INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Channel-wise Gated Res2Net: Towards Robust Detection of Synthetic Speech Attacks

Xu Li[1], *Xixin Wu*[2], *Hui Lu*[1], *Xunying Liu*[1], *Helen Meng*[1]

[1]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, China
[2]Department of Engineering, University of Cambridge, UK
{xuli, luhui, xyliu, hmmeng}@se.cuhk.edu.hk, xw369@cam.ac.uk

Abstract

Existing approaches for anti-spoofing in automatic speaker verification (ASV) still lack generalizability to unseen attacks. The Res2Net approach designs a residual-like connection between feature groups within one block, which increases the possible receptive fields and improves the system's detection generalizability. However, such a residual-like connection is performed by a direct addition between feature groups without channel-wise priority. We argue that the information across channels may not contribute to spoofing cues equally, and the less relevant channels are expected to be suppressed before adding onto the next feature group, so that the system can generalize better to unseen attacks. This argument motivates the current work that presents a novel, channel-wise gated Res2Net (CG-Res2Net), which modifies Res2Net to enable a channel-wise gating mechanism in the connection between feature groups. This gating mechanism dynamically selects channel-wise features based on the input, to suppress the less relevant channels and enhance the detection generalizability. Three gating mechanisms with different structures are proposed and integrated into Res2Net. Experimental results conducted on ASVspoof 2019 logical access (LA) demonstrate that the proposed CG-Res2Net significantly outperforms Res2Net on both the overall LA evaluation set and individual difficult unseen attacks, which also outperforms other state-of-the-art single systems, depicting the effectiveness of our method.

Index Terms: channel-wise gated Res2Net, anti-spoofing, synthetic speech detection, automatic speaker verification

1. Introduction

Spoofing attacks on automatic speaker verification (ASV) have attracted ever-increasing security concerns in recent years, as they pose serious threats to essential applications of ASV, such as e-banking authentication, device activation, etc. These attacks can be categorized into human impersonation [1, 2], audio replay [3, 4], synthetic speech [5, 6] and the recently emerged adversarial attacks [7–12].

In the midst of the arms race between attack and defense for ASV, the speech community has held several ASVspoof Challenges [13–15] to develop countermeasures mainly against audio replay, text-to-speech (TTS) and voice conversion (VC) attacks. ASVspoof 2019 [15] is the latest one that contains two sub-challenges: physical access (PA) and logical access (LA). PA considers spoofing attacks from replay while LA refers to attacks generated with TTS and VC techniques.

A model's generalizability to unseen spoofing attacks is challenging but essential for developing reliable countermeasures [16, 17]. To tackle this issue, previous efforts dedicated to the design of powerful system architectures [18–20] and generalized acoustic features [18, 21] have shown great enhancement of generalization to unseen PA attacks. However, the unseen nature of LA attacks has larger variations due to numerous available speech synthesis algorithms, and some of them are difficult to be detected, e.g. the A17 attack in the LA evaluation set [15]. State-of-the-art (SOTA) countermeasures may easily overfit to the training and development sets, and lack good generalizability to unseen LA attacks [17]. Hence, this work focuses on enhancing generalized detection of LA attacks.

Much promising effort has been dedicated to designing countermeasures against LA attacks [20, 22–25]. Das et al. [26] augments the training data based on signal companding methods to enhance generalization. The RawNet2 architecture [27] is applied to detect synthetic speech directly upon the raw speech waveform. Our earlier work [19] leverages the Res2Net architecture to improve the model's generalizability and demonstrates its superior detection accuracy on unseen LA attacks.

According to [19], Res2Net designs a residual-like connection between different feature groups within one block, which increases the possible receptive fields and helps the system generalize better to unseen attacks. However, such a residual-like connection is performed by a direct addition between feature groups without channel-wise priority. We argue that the information across channels within a group may not contribute to spoofing cues equally, and the less relevant channels are expected to be suppressed before adding to another feature group, so that the system can generalize better to unseen attacks.

From the above motivation, this work extends [19] and proposes a novel network architecture, i.e. channel-wise gated Res2Net (CG-Res2Net). CG-Res2Net modifies the Res2Net block and enables a channel-wise gating mechanism in the residual-like connection between feature groups. This gating mechanism selects the more relevant channels while suppresses the less relevant ones to dynamically adapt to unseen attacks, hence enhances the model's detection generalization. Specifically, we propose and compare three possible gating mechanisms: single-group channel-wise gate (SCG), multi-group channel-wise gate (MCG) and multi-group latent-space channel-wise gate (MLCG). SCG automatically learns a channel-wise gate by a fully-connected layer, based on the current feature group. MCG differs from SCG by additionally considering the next feature group information as reference to compute the gate. Finally, MLCG modifies MCG to firstly project the features of two groups into separate latent spaces, then compute the gate based on the two latent spaces. Three gating mechanisms are integrated with Res2Net to form SCG-Res2Net, MCG-Res2Net and MLCG-Res2Net, respectively. The proposed systems are evaluated on the ASVspoof 2019 LA par-

http://dx.doi.org/10.21437/Interspeech.2021-2125

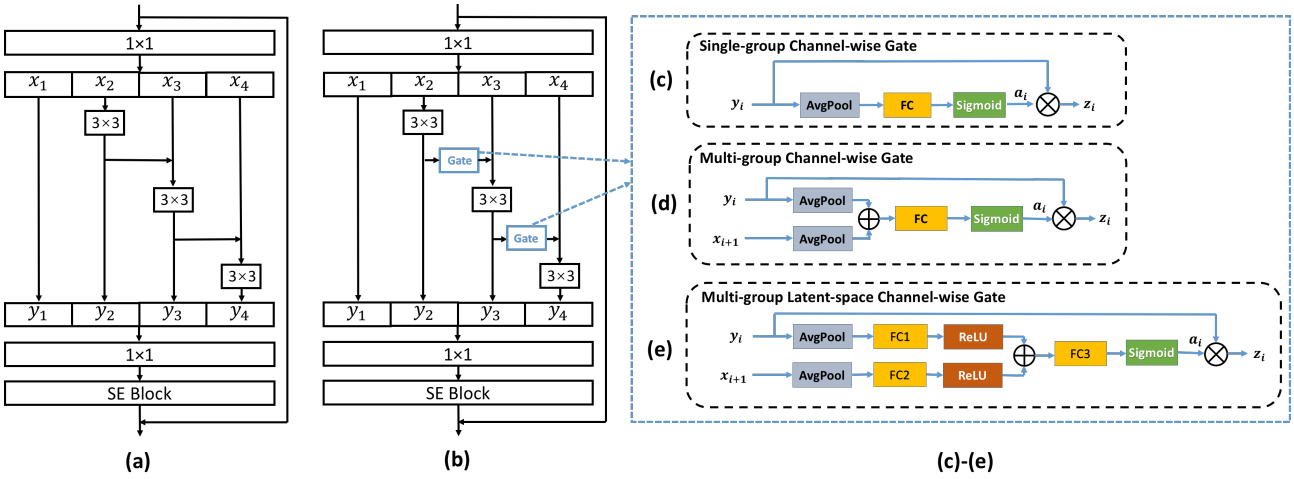

Figure 1: *Illustration of different structures in the modules: (a) Res2Net block; (b) CG-Res2Net block; (c) Single-group Channel-wise Gate; (d) Multi-group Channel-wise Gate; (e) Multi-group Latent-space Channel-wise Gate. (SE Block: the squeeze-and-excitation block [28]; AvgPool: average pooling; FC: fully-connected layer; ReLU and Sigmoid are the two activation functions.)*

tition in terms of the performance on the overall LA evaluation set as well as individual difficult unseen attacks. Experimental results demonstrate the effectiveness of the proposed gating mechanisms. All three proposed CG-Res2Net models outperform other single, SOTA systems on ASVspoof 2019 LA evaluation, depicting the promising performance of the CG-Res2Net models.

The contributions of this work include: 1) Proposing a novel CG-Res2Net architecture which can integrate one of three different channel-wise gating mechanisms into the Res2Net block; 2) Demonstrating that three proposed CG-Res2Net models outperform Res2Net on the overall LA evaluation set as well as individual difficult unseen attacks; 3) The proposed CG-Res2Net models outperform other single, SOTA systems on the ASVspoof 2019 LA evaluation set.

The rest of this paper is organized as follows: Section 2 illustrates the proposed CG-Res2Net architecture and three gating mechanisms. Experimental setup and results are demonstrated in Section 3 and Section 4, respectively. Finally, Section 5 concludes this work.

2. Approach

2.1. Channel-wise gated Res2Net

This section introduces the network structure of proposed CG-Res2Net. CG-Res2Net modifies the Res2Net block to enable a channel-wise gating mechanism in the residual-like connection between feature groups. The comparison between the structures of Res2Net and CG-Res2Net blocks is illustrated in Fig. 1 (a) and (b). After a 1×1 convolution, both models evenly split the input feature map X by the channel dimension into s subsets, denoted by x_i, where $i \in \{1, 2, ..., s\}$. We assume that $X \in \mathbb{R}^{sC \times D \times T}$ and each $x_i \in \mathbb{R}^{C \times D \times T}$, where C, D and T denote the dimensions of channel, spectrum and time, respectively.

Res2Net enables a direct addition between feature groups before a 3×3 convolution. Each y_i is derived as Eq. 1:

$$y_i = \begin{cases} x_i, & i = 1 \\ K_i(x_i), & i = 2 \\ K_i(x_i + y_{i-1}), & 2 < i \leq s \end{cases} \quad (1)$$

where each $K_i()$ denotes a convolutional function with a parameter size of 3×3. CG-Res2Net adopts a gating mechanism in the residual-like connection. Each y_i is derived as follows:

$$y_i = \begin{cases} x_i, & i = 1 \\ K_i(x_i), & i = 2 \\ K_i(x_i + z_{i-1}), & 2 < i \leq s \end{cases} \quad (2)$$

$$z_{i-1} = y_{i-1} \otimes a_{i-1} \quad (3)$$

where z_i scales y_i by a channel-wise gate $a_i \in \mathbb{R}^C$, and $\otimes$ denotes a channel-wise multiplication operation. We expect that such an gating mechanism gives priority to channels that contain most spoofing cues and suppresses the less relevant channels, then enhances the model's generalizability to unseen attacks.

This work proposes three novel channel-wise gating modules to be integrated with the Res2Net block, as shown in Fig. 1 (c)-(e). The detailed functionality of each module is demonstrated in Section 2.2. Our codes have been made open-source[1].

2.2. Channel-wise gating mechanism

2.2.1. Single-group channel-wise gate

As shown in Fig. 1 (c), the single-group channel-wise gate (SCG) automatically learns a gate a_i given the current feature group y_i. The mapping is achieved by a fully-connected layer. y_i is firstly squeezed to the channel dimension by averaging over the spectrum and time dimensions (Eq. 4), and then transformed by a fully-connected layer $W_{fc} \in \mathbb{R}^{C \times C}$ with a sigmoid activation function σ to derive the gate a_i (Eq. 5).

$$F_{ap}(y_i) = \frac{1}{D \times T} \sum_{d=1}^{D} \sum_{t=1}^{T} y_i(:, d, t) \quad (4)$$

$$a_i = \sigma[W_{fc}^T F_{ap}(y_i)] \quad (5)$$

2.2.2. Multi-group channel-wise gate

Since the residual-like connection is operated between y_i and x_{i+1}, it may be helpful to consider x_{i+1} as reference when ap-

[1]https://github.com/lixucuhk/Channel-wise-Gated-Res2Net

plying the gating mechanism. Thus we propose the multi-group channel-wise gate (MCG) where the channel-wise gate is derived from both information of y_i and x_{i+1}, as shown in Fig. 1 (d). This is formulated as Eq. 6:

$$a_i = \sigma\{W_{fc}^T[F_{ap}(y_i) \oplus F_{ap}(x_{i+1})]\} \quad (6)$$

where $\oplus$ is a concatenation function. y_i and x_{i+1} are squeezed to the channel dimension by F_{ap}, then concatenated together and transformed by a linear matrix $W_{fc} \in \mathbb{R}^{2C\times C}$ with sigmoid activation to derive a_i.

2.2.3. Multi-group latent-space channel-wise gate

With the consideration that x_{i+1} provides information as reference while y_i contains information to be re-scaled, the functionalities of them are not symmetric and it may be better to process them independently before concatenation. Thus we propose the multi-group latent-space channel-wise gate (MLCG) that separately project y_i and x_{i+1} into each own latent space before concatenation, as shown in Fig. 1 (e). Moreover, to limit model complexity and aid generalization, the latent space could have a reduced dimension with reduction ratio r. Specifically, y_i and x_{i+1} are squeezed by F_{ap}, then transformed by $W_{fc1} \in \mathbb{R}^{C\times\frac{C}{r}}$ and $W_{fc2} \in \mathbb{R}^{C\times\frac{C}{r}}$ with ReLU activation δ, respectively, as shown in Eq. 7 and 8. The squeezed channel information is concatenated together, to be transformed by $W_{fc3} \in \mathbb{R}^{\frac{2C}{r}\times C}$ with sigmoid activation to derive a_i, as shown in Eq. 9.

$$L_1(y_i) = \delta(W_{fc1}^T F_{ap}(y_i)) \quad (7)$$

$$L_2(x_{i+1}) = \delta(W_{fc2}^T F_{ap}(x_{i+1})) \quad (8)$$

$$a_i = \sigma\{W_{fc3}^T[L_1(y_i) \oplus L_2(x_{i+1})]\} \quad (9)$$

3. Experimental setup

Dataset: Experiments are conducted on the LA partition of ASVspoof 2019 corpus [15], which provides a standard dataset for anti-spoofing. The LA partition consists of bonafide audios and spoofed audios generated by different TTS and VC algorithms. The training and development subsets share the same attack algorithms, while the evaluation subset utilizes 11 unseen attacks (A07-A15, A17 and A18) and two attacks (A16 and A19) from the training set but trained with different data. The detailed information is shown in Table 1. Systems are evaluated by the tandem detection cost function (t-DCF) [15] and equal error rate (EER) [15]. The log-probability of the bonafide class is adopted as the score for t-DCF and EER computation.

Model configuration & Feature extraction: As an extension work of [19], this work adopts the best single system on the LA attacks in [19], i.e. Res2Net50 with squeeze-and-excitation (SE) block, as the baseline. The proposed three gating mechanisms are separately integrated into the baseline model for performance comparison. This work also leverages ResNet50 with SE block for comparison. As indicated in Fig. 1 (a) and (b), all ResNet, Res2Net and CG-Res2Net models in this work integrate the SE block without explicit denotation. The hyperparameter s in the Res2Net block and r in the MLCG mechanism are both experimentally set as 4. For acoustic features, this work adopts the constant-Q transform (CQT), which achieves the best results when incorporating with SERes2Net50 [19]. The CQT is extracted with 16ms step size, Hanning window, 9 octaves with 48 bins per octave. All samples are truncated along the time axis to reserve exactly 400 frames. The samples with less than 400 frames would be extended by repeating their contents.

Table 1: *Summary of the ASVspoof 2019 logical access corpus*

	#Bonafide	#Spoofed	Attack algorithms
Train	2,580	22,800	A01-A06
Dev.	2,548	22,296	A01-A06
Eval.	7,355	63,882	A07-A19

Table 2: *The EER (%) and t-DCF of different network architectures on the ASVspoof 2019 logical access.*

System	# params	Dev. Set		Eval. Set	
		EER (%)	t-DCF	EER (%)	t-DCF
ResNet50	1.09M	1.09	0.037	6.70	0.177
Res2Net50	0.92M	**0.43**	**0.014**	2.50	0.074
SCG-Res2Net50	0.95M	0.59	0.018	2.43	0.076
MCG-Res2Net50	0.96M	0.47	0.015	**1.78**	**0.052**
MLCG-Res2Net50	0.94M	0.86	0.027	2.15	0.069

Training strategy: The training strategy is identical with [19], where binary cross entropy is used to train all models. Adam [29] is adopted as the optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and initial learning rate being 3×10^{-4}. All models are trained for 20 epochs, and the model with lowest EER on development set is chosen to be evaluated.

4. Results

4.1. Effectiveness of channel-wise gated Res2Net

This section evaluates the effectiveness of proposed CG-Res2Net models for enhancing the generalizability to unseen attacks. Table 2 shows the EER and t-DCF performance of different systems. Notice that the robustness of a spoofing countermeasure depends on its effectiveness on detecting unseen attacks in the evaluation set. Consistent with [19], we observe that Res2Net50 performs much better than ResNet50 due to its efficient residual-like connection between feature groups. Compared with Res2Net50, all three CG-Res2Net models demonstrate a superior detection performance on unseen attacks in the evaluation set. Specifically, SCG-Res2Net50 performs slightly better than Res2Net50, while MCG-Res2Net50 and MLCG-Res2Net50 both show a significant improvements over Res2Net50. MCG-Res2Net50 achieves the most promising performance and outperforms Res2Net50 by a relative EER reduction of 28.8% and a relative t-DCF reduction of 29.7%. Such results verify the effectiveness of the proposed gating mechanisms, and the necessity of considering the reference information (x_{i+1} in Fig. 1d and Fig. 1e) in the gating module. MLCG-Res2Net50 did not outperform MCG-Res2Net50 on the overall attacks in the evaluation set, but it has better generalizability to the most difficult unseen attack (A17) in the evaluation set, as will be discussed in Section 4.2. Finally, it is also observed that the model complexity of CG-Res2Net models is comparable to that of Res2Net50 and smaller than that of ResNet50, which verifies the efficiency of the proposed gating mechanisms.

4.2. Detection performance on each unseen attack

As mentioned in Section 3, the LA evaluation set has 11 unseen attacks (A07-A15, A17 and A18) and two attacks (A16 and A19) from the training set but trained with different data. As reported in the official ASVspoof 2019 summary [15], A17 is the most difficult, such that most submitted systems failed

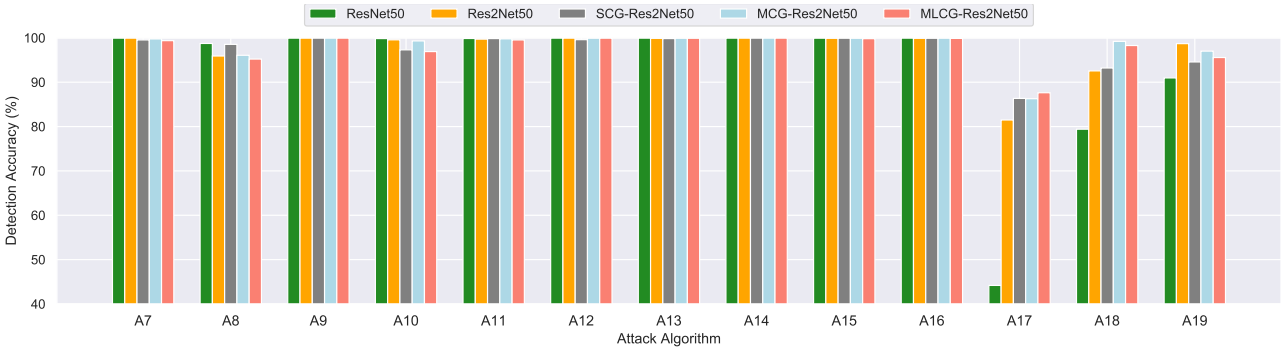

Figure 2: *The detection accuracy on each attack, given different network architectures with EER operating points. A16 and A19 are two attacks from the training set but trained with different training data. A17 is the most difficult unseen attack [15].*

to detect it. To perform a detailed system evaluation on each unseen attack, this section reports the detection accuracy on data of each attack, given the system's operating point in terms of EER, as shown in Fig. 2. It has been observed that A17 and A18 are two difficult unseen attacks. In particular, for the most difficult A17 attack, ResNet50 has a detection accuracy below 50% and Res2Net50 only achieves an accuracy of 81.48%. All the three proposed CG-Res2Net50 models outperform Res2Net50 by a large margin. MLCG-Res2Net50 achieves the highest detection accuracy of 87.63%, which outperforms Res2Net50 by 6.15% in absolute accuracy. For the A18 attack, ResNet50 and Res2Net50 achieve a detection accuracy of 79.41% and 92.55%, respectively. The proposed MCG-Res2Net50 and MLCG-Res2Net50 outperform them with a detection accuracy of 99.21% and 98.27%, respectively. These observations verify the effectiveness of MCG-Res2Net50 and MLCG-Res2Net50 on generalization to difficult unseen attacks. For other easily detectable attacks, CG-Res2Net50 models perform comparably well with Res2Net50.

4.3. Comparison with the state-of-the-art single systems

This section compares the proposed CG-Res2Net models with some reported single, SOTA systems evaluated on the ASVspoof 2019 LA partition, including systems that are submitted to the ASVspoof 2019 competition and those reported in works afterwards (according to our best knowledge). The EER and t-DCF performance are shown in Table 3. The systems are denoted by a name that encodes the input features, system architecture and loss criteria.

We observe that existing efforts dedicated into acoustic features and data augmentation [23, 26, 30], system architecture [19, 27, 31] and loss criteria [31] have achieved very promising performance. As an extension of [19], the proposed CG-Res2Net models outperform other SOTA systems, depicting the effectiveness of the gating mechanisms within the Res2Net block. Moreover, the proposed CG-Res2Net models can be utilized as a backbone network, to be integrated with other effective strategies, e.g. loss criteria, for stronger generalization to unseen attacks.

Table 3: *Performance comparison of CG-Res2Net models to some known state-of-the-art single systems on the ASVspoof 2019 LA evaluation set.*

System	EER (%)	t-DCF
Spec+ResNet+CE [24]	11.75	0.216
Spec+ResNet+CE [23]	9.68	0.274
MFCC+ResNet+CE [23]	9.33	0.204
CQCC+ResNet+CE [23]	7.69	0.217
LFCC+LCNN+A-softmax [20]	5.06	0.100
FFT+LCNN+A-softmax [20]	4.53	0.103
RawAudio+RawNet2+CE [27]	4.66	0.129
FG-CQT+LCNN+CE [30]	4.07	0.102
Spec+LCGRNN+GKDE-Softmax [31]	3.77	0.084
Spec+LCGRNN+GKDE-Triplet [31]	3.03	0.078
DASC-CQT+LCNN+CE [26]	3.13	0.094
CQT+SERes2Net50+CE [19]	2.50	0.074
Ours: CQT+SCG-Res2Net50+CE	2.43	0.076
Ours: CQT+MCG-Res2Net50+CE	**1.78**	**0.052**
Ours: CQT+MLCG-Res2Net50+CE	2.15	0.069

5. Conclusion

This work proposes a novel network architecture, i.e. CG-Res2Net, to enhance the model's generalization to unseen attacks. It modifies the Res2Net block to enable a channel-wise gating mechanism in the residual-like connection between feature groups. Such a gating mechanism dynamically selects channel-wise features based on the input, to suppress the less relevant channels and enhance the detection generalizability. Three gating mechanisms are proposed and verified to be effective in enhancing generalization. In terms of overall performance on the LA evaluation set, MCG-Res2Net achieves the best performance and outperforms the Res2Net by a relative EER reduction of 28.8% and a relative t-DCF reduction of 29.7%. On the most difficult unseen attack (A17), MLCG-Res2Net achieves the best performance, which outperforms Res2Net by 6.15% absolute detection accuracy. The proposed CG-Res2Net models outperform other single, SOTA systems on the ASVspoof 2019 LA evaluation, depicting the effectiveness of our method. Given the superior generalizability of CG-Res2Net, future work will investigate its applicability to other speech applications.

6. Acknowledgement

We sincerely thank Mr. Zhiyuan Peng from the Chinese University of Hong Kong for some meaningful discussions. This work is supported by HKSAR Government's Research Grants Council General Research Fund (Project No. 14208718).

7. References

[1] V. Vestman, T. Kinnunen, R. G. Hautamäki, and M. Sahidullah, "Voice mimicry attacks assisted by automatic speaker verification," *Computer Speech & Language*, vol. 59, pp. 36–54, 2020.

[2] T. Kinnunen, R. G. Hautamäki, V. Vestman, and M. Sahidullah, "Can we use speaker recognition technology to attack itself? enhancing mimicry attacks using automatic target speaker selection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6146–6150.

[3] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *speech communication*, vol. 66, pp. 130–153, 2015.

[4] B. Chettri, S. Mishra, B. L. Sturm, and E. Benetos, "A study on convolutional neural network based end-to-end replay anti-spoofing," *arXiv preprint arXiv:1805.09164*, 2018.

[5] V. Shchemelinin and K. Simonchik, "Examining vulnerability of voice verification systems to spoofing attacks by means of a TTS system," in *ICSC*. Springer, 2013, pp. 132–137.

[6] T. Kinnunen, Z. Wu, K. Lee, F. Sedlak, E. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *ICASSP*. IEEE, 2012, pp. 4401–4404.

[7] R. K. Das, X. Tian, T. Kinnunen, and H. Li, "The attacker's perspective on automatic speaker verification: An overview," *Proc. INTERSPEECH*, 2020.

[8] H. Wu, S. Liu, H. Meng, and H.-y. Lee, "Defense against adversarial attacks on spoofing countermeasures of asv," in *ICASSP*, 2020.

[9] X. Li, J. Zhong, X. Wu, J. Yu, X. Liu, and H. Meng, "Adversarial attacks on GMM i-vector based speaker verification systems," in *ICASSP*. IEEE, 2020, pp. 6579–6583.

[10] H. Wu, X. Li, A. T. Liu, Z. Wu, H. Meng, and H.-y. Lee, "Adversarial defense for automatic speaker verification by cascaded self-supervised learning models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6718–6722.

[11] Z. Peng, X. Li, and T. Lee, "Pairing weak with strong: Twin models for defending against adversarial attack on speaker verification," *Proc. INTERSPEECH*, 2021.

[12] H. Wu, X. Li, A. T. Liu, Z. Wu, H. Meng, and H.-y. Lee, "Adversarial defense for automatic speaker verification by self-supervised learning," *arXiv preprint arXiv:2106.00273*, 2021.

[13] 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 *Proc. INTERSPEECH*, 2015.

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

[15] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. INTERSPEECH*, 2019.

[16] X. Li, N. Li, J. Zhong, X. Wu, X. Liu, D. Su, D. Yu, and H. Meng, "Investigating robustness of adversarial samples detection for automatic speaker verification," *Proc. INTERSPEECH*, 2020.

[17] A. Nautsch, X. Wang, N. Evans, T. 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*, 2021.

[18] X. Cheng, M. Xu, and T. F. Zheng, "Replay detection using cqt-based modified group delay feature and resnewt network in asvspoof 2019," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 540–545.

[19] 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.

[20] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC anti-spoofing systems for the asvspoof2019 challenge," *Proc. INTERSPEECH*, pp. 1033–1037, 2019.

[21] W. Cai, H. Wu, D. Cai, and M. Li, "The DKU replay detection system for the asvspoof 2019 challenge: On data augmentation, feature representation, classification, and fusion," *Proc. INTERSPEECH*, pp. 1023–1027, 2019.

[22] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Interspeech*, 2012.

[23] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," *Proc. INTERSPEECH*, pp. 1078–1082, 2019.

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

[25] 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.

[26] R. K. Das, J. Yang, and H. Li, "Data augmentation with signal companding for detection of logical access attacks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6349–6353.

[27] 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.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ICLR*, 2015.

[30] Z. Wu, R. K. Das, J. Yang, and H. Li, "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks," *Proc. INTERSPEECH*, pp. 1101–1105, 2020.

[31] 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.