INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Advanced RawNet2 with Attention-based Channel Masking for Synthetic Speech Detection

Jing Li[1], *Yanhua Long*[1,2*], *Yijie Li*[2], *Dongxing Xu*[2]

[1]Shanghai Engineering Research Center of Intelligent Education and Bigdata ,
Shanghai Normal University, Shanghai, China
[2]Unisound AI Technology Co., Ltd., Beijing, China

1000511818@smail.shnu.edu.cn, yanhua@shnu.edu.cn, liyijie@unisound.com, xudongxing@unisound.com

Abstract

Automatic speaker verification (ASV) systems are often vulnerable to spoofing attacks, particularly unseen attacks. Due to the diversity of text-to-speech and voice conversion algorithms, how to improve the generalization ability of synthetic speech detection systems is a challenging issue. To address this issue, we propose an advanced RawNet2 (ARawNet2) by introducing an attention-based channel masking (ACM) block to improve the RawNet2, with three main components: the squeeze-and-excitation, the channel masking, and a global-local feature aggregation. The effectiveness of the proposed system is evaluated on both the ASVspoof 2019 and ASVspoof 2021 datasets. Specifically, the ARawNet2 achieves an EER of 4.61% on the ASVspoof 2019 logical access (LA) task, and on the ASVspoof 2021 LA and speech deepfake (DF) tasks, it achieves EER of 8.36% and 19.03%, which obtains relative 12.00% and 14.97% EER reductions over the RawNet2 baseline, respectively.

Index Terms: synthetic speech detection, automatic speaker verification, RawNet2

1. Introduction

In recent years, automatic speaker verification (ASV) [1], aiming at verifying a claimed speaker identity through a spoken utterance, has been widely applied in a variety of domains, such as financial privacy security, personalized service, and audio forensics. ASV systems, however, are vulnerable to spoofing attacks in realistic scenarios, which include four major classes of attacks: impersonation, replay [2], text-to-speech (TTS) [3] and voice conversion (VC) [4]. In order to advance the development of reliable ASV systems, ASVspoof challenge [5, 6, 7, 8] has been held biennially since 2015, indicating that the investigation of countermeasure systems for spoofing attacks is becoming increasingly important.

Although the techniques of spoofing attacks detection have made great progress, the generalization ability of reliable countermeasures against unseen spoof attacks is still a challenging issue. Some countermeasures that achieve good performance in the development set may drastically degrade in the evaluation set. There are five major methods for improving the generalization ability of countermeasures: data augmentation [9, 10], feature engineering [11, 12, 13], system modeling [14, 15], loss function [16] and ensembles [10, 17]. In this study, we focus on the system modeling to improve the system robustness and generalization ability.

In the literature, many end-to-end neural networks for synthetic speech detection have been designed to improve the system modeling ability to defend against unseen spoofing attacks [18, 19, 20]. For example, the RawGAT-ST [18] and AASIST [19], which are based on RawNet2 [21] and graph attention network, have proposed a spectral-temporal attention module to extract the discriminative cues between bonafide and spoofed speech in different temporal intervals and spectral sub-bands. Although their experimental results demonstrate the effectiveness on ASVspoof 2019 LA task, they didn't perform the investigation of model generalization ability on cross-domain tasks. Another work in [20], improved the model cross-domain robustness, by replacing the RawNet2 in the AASIST framework with various self-supervised learning (SSL) front-ends, such as wav2vec [22], etc. However, these pre-trained SSL front-ends require large amounts of external speech data that forbidden by the ASVspoof 2021 challenge [23], or it is difficult to collect under most real-world application scenarios.

In this study, we also aim to improve the generalization and robustness of the end-to-end ASVspoof architecture, especially for the cross-domain tasks with different unseen attack algorithms. Based on the conventional RawNet2 [21], we propose an advanced RawNet2 (ARawNet2) by introducing a simple and effective attention-based channel masking (ACM) block for synthetic speech detection. The ACM block is composed of three key components: 1) The squeeze-and-excitation (SE). It is an effective block that has been widely used in speaker verification committee[24, 25]. In ARawNet2, we use this idea to recalibrate the channel-wise correlation of the high-level global acoustic feature maps; 2) The channel masking. It is designed to randomly mask partial features for enhancing the robustness of the model; 3) The global-local feature aggregation. It is proposed to fully exploit the complementarity between the global and local deep feature maps. All our experiments are performed on both the ASVspoof 2019 [7] and ASVspoof 2021 [8] datasets, where their evaluation sets include many previously unseen spoofing attacks that differ from those in the training and development sets. Experimental results show that, the proposed system significantly outperforms the official ASVspoof challenge baselines. On the ASVspoof 2019 LA task, the proposed ARawNet2 achieves an EER of 4.61%, and on the ASVspoof 2021 LA and DF tasks, it achieves relative EER reductions of 12.00% and 14.97% over the RawNet2 baseline, respectively.

2. RawNet2

RawNet2, proposed in [21], is an end-to-end neural network architecture that has been taken as the official baseline system of ASVspoof 2021 challenge. It mainly consists of two parts: a frame-level feature extractor and a classifier. The feature extractor is composed of four components: a sinc layer, six resid-

*Yanhua Long is the corresponding author, she is also with the Key Innovation Group of Digital Humanities Resource and Research, Shanghai Normal University. The work is supported by the National Natural Science Foundation of China (Grant No.62071302)

10.21437/Interspeech.2023-542

ual blocks [26] with 1-dimensional convolution layers, six feature map scaling (FMS) [27] blocks and a gated recurrent unit (GRU) [28] layer. First, given the raw waveform as input, the sinc layer convolves the waveform with a set of parametrized sinc functions [29] that implement 128 mel-scale band-pass filters, which forces the network to focus on high-level tunable parameters with broad impact on the shape and bandwidth of the resulting filter. Next, the extracted local acoustic features from sinc layer are fed into the six residual blocks to extract frame-level speaker representations. In addition, in order to derive more discriminative speaker information, filter-wise feature map scaling is performed on the output of each residual block. Next, the GRU layer with 1024 hidden nodes is used to aggregate frame-level representations into a single utterance-level representation. Finally, the GRU output is followed by a classifier with two fully-connected layers and a softmax activation function, which can predict whether the input speech is bonafide or spoofed. More details about the RawNet2 system are described in [21].

3. Proposed methods

This section provides a detailed introduction of the architecture of our proposed ARawNet2 model, the key components of ACM block and the implementation of channel masking.

3.1. Architecture

The overall architecture of our proposed ARawNet2 model is shown in the Fig.1, which is mainly designed based on the original RawNet2 system [21], consisting of a frame-level encoder and a classifier. Compared with the original RawNet2, there are three major different points: 1) All 1-dimensional convolution layers in six residual blocks have been replaced with a 2-dimensional convolution layers, by adding a new channel dimension into the original sinc layer to transfer the original three-dimensional feature map into a four-dimensional one; 2) We remove all feature map scaling operations after the output of all residual blocks; 3) All FMS blocks in RawNet2 are replaced with our proposed attention-based channel masking (ACM) blocks that are shown in Fig.1.

3.2. Attention-based channel masking block

The whole structure of the proposed attention-based channel masking (ACM) block is shown in Fig.2. Our idea is motivated by the works in [18, 30], when the cochlea receives bonafide or spoofed audio, the attention mechanism and masking effects of the human auditory system are able to automatically focus on the local and discriminative feature information of temporal intervals and the spectral sub-bands in the global feature information, which makes us to distinguish the bonafide or spoofed audio accurately. Therefore, in the proposed ACM block, we design three key components to simulate the perception ability of human cochlea: the SE block, channel masking, and global-local feature aggregation.

First, the SE block [31], acting as a simple attention mechanism in the ACM, is used to recalibrate the channel-wise discriminative information in both different temporal intervals and spectral sub-bands between bonafide and spoofed speech. Specifically, given the 3-dimensional feature map $X \in \mathbb{R}^{C\times T\times F}$, $X_{GAP} \in \mathbb{R}^{C\times 1}$, derived from the operation of global average pooling on time and frequency dimension, is defined as:

$$X_{GAP} = \frac{1}{T\times F}\sum_{i=1}^{T}\sum_{j=1}^{F} X_{ij} \tag{1}$$

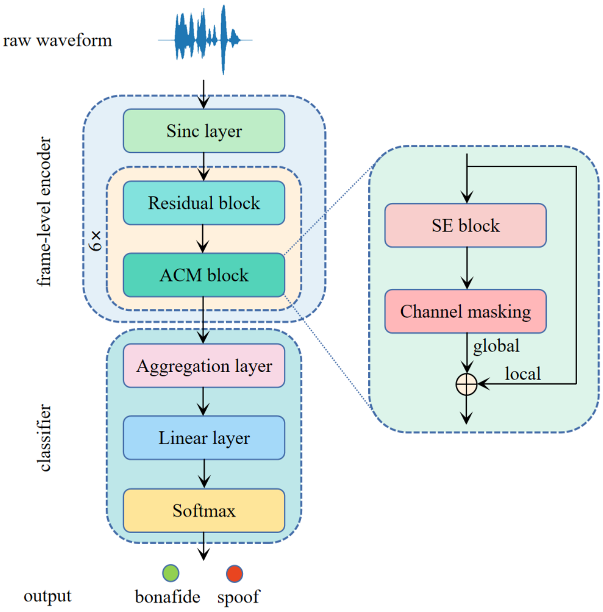

Figure 1: *Overall architecture of the proposed ARawNet2. The proposed Attention-based Channel Masking (ACM) block consists of SE block, channel masking, and global-local feature aggregation.*

where C, T, and F represent the number of channels, frames, and frequency bins, respectively. The recalibrated feature map U with channel-wise interdependence is formulated as:

$$U = X \otimes \delta(W_2^{C\times\frac{C}{r}} \times (relu(W_1^{\frac{C}{r}\times C} \times X_{GAP}))) \tag{2}$$

where × and $\otimes$ refer to matrix multiplication and element-wise multiplication. $W_1^{\frac{C}{r}\times C}$ and $W_2^{C\times\frac{C}{r}}$ are the weights of two fully-connected layers. r is a dimensionality-reduction ratio of the number of channels to control the network parameters. $relu$ and δ denote the Relu activation function and a sigmoid function to scale the channel-wise weights.

Then, as shown in the below part of Fig.2, the output of SE block is followed by the channel masking that is designed to dynamically mask partial global features for enhancing the robustness of our model. The details of the implementation of channel masking will be described in section 3.3.

Finally, each residual block is applied to model the local features with more discriminative cues between temporal intervals and spectral sub-bands by 2-dimensional convolution layers, while the output of channel masking contains more global information derived from the features with the recalibrated channel-wise interdependence. Therefore, to exploit the complementarity between global and local features, we insert a global-local feature aggregation operation in the ACM block to get the final feature matrix Q by element-wise addition between global feature map M and local feature map X, as shown in the Fig.2.

3.3. Implementation of channel masking

Inspired by the frequency masking [32, 33] and temporal masking [34], in order to simulate the masking effects of the human auditory system, in this study, we design a channel masking operation to enhance the model learn more robust deep feature representations. The details of the implementation of channel masking are as follows. Specifically, the channel masking acts

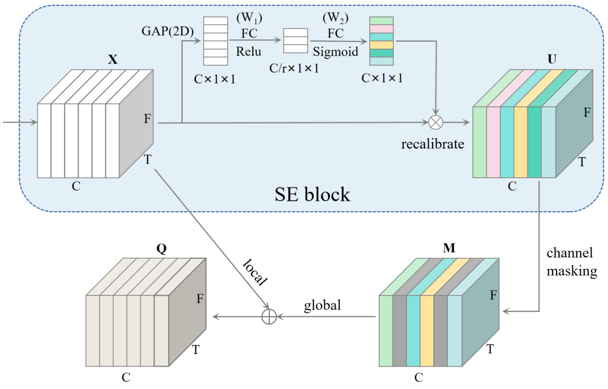

Figure 2: *Illustration of attention-based channel masking (ACM) block.*

as randomly discarding a range of $[C_1, C_2)$ features of channel dimension for the output of SE block (for example, $[2, 4)$ means that the features of the 2nd and the 3rd channels of U will be masked by setting them to zero). During our experiments, we first randomly select a number, f, from a unified distribution between 0 and F, as the total channels that we expect to mask, where $F = 4$ refers to the maximum pre-defined masking channels. Then, the first position of the masking channel C_1 is chosen randomly from $[0, C - f]$, where C represents the total number of the channel dimension of the feature map U, and $C_2 = C_1 + f$. Finally, in all our experiments, the above operations will be implemented twice during each training epoch before the global feature matrix M is finalized. The channel masking is only applied during model training.

4. Experimental setup

4.1. Dataset and evaluation metric

All our experiments are conducted on the datasets of both ASVspoof 2019 [7] and ASVspoof 2021 challenges [8]. Here we focus on two tasks of synthetic speech detection: LA and DF. All our systems and official baselines in these two ASVspoof challenges are trained only using the LA training set of ASVspoof 2019. However, the resulted models are used to evaluate three different evaluation sets from the ASVspoof 2019 LA, ASVspoof 2021 LA and DF to measure their generalization ability under cross-domain/cross-dataset scenarios. The detailed description of the dataset is shown in Table 1.

It is worth noting that those unseen spoofing attacks of three evaluation sets that generated from diverse algorithms are becoming increasingly complex, which differ from those previously existing in training and development sets. The training and development sets of ASVspoof 2019 LA contain bonafide and spoofed data that generated with 6 different algorithms (A01∼A06), while there are 13 different algorithms (A07∼A19) to generate spoofed data of ASVspoof 2019 LA evaluation set. Although the TTS, VC, and hybrid spoofing attacks of ASVspoof 2021 LA evaluation set are the same as those from ASVspoof 2019 LA evaluation set, the difference is the former is degraded by different unknown transmission channels, resulting in a total of 181,566 trials. As for the evaluation set of ASVspoof 2021 DF task, there are more than 100 undisclosed attack algorithms with 611,829 test trials in total. We use equal error rate (EER) [35] as a metric to evaluate the performance of our model.

Table 1: *Summary of the dataset of ASVspoof 2019 LA, ASVspoof 2021 LA evaluation set and ASVspoof 2021 DF evaluation set. 'Train' and 'Dev' refer to training and development sets of ASVspoof 2019 LA, respectively. 'Eval(19LA)', 'Eval(21LA)' and 'Eval(21DF)' represent the evaluation set of ASVspoof 2019 LA, ASVspoof 2021 LA and DF, respectively.*

Subset	Bonafide	Spoof	Attacks
	#Utts	#Utts	#Attack Types
Train	2,580	22,800	6(A01∼A06)
Dev	2,548	22,296	6(A01∼A06)
Eval(19LA)	7,355	63,882	13(A07∼A19)
Eval(21LA)	18,452	163,114	13(A07∼A19)
Eval(21DF)	22,617	589,212	>100(undisclosed)

4.2. Training setup

Our proposed model, ARawNet2, directly uses raw waveforms as input. The duration of all utterances is fixed into 4 seconds by either cropping long utterances or concatenating short utterances [21]. No data augmentation is performed in our experiments. The training set and the development set of ASVspoof 2019 LA are used to train all of our models and to select the best model for system evaluation, respectively. As shown in Table 1, the number of bonafide and spoofed data is heavily imbalanced. Therefore, a weighted cross entropy (WCE) loss function is applied to train our model, where the ratio of weights of bonafide and spoofed trials is set to $9:1$. The dimensionality-reduction ratio r is set to 16. All systems are optimized with ADAM optimiser [36] using a fixed learning rate of 0.0001, a mini-batch size of 8, over 100 epochs and a weight decay of 0.0001.

5. Results and discussion

5.1. Overall results under cross-dataset conditions

Table 2 presents the overall results comparison between our proposed ARawNet2 and the official baseline results of ASVspoof 2019 and 2021 challenges on three different evaluation sets. The CQCC-GMM [11], LFCC-GMM [37], LFCC-LCNN [38] and RawNet2 [21] are the official baseline systems in ASVspoof challenges. The systems trained with LFCC-LCNN and RawNet2 algorithms are only the baseline systems of ASVspoof 2021 challenge, while CQCC-GMM and LFCC-GMM are the both used to build baseline systems of ASVspoof 2019 and ASVspoof 2021 challenges. It is worth noting that all these official baselines are trained on the training set of ASVspoof 2019 LA, that's to say, all the EERs in Table 2 are achieved from the models that trained on the same training set.

By comparing the results of official systems, it's clear that all EERs on the ASVspoof 2019 LA evaluation set are much lower than the ones on other two evaluation sets of ASVspoof 2021. It indicates that the ASVspoof 2019 LA evaluation set is much easier than ASVspoof 2021 LA and DF evaluation sets to the CQCC-GMM and LFCC-GMM systems. This is because the condition/spoofing algorithms between the training and evaluation sets are very close, while the synthetic spoofing algorithms used in ASVspoof 2021 LA and DF evaluation sets deviate far from the ASVspoof 2019 LA training set. Moreover, this finding also tells us that, the models trained on ASVspoof 2019 LA can not generalize well to other evaluation conditions, which demonstrate that they are very vulnerable to cross-dataset application scenarios.

In addition, only from the results of ASVspoof 2021 LA and DF evaluation sets, we see that LFCC-LCNN and RawNet2 achieve similar results, but significantly outperform the CQCC-GMM and LFCC-GMM systems. It indicates that the RawNet2 is a strong baseline that can provide a relatively fair comparison with the proposed ARawNet2. Moreover, by comparing all the results in Table 2, we see the proposed ARawNet2 achieves the best performance, and its effectiveness can generalize well on cross evaluation datasets. Specifically, the proposed ARawNet2 achieves an EER of 4.61% on the ASVspoof 2019 LA task. On the ASVspoof 2021 LA and DF tasks, the ARawNet2 achieves EER of 8.36% and 19.03%, which outperforms the RawNet2 baseline by relative 12.00% and 14.97%, respectively.

Table 2: *Performance (EER%) comparison between our proposed system and official baseline systems on ASVspoof 2019 LA, ASVspoof2021 LA and DF evaluation sets. '19LA', '21LA', and '21DF' represent the ASVspoof 2019 LA, ASVspoof 2021 LA, and ASVspoof2021 DF evaluation set, respectively. LFCC-LCNN and RawNet2 are not the baseline systems of ASVspoof 2019 LA, and we denote the absent result as "-" in the table.*

Method	19LA	21LA	21DF
CQCC-GMM[11]	9.57	15.62	25.56
LFCC-GMM[37]	8.09	19.30	25.25
LFCC-LCNN[38]	-	9.26	23.48
RawNet2[21]	-	9.50	22.38
ARawNet2(ours)	**4.61**	**8.36**	**19.03**

5.2. Results of ablation study

As shown in Table 3, the ablation experiments are performed on the evaluation set of ASVspoof 2021 DF task to investigate how the number of the operation times of channel masking influence the performance of the proposed ARawNet2. It's clear to find that, as the number of operation times of channel masking increases, the value of EER decreases initially and then increases. Specifically, when the channel masking is implemented twice during training, the ARawNet2 achieves the best performance with an EER of 19.03%. This further indicates that the number of channel masking operation times may lead to the perturbation of features in the channel dimension by masking partial features to enhance the robustness of the ARawNet2 model to some extent.

Table 3: *Performance of ARawNet2 with different operation times of channel masking on the evaluation set of ASVspoof 2021 DF task.*

Method	# Operation times of channel masking			
	0	1	2	3
ARawNet2	21.70	20.37	**19.03**	21.33

The experimental results of the ablation study of three key components of the ACM block in ARawNet2 are also performed on the evaluation set of ASVspoof2021 DF, as shown in Table 4. In this table, we denote the proposed ARawNet2 without ACM block as RawNet2*. Compared with the EER of 22.38% of RawNet2 in [21], owing to removing the FMS operation in original RawNet2, the RawNet2* just achieves the EER of 26.39% on the evaluation set of ASVspoof 2021 DF.

However, when the SE block is added to the RawNet2*, the system achieves the EER of 22.09%, which demonstrates that the SE block could make full use of the channel-wise interdependence to capture the salient features and discard the insignificant features, leading to more discriminative representations. When we only add the channel masking into RawNet2*, the EER of the system is reduced to 20.99%, which clearly indicates the channel masking is significant to construct the ACM block. Then, when we insert the SE block and channel masking into the RawNet2* at the same time, performance degrades by 3.90% (22.09% vs. 25.99%) compared with only inserting SE block into RawNet2*. We suspect that the combination of SE block and channel masking may discard partial important features, leading to degraded performance. When compared with the last line result (19.03%) of ARawNet2, the above analysis shows that the global-local feature aggregation is essential for constructing the ACM block to exploit the complementary information between global and local deep feature representations. Finally, when the SE block and global-local feature aggregation are added into the RawNet2* at the same time, compared with the EER of the proposed ARawNet2 (19.03%), the new system only achieves an EER of 21.70%, which further indicates channel masking is of primary importance for the ACM block to enhance the robustness of the ARawNet2. Given the above discussions, it is not difficult to find the fact that all three components of the ACM block are significant for the proposed ARawNet2 to improve the generalization ability.

Table 4: *EER(%) on ablation studies of the three key components of ACM block in ARawNet2 on the evaluation set of ASVspoof2021 DF. SE, CS, and GLA denote the SE block, channel masking, and global-local feature aggregation, respectively. RawNet2* denotes the proposed ARawNet2 without ACM block.*

Method	21DF
RawNet2[21]	22.38
RawNet2*	26.39
RawNet2*+SE	22.09
RawNet2*+CS	20.99
RawNet2*+SE+GLA	21.70
RawNet2*+SE+CS	25.99
RawNet2*+SE+CS+GLA(**ARawNet2**)	**19.03**

6. Conclusion

In this paper, we propose an advanced RawNet2 (ARawNet2) by introducing an attention-based channel masking (ACM) block into the original RawNet2 to improve the performance and robustness of synthetic speech detection system. The idea of ACM block is inspired by the attention mechanism and masking effects of the human auditory system, and three key components are specially designed to enable the ACM block with auditory perception ability: SE block, channel masking, and global-local feature aggregation. All our experiments are conducted on the datasets of ASVspoof 2019 and ASVspoof 2021 challenges. Results show that the proposed ARawNet2 outperforms four types of strong baseline systems, and the significant performance improvements demonstrate the ability of generalization and robustness of the proposed method against unseen spoofing attacks on cross-dataset. Furthermore, the ablation study shows that the combination of SE block, channel masking, and global-local feature aggregation in the ACM block is very important to improve the ARawNet2, and the operation times of channel masking that applied on the outputs of SE block also influences the performance of ARawNet2 to some extent.

7. References

[1] T. H. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech Commun.*, vol. 52, pp. 12–40, 2010.

[2] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, “Audio replay attack detection with deep learning frameworks.” in *Proc. INTERSPEECH*, 2017, pp. 82–86.

[3] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, “TTS synthesis with bidirectional LSTM based recurrent neural networks,” in *Proc. INTERSPEECH*, 2014, pp. 1964–1968.

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

[5] Z. Wu, T. H. Kinnunen, N. W. D. 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, pp. 2037–2041.

[6] 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, pp. 2–6.

[7] 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.

[8] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, “ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection,” in *Proc. ASVspoof Challenge workshop*, 2021, pp. 47–54.

[9] R. K. Das, “Known-unknown Data Augmentation Strategies for Detection of Logical Access, Physical Access and Speech Deepfake Attacks: ASVspoof 2021,” in *Proc. ASVspoof Challenge workshop*, 2021, pp. 29–36.

[10] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, and G. Lavrentyeva, “STC Antispoofing Systems for the ASVspoof2021 Challenge,” in *Proc. ASVspoof Challenge workshop*, 2021, pp. 61–67.

[11] M. Todisco, H. Delgado, and N. Evans, “Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification,” *Computer Speech Language*, vol. 45, pp. 516–535, 2017.

[12] R. K. Das, J. Yang, and H. Li, “Long range acoustic and deep features perspective on asvspoof 2019,” in *Proc. ASRU*, 2019, pp. 1018–1025.

[13] S. Jelil, R. K. Das, S. M. Prasanna, and R. Sinha, “Spoof Detection Using Source, Instantaneous Frequency and Cepstral Features,” in *Proc. INTERSPEECH*, 2017, pp. 22–26.

[14] M. Alzantot, Z. Wang, and M. B. Srivastava, “Deep Residual Neural Networks for Audio Spoofing Detection,” in *Proc. INTERSPEECH*, 2019, pp. 1078–1082.

[15] 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.

[16] 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.

[17] Z. Lei, H. Yan, C. Liu, M. Ma, and Y. Yang, “Two-Path GMM-ResNet and GMM-SENet for ASV Spoofing Detection,” in *Proc. ICASSP*, 2022, pp. 6377–6381.

[18] H. Tak, J. weon Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, “End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection,” in *Proc. ASVspoof Challenge workshop*, 2021, pp. 1–8.

[19] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, “AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks,” in *Proc. ICASSP*, 2022, pp. 6367–6371.

[20] H. Tak, M. Todisco, X. Wang, J. weon Jung, J. Yamagishi, and N. Evans, “Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation,” in *Proc. Odyssey*, 2022, pp. 112–119.

[21] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, “End-to-end anti-spoofing with rawnet2,” in *Proc. ICASSP*, 2021, pp. 6369–6373.

[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[23] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang *et al.*, “ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan,” *arXiv preprint arXiv:2109.00535*, 2021.

[24] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification,” in *Proc. INTERSPEECH*, 2020, pp. 3830–3834.

[25] D. Michelsanti and Z.-H. Tan, “Conditional Generative Adversarial Networks for Speech Enhancement and Noise-Robust Speaker Verification,” in *Proc. INTERSPEECH*, 2017, pp. 2008–2012.

[26] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. CVPR*, 2016, pp. 770–778.

[27] J. weon Jung, S. bin Kim, H. jin Shim, J. ho Kim, and H.-J. Yu, “Improved RawNet with Feature Map Scaling for Text-Independent Speaker Verification Using Raw Waveforms,” in *Proc. INTERSPEECH*, 2020, pp. 1496–1500.

[28] J. weon Jung, H.-S. Heo, J. ho Kim, H. jin Shim, and H.-J. Yu, “RawNet: Advanced End-to-End Deep Neural Network Using Raw Waveforms for Text-Independent Speaker Verification,” in *Proc. INTERSPEECH*, 2019, pp. 1268–1272.

[29] M. Ravanelli and Y. Bengio, “Speaker recognition from raw waveform with sincnet,” in *Proc. SLT*, 2018, pp. 1021–1028.

[30] W. A. Yost, “Pitch perception,” *Attention, Perception, & Psychophysics*, vol. 71, no. 8, pp. 1701–1715, 2009.

[31] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 2011–2023, 2017.

[32] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition,” in *Proc. INTERSPEECH*, 2019, pp. 2613–2617.

[33] I.-Y. Kwak, S. Choi, J. Yang, Y. Lee, S. Han, and S. Oh, “Low-Quality Fake Audio Detection through Frequency Feature Masking,” in *Proc. DDAM workshop*, 2022, pp. 9–17.

[34] S. Yadav and A. Rai, “Frequency and temporal convolutional attention for text-independent speaker recognition,” in *Proc. ICASSP*, 2020, pp. 6794–6798.

[35] J.-M. Cheng and H.-C. Wang, “A method of estimating the equal error rate for automatic speaker verification,” in *Proc. ISCSLP*, 2004, pp. 285–288.

[36] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[37] M. Sahidullah, T. Kinnunen, and C. Hanilçi, “A comparison of features for synthetic speech detection,” in *Proc. INTERSPEECH*, 2015, pp. 2087–2091.

[38] X. Wang and J. Yamagishi, “A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection,” in *Proc. INTERSPEECH*, 2021, pp. 4259–4263.