica Cooper*[1], *Junichi Yamagishi*[1,2]

[1]National Institute of Informatics, Japan [2]SOKENDAI, Japan

{zengchang, wangxin, xiaoxiaomiao, ecooper, jyamagis}@nii.ac.jp

Abstract

The ability of countermeasure models to generalize from seen speech synthesis methods to unseen ones has been investigated in the ASVspoof challenge. However, a new mismatch scenario in which fake audio may be generated from real audio with unseen genres has not been studied thoroughly. To this end, we first use five different vocoders to create a new dataset called CN-Spoof based on the CN-Celeb1&2 datasets. Then, we design two auxiliary objectives for regularization via meta-optimization and a genre alignment module, respectively, and combine them with the main anti-spoofing objective using learnable weights for multiple loss terms. The results on our cross-genre evaluation dataset for anti-spoofing show that the proposed method significantly improved the generalization ability of the countermeasures compared with the baseline system in the genre mismatch scenario.

Index Terms: anti-spoofing, generalization, multi-task learning, DeepFake detection

1. Introduction

With the development of deep neural networks, DeepFake detection has attracted much attention from academia and industry since synthesized media, such as video and speech by artificial intelligence, has brought huge risks. The ASVspoof challenge series [1, 2, 3, 4] has been proposed to explore and promote the investigation of anti-spoofing for automatic speaker verification (ASV) systems. The Audio Deep Synthesis Detection (ADD) challenge [5] was held to study countermeasure models in a scenario with low-quality audio. The generalization ability in anti-spoofing is an important topic, and typical distribution shifts between training and testing data, such as unseen synthesis methods [3] and unseen acoustic environments [4], have been investigated in the challenges. However, there is room for further investigations.

Here, we focus on a new mismatch scenario related to the audio genre, in which fake audio may be generated from real audio with unseen genres because the effects of the speaker's styles and intrinsic factors associated with the specific genre have not been studied thoroughly in anti-spoofing. The definition of audio genre in our study is the same as [6, 7]. To analyze the effects and propose a more robust system, we utilize the copy-synthesis method [8, 9, 10, 11] to produce spoofed data using multiple vocoders from real mel-spectrograms extracted from the waveforms of the CN-Celeb1&2 datasets [6, 7], which contain more than $600,000$ utterances with 11 different genres. The new dataset is called CN-Spoof. To visually show the genre mismatch, we randomly select eight genres from the CN-Celeb1&2 and CN-Spoof datasets and train a lightweight convolutional neural network (LCNN) [3, 12]. We use the well-trained LCNN to extract the countermeasure embeddings from the waveforms with two seen and two unseen genres and visualize them in two-dimensional space by T-SNE [13] as Fig. 1 shows. From the figure, it is obvious that the embedding with seen "play" and "speech" genres can be classified well by the LCNN model. However, for the embedding with unseen "singing" and "recitation" genres, part of the real and fake embeddings overlap, which means it is hard to distinguish the fake audio samples with unseen genres by the LCNN model.

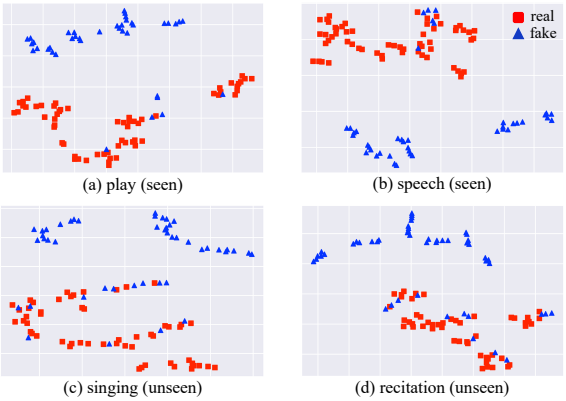

Figure 1: *Visualization of LCNN countermeasure embedding by T-SNE.*

To address this genre mismatch, we make four training protocols by grouping different genres. For each protocol, we randomly select utterances from several genres as the training dataset, and all training protocols share the same evaluation dataset. On the basis of the training and evaluation datasets, in this paper, we propose a novel multi-task learning method and design two auxiliary regularization objectives in addition to the main anti-spoofing objective to improve the generalization ability of the countermeasure model. The first objective is to simulate the genre mismatch scenario in the training stage by meta-optimization [14, 15, 16, 17]. In addition, for the second objective, we utilize a genre alignment module that contains a gradient reversal layer (GRL) [18, 19] to remove the genre information in the countermeasure embedding. Finally, the two auxiliary regularization objectives are combined with the main anti-spoofing objective by using uncertainty loss weights [20, 21], which can be learned from the data instead of setting them by hand resulting in inferior optimization. The experimental results on the evaluation dataset show that the proposed method significantly improves the genre generalization ability compared with the baseline system.

*This study is partially supported by JST CREST Grants (JPMJCR18A6 and JPMJCR20D3), and MEXT KAKENHI Grants (21K17775, 21H04906, 21K11951, 22K21319).

10.21437/Interspeech.2023-125

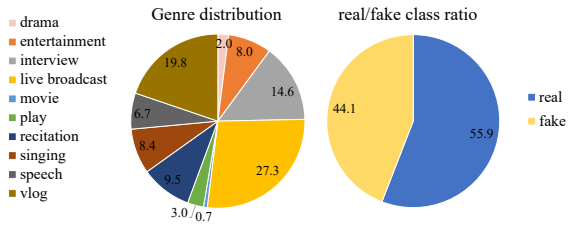

Figure 2: *Genre distribution and real/fake class ratio for the evaluation dataset.*

Table 1: *Genre group division*

Group	Genre Types
Group I	drama (dr), vlog (vl), speech (sp)
Group II	entertainment (en), interview (in), play (pl)
Group III	live broadcast (lb), movie (mo)
Group IV	singing (si), recitation (re)

The rest of this paper is organized as follows. In Section 2, we describe how to generate the CN-Spoof dataset and protocols for studying genre mismatch. The proposed multi-task learning method is illustrated in Section 3. The experimental results are shown in Section 4. Finally, we conclude the paper in Section 5.

2. CN-Spoof dataset and protocols

Since there are no related datasets in the research on anti-spoofing for the genre mismatch scenario, we leverage three pretrained neural vocoders (Multi-band MelGAN [22], Parallel WaveGAN [23], and HiFiGAN [24]), and two DSP vocoders (WORLD [25] and Griffin-Lim [26]), to reconstruct the waveforms from the real mel-spectrograms of the CN-Celeb1&2 datasets which are collected for speaker verification in the multi-genre scenario. Specifically, for each vocoder, we randomly select $20,000$ and $80,000$ utterances from the CN-Celeb1 and CN-Celeb2 datasets, respectively. Then, vocoded waveforms of the selected utterances using the above five vocoders are treated as fake data. As a result, the CN-Spoof dataset contains $500,000$ fake utterances in total.

After generating the CN-Spoof dataset, we combine it with the CN-Celeb1&2 datasets and randomly sample $200,000$ utterances from all genres as the evaluation dataset, whose genre distribution and real/fake class ratio are shown in Fig. 2. The remaining portions are utilized to construct the training dataset for different cross-genre protocols (CGP). We first divide ten genres included in the remaining portions into four groups, shown in Table 1. Note that the "advertisement" genre is discarded from the dataset since its number is $2,929$, which is limited. For each CGP, we randomly sample $660,000$ utterances with three genre groups as the training dataset from the remaining portions. In this way, each training dataset has an unseen genre group that exists in the evaluation dataset, as Table 2 shows.

3. Proposed multi-task learning method

The proposed multi-task learning method is illustrated concretely in this section. Since our method contains meta-optimization and genre alignment regularization objectives, we first describe a genre sampling strategy for constructing a task set for all objectives. Then, the main objective and auxiliary regularization objectives are depicted, respectively. Finally, we give details on combining multiple objectives with the uncertainty loss weights.

Table 2: *Cross-genre protocols (CGP)*

CGP	Seen Genres	Unseen Genres
CGP I	Group I, Group II, Group III	Group IV
CGP II	Group I, Group II, Group IV	Group III
CGP III	Group I, Group III, Group IV	Group II
CGP IV	Group II, Group III, Group IV	Group I

3.1. Genre sampling

As shown in Fig. 3, we randomly sample data from the training dataset $\mathcal{D} = \{\mathcal{D}_1, ..., \mathcal{D}_G | G > 1\}$, which includes G seen genres, to construct a task set $\mathcal{T} = \{\mathcal{T}_1, ..., \mathcal{T}_K | K > 1\}$, which contains K tasks. A task is composed of a meta-train dataset that contains $G_{mtr} (G_{mtr} < G)$ genres and is used for the main anti-spoofing objective as well as the auxiliary genre alignment objective, and a meta-test dataset is used that contains the remaining $G_{mte} = G - G_{mtr}$ genres for the auxiliary meta-optimization objective. In our experiments, for each training protocol, we set G_{mtr} as $G - 1$ for the meta-train dataset, and the remaining one is used as the meta-test dataset.

3.2. Main anti-spoofing objective

The main anti-spoofing objective is to distinguish whether input speech is real. In our proposed method, LCNN [3, 12] is selected as the backbone, as shown in Fig. 3. Countermeasure embeddings are extracted by the LCNN model from the meta-train dataset, which is represented by a formula:

$$\boldsymbol{e}_{mtr} = f_{\boldsymbol{\theta}_b}(\boldsymbol{x}_{mtr}), \tag{1}$$

where $\boldsymbol{e}_{mtr}$ and $\boldsymbol{x}_{mtr}$ represent the countermeasure embedding and corresponding speech in the meta-train dataset, respectively. $f_{\boldsymbol{\theta}_b}(\cdot)$ means the transformation function of the LCNN backbone, whose parameters are $\boldsymbol{\theta}_b$.

Then, a binary classifier, including an affine layer and a Sigmoid function, transforms the embeddings $\boldsymbol{e}_{mtr}$ into probability values, representing the possibility of speech being real. Finally, we perform a binary cross-entropy (BCE) loss function on the probability values and the corresponding labels. The process can be formulated as:

$$P(\boldsymbol{e}_{mtr}) = \frac{1}{1 + \exp(-\boldsymbol{e}_{mtr}^{\top} \boldsymbol{\theta}_c)}, \tag{2}$$

$$\begin{aligned} \mathcal{L}_{main} = -\frac{1}{N} \sum_{i=1}^{N} [&\mathcal{I}(y_{mtr}^i = 1) \log P(\boldsymbol{e}_{mtr}^i) \\ &+ \mathcal{I}(y_{mtr}^i \neq 1) \log(1 - P(\boldsymbol{e}_{mtr}^i))], \end{aligned} \tag{3}$$

where $\mathcal{L}_{main}$ represents the cost of the main anti-spoofing objective on the meta-train dataset, which has N speech samples. $\boldsymbol{\theta}_c$ denotes the parameters of the classifier. Note that for concise description, we ignore the bias parameter of the affine layer in Eq. (2). $\mathcal{I}(\cdot)$ is an indicator function that returns 1 if the condition is true and 0 otherwise, and y_{mtr}^i is the i-th label. When $\boldsymbol{e}_{mtr}^i$ is from real speech, y_{mtr}^i equals 1, otherwise 0. In the following part, we use $\boldsymbol{\theta}$ to denote the $\boldsymbol{\theta}_b$ and $\boldsymbol{\theta}_c$.

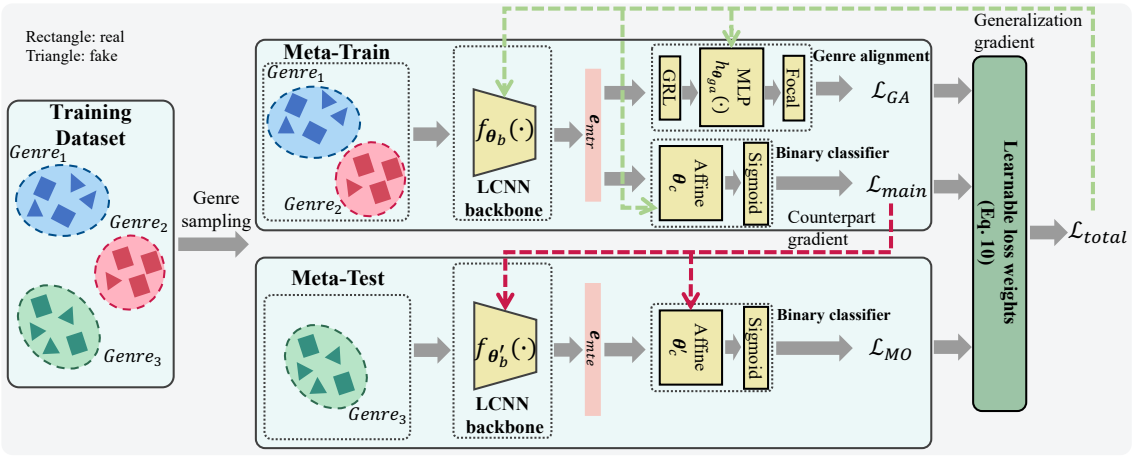

Figure 3: *Architecture of proposed multi-task learning method.*

3.3. Meta-optimization regularization objective

In the genre sampling stage, we divide a task into a meta-train and a meta-test dataset without overlapping genres. In this way, we can simulate the genre mismatch in the training stage by evaluating the performance on the meta-test dataset of the model trained on the meta-train dataset. Specifically, as Fig. 3 shows, we maintain a counterpart of the model parameters and update it by the loss $\mathcal{L}_{main}$ with the learning rate β:

$$\boldsymbol{\theta}' = \boldsymbol{\theta} - \beta \cdot \frac{\partial \mathcal{L}_{main}}{\partial \boldsymbol{\theta}}, \tag{4}$$

where $\boldsymbol{\theta}'$ denotes the updated parameters of the counterpart, including $\boldsymbol{\theta}'_b$ and $\boldsymbol{\theta}'_c$. Next, we evaluate the model performance on the meta-test dataset and compute the loss:

$$\boldsymbol{e}_{mte} = f_{\boldsymbol{\theta}'_b}(\boldsymbol{x}_{mte}), \tag{5}$$

$$P(\boldsymbol{e}_{mte}) = \frac{1}{1+\exp(-\boldsymbol{e}_{mte}^{\top}\boldsymbol{\theta}'_c)},$$

$$\mathcal{L}_{MO} = -\frac{1}{M}\sum_{i=1}^{M}[\mathcal{I}(y^i_{mte}=1)\log P(\boldsymbol{e}^i_{mte}) \tag{6}$$

$$+\mathcal{I}(y^i_{mte}\neq 1)\log(1-P(\boldsymbol{e}^i_{mte}))], \tag{7}$$

where $\boldsymbol{x}^i_{mte}$ and y^i_{mte} are the i-th speech sample and corresponding label in the meta-test dataset, which has M samples. $\boldsymbol{e}_{mte}$ is the corresponding countermeasure embedding. $\mathcal{L}_{MO}$ denotes the loss on the meta-test dataset for the meta-optimization regularization objective.

3.4. Genre alignment regularization objective

In addition to simulating the genre mismatch in the training stage by meta-optimization, we use another auxiliary genre alignment regularization objective that is realized by using a multi-layer perceptron (MLP) with the GRL component to improve the generalization ability of the LCNN model because it can remove the genre information contained in countermeasure embeddings by adversarial training [18]. Before inputting the countermeasure embedding $\boldsymbol{e}_{mtr}$ to the MLP for genre classification, we first apply the GRL component to it as Fig. 3 shows. The MLP contains three layers, and each has 128 neural units. Instead of utilizing plain cross-entropy as the loss function, here we use focal loss [27] because it can improve the discriminative ability of the genre alignment module, which is beneficial for further filtering out the genre information in countermeasure embeddings. The loss function of the genre alignment module is formulated as:

$$P(\boldsymbol{e}_{mtr}, g_c) = \frac{\exp(h_{\boldsymbol{\theta}^{g_c}_{ga}}(\boldsymbol{e}_{mtr}))}{\sum_{j=1}^{G}\exp(h_{\boldsymbol{\theta}^{g_j}_{ga}}(\boldsymbol{e}_{mtr}))}, \tag{8}$$

$$\mathcal{L}_{GA} = -\frac{1}{N}\sum_{i}^{N}(1-P(\boldsymbol{e}^i_{mtr}, g^i_c))^{\gamma}\log P(\boldsymbol{e}^i_{mtr}, g^i_c), \tag{9}$$

where g_c is the genre label of the countermeasure embedding $\boldsymbol{e}_{mtr}$. $h_{\boldsymbol{\theta}^{g_c}_{ga}}(\cdot)$ denotes a transformation for outputting the unnormalized possibility of the c-th genre from the genre alignment module, whose parameters are $\boldsymbol{\theta}_{ga}$. $P(\boldsymbol{e}_{mtr}, g_c)$ is the probability that the countermeasure embedding $\boldsymbol{e}_{mtr}$ belongs to the correct genre, and γ is a hyper-parameter that can adjust the gradient contribution of different samples in accordance with their difficulties. Note here that we ignore the hyper-parameter α in [18] since all genres are treated equally in our method. Due to the GRL component, the gradient after the GRL in the backward propagation is reversed compared with that without the GRL.

3.5. Learnable loss weights

As the proposed method has multiple loss terms, the model performance is extremely sensitive to loss weight selection [20, 21]. To this end, we use a common strategy in multi-task learning to combine multiple loss terms with learnable loss weights, the aim of which is learning the optimal weights. As our auxiliary objectives can be discarded in the inference stage, we treat the loss of auxiliary objectives as the regularization terms [21]. Thus, the total loss can be formulated as:

$$\begin{aligned}\mathcal{L}_{total} =& \frac{1}{2\cdot\lambda^2_{main}}\cdot\mathcal{L}_{main} + \ln(1+\lambda^2_{main}) \\ &+ \frac{1}{2\cdot\lambda^2_{MO}}\cdot\mathcal{L}_{MO} + \ln(1+\lambda^2_{MO}) \\ &+ \frac{1}{2\cdot\lambda^2_{GA}}\cdot\mathcal{L}_{GA} + \ln(1+\lambda^2_{GA}),\end{aligned} \tag{10}$$

where λ_{main}, λ_{MO}, and λ_{GA} are the learnable parameters for each loss term, respectively. As for the logarithmic terms in Eq. (10), they are constraint conditions for avoiding trivial solutions [20, 21].

Table 3: *EER (%) of experimental results on CGP. For each protocol, the genre group in the bracket does not appear in the training dataset. A bold number means the best performance of this genre.*

Protocol	System	Overall	Group I			Group II			Group III		Group IV	
			dr	vl	sp	en	in	pl	lb	mo	si	re
CGP I (Group IV)	$\mathcal{L}_{main}$	8.299	6.890	9.124	6.582	7.505	7.799	6.876	7.933	7.960	9.517	9.779
	$\mathcal{L}_{MO}, \mathcal{L}_{main}$	7.863	5.626	9.053	6.565	5.962	6.818	6.148	7.929	6.799	8.761	9.385
	$\mathcal{L}_{GA}, \mathcal{L}_{main}$	8.238	6.831	9.082	6.904	7.399	7.672	6.855	7.903	8.031	9.063	9.615
	$\mathcal{L}_{MO}, \mathcal{L}_{GA}, \mathcal{L}_{main}$	**7.511**	**5.109**	**8.827**	**6.362**	**5.508**	**6.266**	**5.216**	**7.577**	**6.799**	**8.248**	**9.157**
CGP II (Group III)	$\mathcal{L}_{main}$	8.566	7.176	9.281	6.887	7.601	8.147	7.414	8.919	7.794	9.053	8.996
	$\mathcal{L}_{MO}, \mathcal{L}_{main}$	8.181	5.926	9.387	6.797	6.456	7.184	6.408	8.682	6.965	8.916	8.788
	$\mathcal{L}_{GA}, \mathcal{L}_{main}$	8.481	7.176	9.369	7.073	7.320	7.863	6.855	8.787	7.334	9.157	9.110
	$\mathcal{L}_{MO}, \mathcal{L}_{GA}, \mathcal{L}_{main}$	**7.764**	**5.788**	**8.959**	**6.664**	**5.770**	**6.676**	**5.365**	**8.339**	**6.347**	**8.314**	**8.425**
CGP III (Group II)	$\mathcal{L}_{main}$	8.599	7.922	9.099	7.035	8.620	8.983	8.905	8.112	8.679	9.424	8.603
	$\mathcal{L}_{MO}, \mathcal{L}_{main}$	8.182	7.118	8.942	6.823	7.657	8.236	7.489	**7.641**	**7.772**	9.277	8.693
	$\mathcal{L}_{GA}, \mathcal{L}_{main}$	8.505	8.266	8.915	6.785	8.365	9.009	8.420	7.947	8.808	9.658	8.657
	$\mathcal{L}_{MO}, \mathcal{L}_{GA}, \mathcal{L}_{main}$	**8.032**	**6.889**	**8.746**	**6.447**	**7.464**	**8.192**	**7.202**	7.738	8.126	**8.983**	**8.060**
CGP IV (Group I)	$\mathcal{L}_{main}$	8.160	6.339	9.322	7.242	6.995	7.639	7.261	7.886	8.290	8.966	8.657
	$\mathcal{L}_{MO}, \mathcal{L}_{main}$	7.827	**5.454**	9.357	7.174	6.141	6.765	**5.602**	7.850	6.965	8.502	8.790
	$\mathcal{L}_{GA}, \mathcal{L}_{main}$	7.944	6.028	9.334	7.124	6.954	7.131	6.483	**7.819**	6.799	8.804	**8.299**
	$\mathcal{L}_{MO}, \mathcal{L}_{GA}, \mathcal{L}_{main}$	**7.739**	5.568	**9.182**	**7.073**	**6.004**	**6.489**	5.700	7.840	**6.136**	**8.341**	8.633

4. Experimental results

4.1. Experimental setup

As described in Section 2, the CN-Spoof dataset was combined with the CN-Celeb1&2 datasets, and we divided these data in accordance with the cross-genre protocols. For each protocol, the training dataset contained $660,000$ utterances. All protocols shared the same evaluation dataset, which contained $200,000$ utterances from all genres.

As for the systems used for comparison in the paper, the LCNN model was selected as the baseline system. Additionally, we also constructed two other experimental systems: one incorporating the $\mathcal{L}_{main}$ and $\mathcal{L}_{MO}$ losses, and the other incorporating the $\mathcal{L}_{main}$ and $\mathcal{L}_{GA}$ losses. These systems were developed to thoroughly examine the impact of various regularizations.

4.2. Training methodology

For the systems without meta-optimization, we randomly selected 64 samples from the dataset as a mini-batch. For the systems including meta-optimization, we randomly selected 64 samples from the dataset. One genre was randomly selected from this subset as the meta-test dataset, while the remaining samples were used as the meta-train dataset. We trained all systems for 40 epochs using an SGD optimizer [28] with a 0.001 initial learning rate, 0.9 momentum, and 0.0001 L2 regularization. The learning rate was decayed by 0.9 every epoch. As for other hyper-parameters, β in Eq. (4) was set to 0.001, and γ in Eq. (9) was set to 5.

4.3. Results and analysis

The experimental results are shown in Table 3. As we can see, for each protocol, the countermeasure performance without the regularization terms on the unseen genres was generally worse than the counterparts for the other protocols except for the "movie" genre in CGP II and "drama" genre in CGP IV. This result proves that a model without the regularization terms cannot generalize well on unseen genres, which is consistent with our hypothesis.

Although the GRL component has shown the capacity to generalize well on unseen domains in the speaker verification task [19], we found that the genre alignment loss $\mathcal{L}_{GA}$ only slightly improved the generalization ability compared with the baseline model in the genre mismatch scenario. In contrast, the meta-optimization loss $\mathcal{L}_{MO}$ improved the EER numbers on some unseen genres, such as the "singing" genre for CGP I and "movie" genre for CGP II. However, there are some other genres that the system had difficulty generalizing well, such as the "speech" genre for CGP III.

As for our multi-task learning method that integrates the meta-optimization and genre alignment regularization objectives using learnable loss weights, it significantly improved the generalization ability in the genre mismatch scenario by comparing its result with the baseline system for unseen genres. Even for some challenging genres like the "recitation" and "speech" genres, which cannot be generalized well by using either $\mathcal{L}_{MO}$ or $\mathcal{L}_{GA}$ only, combining them further improve the generalization ability.

Moreover, we can see that our proposed system not only performs well on unseen genres but also on seen genres. For protocol CGP I and II, our system obtained the best performance on all genres compared with the other systems. As for the protocol CGP III and IV, although our approach performed slightly worse than the other systems for several genres, it still obtained the best EER numbers for most genres.

5. Conclusions

In this paper, we explored a new mismatch scenario for the anti-spoofing objective, in which the fake speech may come from the real speech with unseen genres. Since there is no anti-spoofing data related to this scenario, we utilized the copy-synthesis method to create a spoofed dataset called CN-Spoof based on the CN-Celeb1&2 datasets. Our proposed multi-task learning method, which combines the meta-optimization loss and genre alignment loss as the regularization terms by using learnable loss weights, shows the potential to improve the generalization ability of the countermeasure models under this scenario. The experimental results on four different cross-genre protocols proved that our method is more robust than the baseline system, even when facing difficult genres.

6. References

[1] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Sixteenth annual conference of the international speech communication association*, 2015.

[2] T. H. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. W. D. Evans, J. Yamagishi, and K.-A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Interspeech*, 2017.

[3] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[4] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*, 2021.

[5] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan *et al.*, "Add 2022: the first audio deep synthesis detection challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9216–9220.

[6] 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 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7604–7608.

[7] 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.

[8] W. J. Holmes, "Copy synthesis of female speech using the jsru parallel formant synthesiser." in *EUROSPEECH*, 1989, pp. 2513–2516.

[9] M. Pal, D. Paul, and G. Saha, "Synthetic speech detection using fundamental frequency variation and spectral features," *Computer Speech & Language*, vol. 48, pp. 31–50, 2018.

[10] J. C. Frank and L. Schönherr, "Wavefake: A data set to facilitate audio deepfake detection," in *Proc. NeurIPS Datasets and Benchmarks 2021*, 2021.

[11] X. Wang and J. Yamagishi, "Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (accepted)*. IEEE, 2023.

[12] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC Antispoofing Systems for the ASVspoof2019 Challenge," in *Proc. Interspeech 2019*, 2019, pp. 1033–1037.

[13] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[14] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[15] J. Kang, R. Liu, L. Li, Y. Cai, D. Wang, and T. F. Zheng, "Domain-invariant speaker vector projection by model-agnostic meta-learning," in *Interspeech*, 2020.

[16] H. Zhang, L. Wang, K. A. Lee, M. Liu, J. Dang, and H. Chen, "Learning domain-invariant transformation for speaker verification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7177–7181.

[17] Zhang, Hanyi and Wang, Longbiao and Lee, Kong Aik and Liu, Meng and Dang, Jianwu and Chen, Hui, "Meta-learning for cross-channel speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5839–5843.

[18] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[19] 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.

[20] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.

[21] L. Liebel and M. Körner, "Auxiliary tasks in multi-task learning," *arXiv preprint arXiv:1805.06334*, 2018.

[22] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band melgan: Faster waveform generation for high-quality text-to-speech," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 492–498.

[23] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[24] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[25] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[26] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 8, pp. 1223–1235, 1988.

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[28] S.-i. Amari, "Backpropagation and stochastic gradient descent method," *Neurocomputing*, vol. 5, no. 4-5, pp. 185–196, 1993.