An Empirical Study on Channel Effects for Synthetic Voice Spoofing Countermeasure Systems

You Zhang[1], *Ge Zhu*[1], *Fei Jiang*[1,2], *Zhiyao Duan*[1]

[1]University of Rochester, Rochester, NY, USA
[2]Beijing Institute of Technology, Beijing, China
{you.zhang, ge.zhu, fei.jiang, zhiyao.duan}@rochester.edu

Abstract

Spoofing countermeasure (CM) systems are critical in speaker verification; they aim to discern spoofing attacks from bona fide speech trials. In practice, however, acoustic condition variability in speech utterances may significantly degrade the performance of CM systems. In this paper, we conduct a cross-dataset study on several state-of-the-art CM systems and observe significant performance degradation compared with their single-dataset performance. Observing differences of average magnitude spectra of bona fide utterances across the datasets, we hypothesize that channel mismatch among these datasets is one important reason. We then verify it by demonstrating a similar degradation of CM systems trained on original but evaluated on channel-shifted data. Finally, we propose several channel robust strategies (data augmentation, multi-task learning, adversarial learning) for CM systems, and observe a significant performance improvement on cross-dataset experiments.

Index Terms: spoofing countermeasure, channel variation, cross dataset, data augmentation, deep learning

1. Introduction

Automatic speaker verification (ASV) systems are vulnerable to spoofing attacks, where attackers pretend to be the target speaker by presenting false but similar-to-bona-fide speech trials [1]. Spoofing countermeasure (CM) systems aim to detect such attacks. Spoofing attacks are considered physical access (PA) if they are utterances presented to the microphone of the ASV system, and logical access (LA) if they bypass the microphone and feed to the verification algorithm directly. In a common anti-spoofing setup as in ASVspoof2019 [2], the PA scenario features replay attacks using various playback devices, while the LA scenario features synthetic utterances generated by text-to-speech (TTS) and voice conversion (VC) algorithms.

Recently, deep learning technologies have shown great success in learning discriminative speaker embeddings to classify spoofing attacks from bona fide speech in the LA scenario [3]. The CM community has been exploring the usage of different input speech features [4, 5, 6, 7], model architectures [8, 9, 10, 11, 12], and loss functions [13, 14, 15] to improve the performance on detecting synthetic attacks.

However, several cross-dataset studies [16, 17, 18, 19] in anti-spoofing show significant performance degradation from single-dataset studies. For example, when systems are trained on LA but tested on PA, performance degradation happens, and suggested solutions include the usage of more generalized speech features [17, 19] and domain adaptation [18]. Another more surprising performance degradation happens when a state-of-the-art CM system is trained and tested on different LA datasets [20]. The authors suggested that it is because some unseen attacks are more challenging. While we agree that this can be an important reason, we also think that other differences, such as channel variation across datasets could be possible reasons. This motivates us to systematically conduct a cross-dataset study on synthetic voice spoofing CM systems.

Such cross-dataset studies are important in the design of robust CM systems. Due to limited access of training data and its channel variation, CM systems may be frail in practice. Here *channel effects* refer to audio effects imposed onto the speech signal throughout the entire recording and transmission process, including reverberation of recording environments, frequency responses of recording devices, and compression algorithms in telecommunication. Without properly considering and compensating for these effects, CM systems may overfit to the limited channel effects presented in the training set and fail to generalize to unseen channel variation. This issue has been studied in replay attacks [21], but little attention is paid in the LA scenario. For example, the bona fide speech utterances in ASVspoof2019LA were all from the VCTK corpus, and the TTS/VC systems used in generating LA attacks were all trained on the VCTK corpus. This may introduce strong biases to CM systems on the limited channel variation.

In this work, we first conduct a cross-dataset study of three state-of-the-art CM systems between ASVspoof2019LA, ASVspoof2015, and VCC2020. Observing significant performance degradation on all CM systems, we hypothesize that the channel effect mismatch between these datasets is one important reason for the degradation. To test our hypothesis, we first compare the average magnitude spectra across all bona fide utterances among these three datasets and observe significant mismatches. We then conduct a controlled cross-channel experiment by training the three CM systems on ASVspoof2019LA and evaluating them on the evaluation set of its channel-augmented version, ASVspoof2019LA-Sim, which is generated by passing ASVspoof2019LA utterances through an acoustic simulator [22]; our hypothesis is again verified by consistent performance degradation across the three CM systems. Finally, we propose several strategies to improve channel robustness leveraging the channel-augmented data. Results show that these strategies successfully improve the cross-dataset performance of all three CM systems.

As we conduct this study, we notice that the LA subchallenge of ASVspoof 2021 [23] also intends to consider channel robustness in its evaluation mechanism. We believe that our study will provide useful insights into this research direction.

2. Cross-Dataset Studies

In this section, we take three state-of-the-art CM systems and three commonly used anti-spoofing datasets to extensively study the performance degradation issue in cross-dataset evalu-

ation for synthetic voice spoofing CM systems.

2.1. Datasets

We employ three datasets containing both bona fide speech and synthetic speech generated by TTS or VC algorithms.

ASVspoof2019LA [24] is a large-scale dataset used in the LA sub-challenge of ASVspoof2019. It contains a large variety of up-to-date TTS and VC algorithms forming a diverse collection of attacks. The bona fide speech was collected from the VCTK corpus [25]. Training and Development sets share the same attacks (A01-A06), but the evaluation set contains totally different attacks (A07-A19).

ASVspoof2015 [26] is the database for the 2015 edition of the ASVspoof challenge, which only deals with synthetic voice spoofing attacks. The training set includes S01-S05 attacks and the evaluation set includes S01-S10, which have both known and unknown attacks.

VCC2020 (Voice Conversion Challenge 2020) [27] distributed a new dataset that aims to develop systems for converting speech from a source speaker to a target speaker. The participating teams submitted their converted speech developed on the training data. The utterances in the training data provided by the organizers are considered bona fide trials, while the converted utterances generated by each submitted VC system are considered spoofing attacks. Different from the previous two datasets, VCC2020 is multilingual.

2.2. Experimental Setup

We select three state-of-the-art CM systems that show top performance on the ASVspoof2019LA database: LCNN [9], ResNet [10], and ResNet-OC [15]. These CM systems are trained on the training set of ASVspoof2019LA [24], and validated on its development set. For evaluation, we use only the evaluation sets of ASVspoof2019LA and ASVspoof2015, and the complete VCC2020. In this way, all spoofing attacks in evaluation are unknown.

For all of the CM systems, the 60-d linear-frequency cepstral coefficients (LFCC) are extracted as speech features from each frame of the utterances. The frame length is 20ms and the hop size is 10ms. To form batches, we set 750 frames as the fixed length; We use repeat padding for shorter trials, and we randomly choose a consecutive 750-frame segment for longer trials. The learning rate is initially set to 0.0003 with 50% decay for every 10 epochs. We train the network for 100 epochs on a single NVIDIA GTX 1080 Ti GPU. Finally, we select the model with the lowest validation loss for evaluation.

Each CM system outputs a score to indicate the confidence that the given utterance is bona fide. Equal Error Rate (EER) is calculated by setting a threshold on the CM score such that the false alarm rate is equal to the miss rate. We use EER since it is used across all ASVspoof challenge series.

2.3. Results and Analyses

In Table 1, we demonstrate EER degradation across datasets for all three CM systems. As a sanity check, the result for LCNN is consistent with that in [19] on ASVspoof2015 and [20] on VCC2020. For the sake of space and without loss of generality, here we only analyze the ResNet-OC CM system, as it achieves the lowest EER among all single-system CM on the evaluation set of the ASVspoof2019LA dataset [15].

The EER degradation could be because that the score distribution of spoof trials shifts up, or that of bona fide trials shifts down. We hence plot the score distributions of ResNet-OC in Figure 1. We observe that the unknown spoofing attacks are mostly correctly scored across the three datasets, verifying that ResNet-OC is capable of detecting unseen attacks. However, the score distribution of bona fide trials shifts down significantly from ASVspoof2019LA to ASVspoof2015 and VCC2020, which would cause many false alarm errors. This suggests that the main cause of the EER degradation is some differences in bona fide speech, among which, channel variation is worth checking.

Table 1: *EER performance across different evaluation datasets (ASVspoof2019LA-eval, ASVspoof2015-eval, VCC2020). All of the three CM systems are trained on the training set of ASVspoof2019LA and validated on its development set.*

EER (%)	CM Systems		
Evaluation Datasets	LCNN [9]	ResNet [10]	ResNet-OC [15]
2019LA-eval	3.25	5.23	2.29
2015-eval	24.55	37.11	26.30
VCC2020	33.78	36.09	41.66

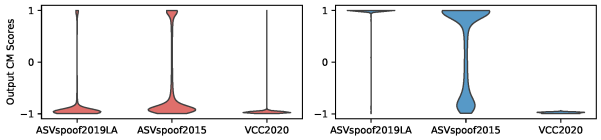

Figure 1: *Score distributions of ResNet-OC method on spoofing attacks (left) and bona fide (right) of cross-dataset evaluation.*

In Figure 2, we show the average magnitude spectrum across all bona fide utterances of each dataset as an indication of its channel effect. We can see that the average spectra are very different among the three datasets . With this observation, we hypothesize that channel mismatch is an important reason for the EER degradation, and will test this hypothesis through a controlled experiment next.

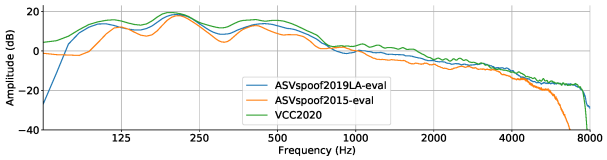

Figure 2: *Average magnitude spectra of bona fide utterances across different datasets.*

2.4. Channel Simulation

We augment the channel effects of the ASVspoof2019LA dataset through an open-source channel simulator [22]. This channel simulator provides three types of degradation processes including additive noise, telephone and audio codecs, and device/room impulse responses (IRs). In our simulation, we choose 12 out of 74 different device IRs and apply each of them to all utterances of ASVspoof2019LA. This channel-augmented dataset is named **ASVspoof2019LA-Sim**, and the same train/dev/eval split is followed from ASVspoof2019LA. We do not use additive noise, audio codecs, or reverberation IRs in our simulation. The average magnitude spectra of the

augmented bona fide utterances in the ASVspoof2019LA-Sim evaluation set using each device IR are plotted in Figure 3.

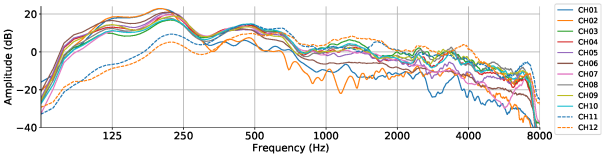

Figure 3: *Average magnitude spectra of channel-shifted bona fide utterances in the evaluation set of ASVspoof2019LA-Sim using different channel IRs.*

We then test the performance of the three CM systems selected in Section 2.2 on the evaluation set of ASVspoof2019LA-Sim. As a reminder, all three CM systems are trained and validated on the training and development sets of ASVspoof2019LA (i.e., without augmented channel-shifts). Table 2 shows the results. The average and standard deviation of EERs across all of the 12 simulated channels are calculated. We observe that the average EER drops significantly from ASVspoof2019LA-eval and the standard deviation is quite large. We conclude that the channel mismatch between training and evaluation is indeed an important reason for the performance degradation.

Table 2: *EER performance on ASVspoof2019LA-Sim-eval. Average and standard deviation EERs are calculated across the 12 simulated channels. All of the three CM systems are trained on ASVspoof2019LA-train.*

EER (%)	CM Systems		
Statistics	LCNN [9]	ResNet [10]	ResNet-OC [15]
Avg. (CH01-CH12)	27.75	48.78	40.46
Std. (CH01-CH12)	7.44	18.80	11.22

3. Channel Robust Strategies

We propose several strategies to improve the channel robustness of CM systems using the channel-shifted data. Specifically, we create a *channel-augmented training set* containing the original ASVspoof2019LA training data and only 10 out of the 12 channel shifts of the ASVspoof2019LA-Sim training data. For evaluation, we use all 12 channel shifts of the ASVspoof2019LA-Sim evaluation data and their original utterances.

Similar to Section 2.3, without loss of generality and for the sake of space, we only demonstrate these strategies on the ResNet-OC CM system [15]. Its model architecture is illustrated on the left side of Figure 4. An embedding network based on ResNet18 with attentive pooling, parameterized by θ_e, aims to learn a discriminative speech embedding, which is then classified by another fully connected (FC) layer, parameterized by θ_{cm}, into spoofing attacks or bona fide speech. The loss function is OC-Softmax. Without the channel robust strategies, ResNet-OC is trained only on the ASVspoof2019LA training set, and this model is named the **Vanilla** model as a baseline.

3.1. Proposed Strategies

Utilizing utterances from the ASVspoof2019LA-Sim training set, we proposed three channel-robust strategies.

Augmentation (AUG) uses the same model architecture as

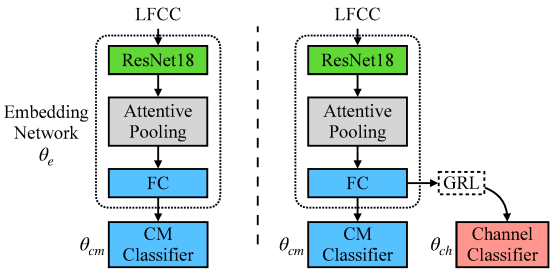

Figure 4: *Model structure of the proposed channel robust strategies. Left: Vanilla model and AUG. Right: MT-AUG (w/o GRL) and ADV-AUG (w/ GRL).*

the Vanilla model (Figure 4 left side) but is trained on the channel augmented training set mentioned above.

Multi-Task Augmentation (MT-AUG) adds a channel classifier, parameterized by θ_{ch}, to the vanilla model architecture to form a multi-task learning setup. The overall model structure is shown on the right side of Figure 4, skipping the "GRL" module. This channel classifier uses two fully connected layers to map deep speech embeddings to the channel labels, and uses the cross entropy loss. The overall training objective of MT-AUG is thus:

$$(\hat{\theta}_e, \hat{\theta}_{cm}, \hat{\theta}_{ch}) = \arg\min_{\theta_e, \theta_{cm}, \theta_{ch}} \mathcal{L}_{cm}(\theta_e, \theta_{cm}) + \lambda \mathcal{L}_{ch}(\theta_e, \theta_{ch}) . \tag{1}$$

Adversarial Augmentation (ADV-AUG) inserts a Gradient Reversal Layer (GRL) [28] between the embedding network and the channel classifier. The loss of the channel classifier is now backpropagated through the GRL, with the sign reversed, to the embedding network. Therefore, the embedding network aims to maximize the channel classification error while the channel classifier aims to minimize it, forming an adversarial training paradigm. When equilibrium is reached, the learned speech embeddings would be channel-agnostic, making the CM classifier robust to channel variation.

$$\begin{aligned} (\hat{\theta}_e, \hat{\theta}_{cm}) &= \arg\min_{\theta_e, \theta_{cm}} \mathcal{L}_{cm}(\theta_e, \theta_{cm}) - \lambda \mathcal{L}_{ch}(\theta_e, \hat{\theta}_{ch}) \\ (\hat{\theta}_{ch}) &= \arg\min_{\theta_{ch}} \mathcal{L}_{ch}(\hat{\theta}_e, \theta_{ch}) \end{aligned} \tag{2}$$

It is noted that GRL has been employed in various speech processing tasks such as phonetic-informed speaker embedding [29], channel-invariant speaker verification [30, 31], speaker-invariant speech emotion recognition [32, 33], and cross-domain replay spoofing attack detection [21].

3.2. In-Domain Test

In this test, we evaluate the proposed strategies on the evaluation set of our simulated ASVspoof2019LA-Sim dataset. Out of the 12 channel shifts in the evaluation set, 10 have also been used in forming the channel-augmented training set, with which we trained the above strategies. Therefore, our test results contain both seen (CH01-10) and unseen (CH11-12) channel shifts.

Table 3 shows EER statistics. We can see that the average EER on seen channels and the EERs on unseen channels of all three training strategies decrease much from the vanilla model. This shows that the proposed strategies do improve the channel robustness of the CM system. The significant decrease of the standard deviation of EER across the 10 seen channels also

suggests that the strategies make the CM system less sensitive to channel variation. Comparing the three strategies in terms of EER, ADV-AUG performs the best on seen channels, while AUG performs the best on unseen channels.

Table 3: *EER performance comparison of the proposed strategies and the vanilla model on ASVspoof2019LA-Sim-eval. The proposed strategies are trained on the augmented training set.*

EER (%)	Methods			
	Vanilla	AUG	MT-AUG	ADV-AUG
Avg. (CH01-10)	38.14	4.43	4.29	3.92
Std. (CH01-10)	10.83	0.75	0.46	0.43
CH11	54.98	3.58	4.59	3.78
CH12	49.17	4.41	7.08	6.28

The detection error tradeoff (DET) curve is often used to show the tradeoff between miss and false alarm errors in detection tasks [34]. We compute a DET curve for the original and each channel-shifted data subset . In Figure 5, we plot the DET curve on the original ASVspoof2019LA-eval dataset, the average of the DET curves across all seen channels (CH01-CH10) with its one standard deviation (SD) above and below, and the DET curves of the two unknown channels (CH11, CH12). All of the curves are shown in the normal deviate scale.

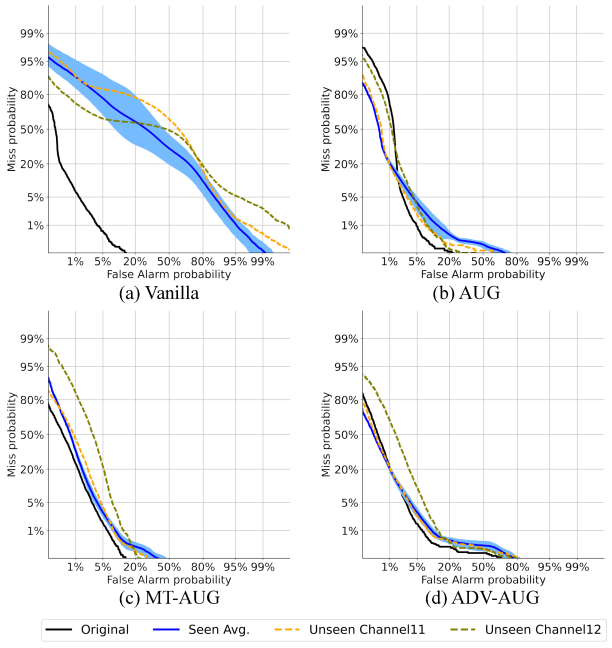

Figure 5: *DET curves of the vanilla model and the proposed channel robust strategies, evaluated on the original ASVspoof2019LA-eval and the simulated ASVspoof2019LA-Sim-eval (with 12 channel effects (10 seen, 2 unseen))*

For channel-robust CM systems, the SD region of DET curves of seen channels is expected to be narrow, and the curves of the original data and unknown channels are expected to be within or close to the SD region. From Figure 5, the vanilla system does not show such properties. The proposed approaches, in contrast, all show much narrower SD regions, and the DET curves for the original data and unseen channels are also much closer to the SD region. This again suggests that the proposed approaches improve the channel robustness of the CM system.

3.3. Out-of-Domain (Cross-Dataset) Test

We perform a cross-dataset evaluation as in Section 2, and the EER results are shown in Table 4. Compared to the Vanilla model, our proposed channel-robust strategies show slight degradation on the in-domain dataset, ASVspoof19LA-eval. However, they show significant improvement on both out-of-domain datasets, ASVspoof2015-eval and VCC2020, verifying our hypothesis of channel mismatch among these datasets and the effectiveness of the proposed strategies.

Table 4: *EER comparison of the proposed strategies and the vanilla model on cross-dataset evaluation.*

EER(%)	Methods			
Evaluation Datasets	Vanilla	AUG	MT-AUG	ADV-AUG
2019LA-eval	2.29	2.92	3.41	3.23
2015-eval	26.30	16.25	22.10	14.38
VCC2020	41.66	30.51	28.85	27.07

As ADV-AUG achieves the best cross-dataset performance among all channel robust strategies, we show its new score distribution in Fig. 6. Compared with Fig. 1, the score distributions of spoofing attacks are not changed much, while those of bona fide trials in ASVspoof2015 and VCC2020 are all moved up. This suggests that the channel mismatch issue mainly resides on bona fide speech, and the proposed strategies are useful. On the other hand, this move-up is not sufficient for VCC2020, suggesting the limitation of the proposed strategies. The remaining performance gap may also be due to other unique properties of VCC2020 such as its multilingual and more challenging attacks.

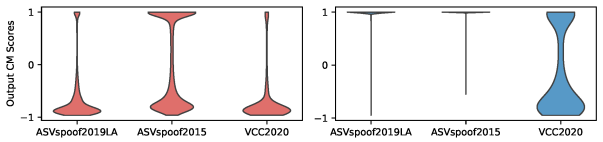

Figure 6: *Score distributions of ADV-AUG strategy on spoofing attacks (left) and bona fide (right) of cross-dataset evaluation.*

4. Conclusions

In this paper, we observed significant performance degradation of several state-of-the-art CM systems when they are trained on ASVspoof2019LA and tested on ASVspoof2015 and VCC2020. We then hypothesized that the channel effect is one reason for the performance degradation, after observing that the average magnitude spectrum of bona fide speech is different across the datasets. We further verified this hypothesis by testing a CM system on a channel-shifted dataset using various device IRs and observing a similar performance degradation. We then proposed several strategies to improve the robustness of CM systems to channel variation, and obtained significant improvement in both in-domain and cross-dataset tests. For future work, we plan to investigate other potential factors that may cause cross-dataset performance degradation of CM systems.

5. Acknowledgements

This work was supported by National Science Foundation grant No. 1741472 and funding from Voice Biometrics Group. The authors would also like to thank Dr. Xin Wang from National Institute of Informatics, Tokyo, Japan for valuable discussions.

6. References

[1] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. De Leon, "Speaker recognition anti-spoofing," in *Handbook of biometric anti-spoofing*. Springer, 2014, pp. 125–146.

[2] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Interspeech*, 2019, pp. 1008–1012.

[3] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[4] R. K. Das, J. Yang, and H. Li, "Long range acoustic and deep features perspective on ASVspoof 2019," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 1018–1025.

[5] J. Yang, R. K. Das, and H. Li, "Significance of subband features for synthetic speech detection," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2160–2170, 2019.

[6] J. Yang and R. K. Das, "Long-term high frequency features for synthetic speech detection," *Digital Signal Processing*, vol. 97, p. 102622, 2020.

[7] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6369–6373.

[8] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "ASSERT: Anti-spoofing with squeeze-excitation and residual networks," in *Proc. Interspeech*, 2019, pp. 1013–1017.

[9] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC antispoofing systems for the ASVspoof2019 challenge," in *Proc. Interspeech*, 2019, pp. 1033–1037.

[10] J. Monteiro, J. Alam, and T. H. Falk, "Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers," *Computer Speech & Language*, vol. 63, p. 101096, 2020.

[11] P. Parasu, J. Epps, K. Sriskandaraja, and G. Suthokumar, "Investigating light-ResNet architecture for spoofing detection under mismatched conditions," in *Proc. Interspeech*, 2020, pp. 1111–1115.

[12] X. Li, N. Li, C. Weng, X. Liu, D. Su, D. Yu, and H. Meng, "Replay and synthetic speech detection with Res2Net architecture," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6354–6358.

[13] A. Gomez-Alanis, J. A. Gonzalez-Lopez, and A. M. Peinado, "A kernel density estimation based loss function and its application to ASV-spoofing detection," *IEEE Access*, vol. 8, pp. 108 530–108 543, 2020.

[14] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of audio deepfake detection," in *Proceedings of the Odyssey Speaker and Language Recognition Workshop, Tokyo, Japan*, 2020, pp. 1–5.

[15] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[16] P. Korshunov and S. Marcel, "Cross-database evaluation of audio-based spoofing detection systems," in *Proc. Interspeech*, 2016, pp. 1705–1709.

[17] D. Paul, M. Sahidullah, and G. Saha, "Generalization of spoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2017, pp. 2047–2051.

[18] I. Himawan, F. Villavicencio, S. Sridharan, and C. Fookes, "Deep domain adaptation for anti-spoofing in speaker verification systems," *Computer Speech & Language*, vol. 58, pp. 377–402, 2019.

[19] R. K. Das, J. Yang, and H. Li, "Assessing the scope of generalized countermeasures for anti-spoofing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6589–6593.

[20] R. K. Das, T. Kinnunen, W.-C. Huang, Z.-H. Ling, J. Yamagishi, Z. Yi, X. Tian, and T. Toda, "Predictions of subjective ratings and spoofing assessments of voice conversion challenge 2020 submissions," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 2020, pp. 99–120.

[21] H. Wang, H. Dinkel, S. Wang, Y. Qian, and K. Yu, "Cross-domain replay spoofing attack detection using domain adversarial training," in *Proc. Interspeech*, 2019, pp. 2938–2942.

[22] M. Ferras, S. Madikeri, P. Motlicek, S. Dey, and H. Bourlard, "A large-scale open-source acoustic simulator for speaker recognition," *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 527–531, 2016.

[23] ASVspoof 2021. [Online]. Available: https://www.asvspoof.org/

[24] 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.

[25] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[26] 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, pp. 2037–2041.

[27] Z. Yi, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020—intra-lingual semi-parallel and cross-lingual voice conversion—," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 2020, pp. 80–98.

[28] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1180–1189.

[29] S. Wang, J. Rohdin, L. Burget, O. Plchot, Y. Qian, K. Yu, and J. Cernockỳ, "On the usage of phonetic information for text-independent speaker embedding extraction." in *Proc. Interspeech*, 2019, pp. 1148–1152.

[30] Z. Meng, Y. Zhao, J. Li, and Y. Gong, "Adversarial speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6216–6220.

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

[32] H. Li, M. Tu, J. Huang, S. Narayanan, and P. Georgiou, "Speaker-invariant affective representation learning via adversarial training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7144–7148.

[33] Y. Yin, B. Huang, Y. Wu, and M. Soleymani, "Speaker-invariant adversarial domain adaptation for emotion recognition," in *International Conference on Multimodal Interaction*, 2020, pp. 481–490.

[34] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," National Inst of Standards and Technology Gaithersburg MD, Tech. Rep., 1997.