Multi-perspective Information Fusion Res2Net with Random Specmix for Fake Speech Detection

Shunbo Dong[1,†], Jun Xue[1,†], Cunhang Fan[1,*], Kang Zhu[1], Yujie Chen[1] and Zhao Lv[1,*]

[1]Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University (AHU), 11 Jiulong Road, Hefei, 230601, China

Abstract

In this paper, we propose the multi-perspective information fusion (MPIF) Res2Net with random Specmix for fake speech detection (FSD). The main purpose of this system is to improve the model's ability to learn precise forgery information for FSD task in low-quality scenarios. The task of random Specmix, a data augmentation, is to improve the generalization ability of the model and enhance the model's ability to locate discriminative information. Specmix cuts and pastes the frequency dimension information of the spectrogram in the same batch of samples without introducing other data, which helps the model to locate the really useful information. At the same time, we randomly select samples for augmentation to reduce the impact of data augmentation directly changing all the data. Once the purpose of helping the model to locate information is achieved, it is also important to reduce unnecessary information. The role of MPIF-Res2Net is to reduce redundant interference information. Deceptive information from a single perspective is always similar, so the model learning this similar information will produce redundant spoofing clues and interfere with truly discriminative information. The proposed MPIF-Res2Net fuses information from different perspectives, making the information learned by the model more diverse, thereby reducing the redundancy caused by similar information and avoiding interference with the learning of discriminative information. The results on the ASVspoof 2021 LA dataset demonstrate the effectiveness of our proposed method, achieving EER and min-tDCF of 3.29% and 0.2557, respectively.

Keywords

multi-perspective information fusion, fake speech detection task, random Specmix strategy,

1. Introduction

Automatic speaker verification (ASV) [1] is a technology that verifies whether a person's voice matches their voiceprint model. ASV systems are currently vulnerable to three types of attack methods: audio replay, text-to-speech (TTS), and voice conversion (VC). In order to prevent these technologies from being abused and posing a threat to social security, researchers have noticed this issue and the biennial ASVspoof challenge is held to promote the development of countermeasures. The latest competition was held in 2021[5], and the first started in 2015 [2, 3, 4]. The first audio deepfake detection [6] challenge was successfully held in 2022.

Existing studies [7, 8, 9, 10, 32] aim to propose a system that can be universally applied to synthesis speech with the unknown attack types in the clean scenarios. Regarding the scenarios with poor quality, it introduces various interference to challenge the generalization of countermeasures, and the methods in the studies mentioned above will decrease their performance under this circumstance. For the speeches in the low-quality scenarios, researchers try to use data augmentation (DA) methods to improve the robustness of fake speech detection (FSD) task. For example, Tak et al. [11] proposed a data boosting method, Rawboost, on the raw audio for the reliable system. This technique improved performance of FSD task greatly. Park et al. [12] simply acted the masking method on log mel spectrogram, entailed the system to maintain robustness in the face of incomplete frequency information. In [13], they use frequency feature masking (FFM) to cover the information of spectrogram. Kim et al. [20] proposed the Specmix for the spectral correlation by applying time-frequency masks. However, these DA methods are always conducted for all samples, this may affect the original data distribution characteristics, resulting in performance degradation.

Efficient classification models have always been the research subject on FSD task. Light convolution neural network (LCNN) [14] can separate the noise signal and information signal, and is helpful for the feature selection. Alzantot et al. [15] built three variants based on the residual convolution network for performance improvement. Lai et al. [16] proposed the use of SE-Net to detect speech forgery, which assigns weights according to global attention dimensions. Res2Net [17] was proposed by Gao et al. to enhance its ability to capture multi-scale information by transferring information among channel groups.

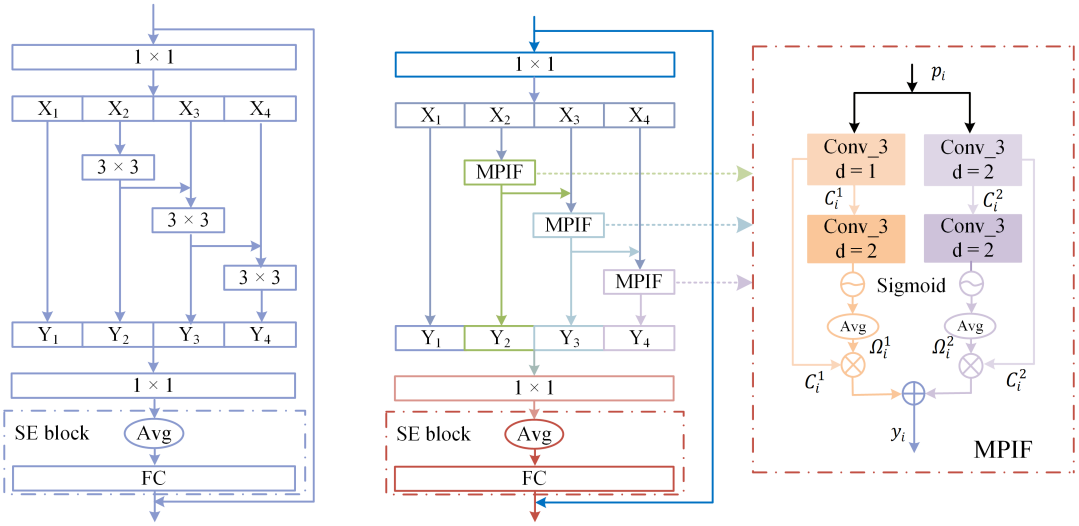
IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023), August 19, 2023, Macao, S.A.R

*Corresponding author.

†These authors contributed equally.

e22201030@stu.ahu.edu.cn (S. Dong); e21201068@stu.ahu.edu.cn (J. Xue); cunhang.fan@ahu.edu.cn (C. Fan); e22201061@stu.ahu.edu.cn (K. Zhu); e22201148@stu.ahu.edu.cn (Y. Chen); kjlz@ahu.edu.cn (Z. Lv)

0000-0002-0971-3272 (S. Dong)



(a) Res2Net backbone (b) MPIF-Res2Net

Figure 1: Illustration of Res2Net backbone(a) and the proposed MPIF-Res2Net(b). (SE Block: the squeeze-and-excitation block [31]; Avg is the AdaptiveAvgPool2d function; FC is full-connection layer; Sigmoid is a activate function; the MPIF-Res2Net we proposed is shown on the left of Figure 1(b); the MPIF module is shown on the right of Figure 1(b).)

This enables the model to acquire more comprehensive information of features. Li et al. [18] investigated the effectiveness of Res2Net in conjunction with different acoustic features. Li et al.[19] proposed a channel-wise gating mechanism to suppress channels with lower correlations which they thought not useful. However, the models mentioned above may not achieve better results in the low-quality scenarios as they only conducted their experiments in the celan scenarios.

In this work, we propose multi-perspective information fusion Res2Net (MPIF-Res2Net) with random Specmix. Spoofing information from a single perspective is always similar during learning process, which causes redundant information and blurs the truly discriminative information. The MPIF module fuses the information from different receptive field to reduce the redundant spoofing cues and enhance the robustness of system in the poor-quality scenarios. Specmix can increase the diversity of training data, thereby improving the generalization ability of the model. The generated spectrogram will incorporate information from other spectrograms, allowing the model to pay attention to the noteworthy information. And it performs cut and paste among spectrograms without introducing data that was not present in the original dataset with a modest impact on the original dataset. DA method conducted on all the samples may affect the distribution characteristics of the original data. For this issue, we randomly choose samples according to the probability p_hyper in advance to conduct Specmix to prevent excessive augmentation methods from weakening the fitting ability of system. Specifically, we randomly cover the part of the frequency information with corresponding frequency information of another sample in the same batch. This approach can improve the performance greatly. Our proposed method has been shown to be effective on the ASVspoof 2021 LA dataset, with achieved EER and min-tDCF results of 3.29% and 0.2557, respectively.

2. Methodology

2.1. Proposed method

In this section, we introduce the structure of the proposed MPIF-Res2Net, it reduces redundancy caused by learning single-perspective forgery information by integrating information from multiple perspectives. The convolutional operations with single kernel size are learning the similar forgery clues, producing too much redundant information and obscuring the important discriminative information. Therefore, the MPIF-Res2Net as shown on the left of Figure 1(b) is proposed to fuse information from different convolutional operations with different kernel size. The architecture of Res2Net is shown as Figure 1(a), the outcome from the 1×1 convolution was splited into n equal parts by the channel dimension, denoted as p_i, where i is the integer between 1 to n. And each part has p (Eq. 1) channels.

$$p = \frac{\#channel}{n} \tag{1}$$

where $\#channel$ means the total number of channels. Res2Net uses the residual-like connection to perform addition between the channel groups. The following formulation can be used to describe this process:

$$y_i = \begin{cases} p_i, & i = 1 \\ K_i\left(p_i\right), & i = 2 \\ K_i\left(p_i + y_{i-1}\right), & 2 < i \leq n \end{cases} \tag{2}$$

Where K_i represents the convolution operation. The proposed MPIF-Res2Net replaces the K_i operation with MPIF module as shown on the right of Figure 1(b). The y_i is calculated as fowllows:

$$y_i = \begin{cases} p_i, & i = 1 \\ MPIF_i\left(p_i\right), & i = 2 \\ MPIF_i\left(p_i + y_{i-1}\right), & 2 < i \leq n \end{cases} \tag{3}$$

2.2. Multi-perspective Information Fusion Module

As shown on the right of Figure 1(b), MPIF module in current channel group i performs different convolution operation on p_i or $p_i + y_{i-1}$ to get the spoofing information from different perspective. Firstly, p_i is sent into the convolution operations with different dilation parameter j, where $j \in [1, 2]$, at the beginning of MPIF module (Eq. 4). The results are then passing through the dilated convolution to recalculate the energy distribution of each channel, normalize them through the Sigmoid function, and then the average pooling layer is used to get the results ω_k^j as the weight of each channel k from $Conv2d$. And the purpose of using dilated convolution is to increase the receptive field, ensuring that each convolution output contains information from a larger range while keeping the parameter and computation cost constant. The weighting factor ω_k^j of each channel k is calculated by Eq. 5.

$$c^j = (Conv2d_j(p_i))) \tag{4}$$

$$\omega_k^j = \mathcal{A}vg(Sigmoid(Conv2d(c_k^j))) \tag{5}$$

$Conv2d_j$ at the beginning of MPIF module takes p_i as input and outputs c^j. c_k^j is the kth channel in $conv^j$. $Conv2d$ is a convolutional operation to recalculate energy distribution. $Sigmoid$ is the $Sigmoid$ function. Avg denotes the $AdaptiveAvgPool2d$ function.

After the weighting factor ω_k^j, we perform multiplication on c_k^j and ω_k^j, and sum up the results. Then we can get the result $MPIF_i(p_i)$ of i-th channel group as follows:

$$MPIF_i(p_i) = \sum_{j=1}^{2} C_i^j \times (\Omega_i^j) \tag{6}$$

where C_i^j is a matrix composed with c_k^j, $\Omega_i^j \in \mathbb{R}^{p \times 1 \times 1}$ is the weight matrix with ω_k^j. The p is the number of the channel of p_i.

Table 1
The Proposed MPIF-Res2net Model Architecture and Configuration. the Dimensions Are Arranged in the Order of Channels, Frequency, and Time). BN Denotes Batch Normalization and ReLU denotes Rectified Linear Unit, MPIF and SE Are the Multi-perspective Information Module and the Squeeze And Excitation Layer, Respectively.

Layer	Input:27000 samples	Output shape
Front-end	F0 subband	(45,600)(F,T)
Pre-processing	Channel expansion Conv2D_1 BN & ReLU	(1,45,600) (16,45,600)
Layer1 &Layer3	$1 \times \begin{cases} Conv2D\_1 \\ Conv2D\_3 \\ Conv2D\_1 \\ SE \end{cases}$	Layer1 (32,45,600) Layer3(128,12,150)
	$1 \times \begin{cases} Conv2D\_1 \\ MPIF \\ Conv2D\_1 \\ SE \end{cases}$	
Layer2 &Layer4	$1 \times \begin{cases} Conv2D\_1 \\ Conv2D\_3 \\ Conv2D\_1 \\ SE \end{cases}$	Layer2(64,23,300) Layer4(256,6,75)
	$2 \times \begin{cases} Conv2D\_1 \\ MPIF \\ Conv2D\_1 \\ SE \end{cases}$	
Output	Avgpool2D(1, 1) AngleLinear	(256,1,1) 2

2.3. Random Specmix Strategy

In this work, we use a random Specmix strategy to help the model to locate the discriminative information and enhance the generalization of the model. For the training of deep neural networks, we always transform the raw audio from time domain into time-frequency domain. And inspired by [20], we conduct Specmix on the frequency dimension of the F0 subband [30], which is a subband of amplitude spectrum, and the maximum span of Specmix operation is no more than 10. Specmix cuts and pastes spectrograms among themselves in the same batch to help the model focus on the discriminative regions that may be worth to attend to. And different from [20], there is no Specmix operation on labels. We cover the information on frequency dimension with the corresponding parts of other samples in the same batch. At the same time, to avoid the conduction of Specmix on all the samples, inspired by [21], we randomly choose speech samples according to the hyperparameter p_hyper in advance to conduct Specmix operation. For a batch of speech samples, the probability of them

Table 2
Results of Ablation Experiments for Our Proposal Module. (p_hyper) Means the Probability of Application of random Specmix strategy. Res2Net_k3 Denotes the Kernel Sizes of Convolution Are 3 In the channel groups; Res2Net_k5 Denotes the Kernel Sizes of Convolution Are 3 with the Dilation Parameter Is 2 In the channel groups.

Ablation Experiments Results On ASVspoof 2021 LA dataset						
(p_hyper)	MPIF-Res2Net		Res2Net_k3		Res2Net_k5	
	EER(%)	min-tDCF	EER(%)	min-tDCF	EER(%)	min-tDCF
0	4.04	0.2713	4.26	0.2750	4.00	0.2702
0.1	3.57	0.2577	4.16	0.2760	4.44	0.2806
0.2	3.96	0.2693	4.29	0.2739	4.08	0.2711
0.3	3.98	0.2692	4.08	0.2753	4.18	0.2762
0.4	3.70	0.2638	3.87	0.2688	4.08	0.2719
0.5	**3.29**	**0.2557**	4.23	0.2738	4.24	0.2756
0.6	3.65	0.2679	3.98	0.2708	3.54	0.2598
0.7	3.69	0.2737	4.25	0.2775	3.56	0.2660
0.8	3.73	0.2692	4.25	0.2804	3.58	0.2670
0.9	4.02	0.2696	4.23	0.2758	4.13	0.2731
1(no Specmix)	3.70	0.2707	4.58	0.2811	4.49	0.2859

undergoing random Specmix is p, when p is bigger than p_hyper, Specmix was conducted on them, otherwise no conduction with Specmix. And in the evaluation phase, we do not use the random Specmix strategy.

3. Expriments And Results

3.1. Experimental Setup

It must be a challenging task to learn a robust countermeasure suitable to low-quality scenario trained on the training set without same interference conditions. In this work, we use the Rawboost [11] DA method to train the model, this technique can enhance the accuracy in the low-quality scenarios. To be more precise, the impulsive signal-dependent (ISD) additive noise and stationary signal-independent (SSI) additive noise are added to the raw waveform. After the STFT operation with the window length is 1728 and the hop length is 130, we got a spectrogram of size 865. We then truncate or concatenate the spectrogram to fix the number of frames at 600. We utilize the 0-400 Hz LPS feature with the first 0-45 dimension as our F0 subband feature. The resulting feature size of F0 subband is 45×600. Then, we determine whether to conduct random Specmix on the samples in the current batch by setting the hyperparameter (p_hyper), the probability whether to conduct the random Specmix strategy. Considering that the F0 feature is a subband of the amplitude spectrum, we set the maximum span for Specmix to be no more than 10.

In this article, we propose MPIF-Res2Net to fuse the information from different perspective to reduce the redundant spoofing cues and introduce random Specmix to improve the generalization ability of the model. Table 1 presents the design of MPIF-Res2Net, which includes details on channels, convolution kernels, and repetition frequency. In our experiments, Adam is utilized as the optimizer, with the following parameter settings: $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$, and weight decay is 10^{-4}. The epoch is set to 32. And the number of channel groups is set to 8. The batch size is 16.

3.2. Dataset

The data in the ASVspoof 2019 logical access (LA) dataset is divided into three subsets: training set, development set, and evaluation set. The spoof speech in the training

and development sets comes from six speech synthesis and speech conversion technologies, which are known attack types. The evaluation set contains audio generated by 11 unknown attack types. We trained our model on the ASVspoof 2019 training set and selected the best performing model on the development set. As stated in [33], the ASVspoof 2021 LA dataset is designed for developing anti-spoofing methods that can effectively adapt to unknown channel variations and does not provide new matching training or development data. The speech samples from the ASVspoof 2021 evaluation set were transmitted via actual telephone systems utilizing various bandwidths and codecs. The data in the 2019 LA training and development subset does not have similar encoding and transmission, and these subsets only contain clean data. Equal error rate (EER) and minimum tandem detection cost function (min t-DCF) are used as the metrics.

3.3. Experimental Results

3.3.1. Ablation Study

Firstly, different values of the probability (p_hyper) should be considered as the guidance of the experimental conduction to obtain the best p_hyper. Table 2 shows the EER results of conduction with random Specmix strategy for different values of p_hyper. The MPIF-Res2Net with p_hyper=0.5 has the best performance whose EER result is 3.29%, and the min t-DCF result is 0.2557, which means a relatively higher reliability of the countermeasure system when it is applied with an ASV system. For experiments involving information from a single receptive field, we set up two models, Res2Net_k3 and Res2Net_k5, with the parameter kernel sizes and dilations are 3 and 1, 3 and 2, respectively. The EER result of MPIF-Res2Net with the p_hyper equal to 1 is 3.70%, however, the corresponding EER results of the other two systems are 4.58% and 4.49% respectively, which verifies the MPIF-Res2Net we proposed do have the ability by fusing information from different perspective to reduce the redundancy caused by learning the similar spoofing clues with the single kernel size. The EER results of Res2Net_k3 and Res2Net_k5 undergoing Specmix demonstrate that Specmix can help help the model to locate the forgery information and improve model generalization performance.

For random Specmix strategy, the MPIF-Res2Net with p_hyper is 0 got the EER result of 4.04%, the p_hyper with 0 means the Specmix conduction was conducted on all of the samples, this indicates that all the samples undergoing Specmix cause the serious performance degradation of system. Overall, the random Specmix has improved the model's generalization ability and enhanced its performance.

Table 3
Results Comparison with Fusion Systems on the Performance of ASVspoof2021 Dataset

System	t-DCF	EER(%)
T23 [22]	0.2177	1.32
T20 [23]	0.2608	3.21
T04 [24]	0.2747	5.58
T06 [25]	0.2853	5.66
T35 [22]	0.2480	2.77
T19 [22]	0.2495	3.13
Fusion systems [27]	0.2882	4.66
MPIF-Res2Net **ours**	**0.2557**	**3.29**

Table 4
Results Comparison with Single System on the Performance of ASVspoof2021 Dataset

System	t-DCF	EER(%)
B03 [28]	0.3445	9.26
B04 [28]	0.4257	9.5
B01 [28]	0.4974	15.62
B02 [28]	0.5758	19.3
RawNet2 [29]	0.3069	8.05
LFCC-LCNN [29]	0.3152	8.90
MPIF-Res2Net **ours**	**0.2557**	**3.29**

Experimental results show that our proposed MPIF-Res2Net with random Specmix enhancement methods can improve performance for FSD task in the low-quality scenarios.

3.3.2. Performance Comparison With Other Systems

The Table 3 shows the results on ASVspoof 2021 LA dataset of different fusion systems. Although the fusion systems T23 [22], T20 [23], T35 [22] and T19 [22] outperform than our proposed MPIF-Res2Net, but their fusion ways are very complicated. Such as the T23, it is composed by 12 other systems trained separately, and got fused with finely adjusted weight assignment at the score stage. The method we proposed is based on a single system, which is less complicated compared to the fusion systems. Table 4 shows the EER result of single systems, the best EER result of other systems is 8.05%, the method we proposed has improved the performance by 59% relative to the RawNet2[29]system.

4. Conclusion

In this paper, we achieve accurate and useful information discrimination from two aspects. On the one hand, Specmix helps the model to focus on the location of key information in the sample by mixing information between samples, and randomly selects samples

for Specmix operations, effectively avoiding the phenomenon of performance degradation caused by the destruction of original data. On the other hand, MPIF-Res2Net reduces redundant information caused by learning similar information from a single perspective by fusing information from multiple perspectives, removing the influence of redundant information on the learning of key information. The effectiveness of our method has been demonstrated by experiments. The effectiveness of our proposed method was verified by the experiment results.

References

[1] Naika R. An overview of automatic speaker verification system[C]//Intelligent Computing and Information and Communication: Proceedings of 2nd International Conference, ICICC 2017. Springer Singapore, 2018: 603-610.

[2] Wu Z, Kinnunen T, Evans N, et al. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge[C]//Sixteenth annual conference of the international speech communication association. 2015.

[3] Kinnunen T, Sahidullah M, Delgado H, et al. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection[J]. 2017.

[4] Todisco M, Wang X, Vestman V, et al. ASVspoof 2019: Future horizons in spoofed and fake audio detection[J]. arXiv preprint arXiv:1904.05441, 2019.

[5] Yamagishi J, Wang X, Todisco M, et al. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection[J]. arXiv preprint arXiv:2109.00537, 2021.

[6] 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. IEEE, 2022, pp. 9216– 9220.

[7] Arif T, Javed A, Alhameed M, et al. Voice spoofing countermeasure for logical access attacks detection[J]. IEEE Access, 2021, 9: 162857-162868.

[8] Zhang Y, Jiang F, Duan Z. One-class learning towards synthetic voice spoofing detection[J]. IEEE Signal Processing Letters, 2021, 28: 937-941.

[9] Das R K, Yang J, Li H. Long range acoustic and deep features perspective on ASVspoof 2019[C]//2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019: 1018-1025.

[10] Nautsch A, Wang X, Evans N, et al. ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2021, 3(2): 252-265.

[11] Tak H, Kamble M, Patino J, et al. Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification antispoofing[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 6382-6386.

[12] Park D S, Chan W, Zhang Y, et al. Specaugment: A simple data augmentation method for automatic speech recognition[J]. arXiv preprint arXiv:1904.08779, 2019.

[13] Kwak I Y, Choi S, Yang J, et al. CAU_KU team's submission to ADD 2022 Challenge task 1: Low-quality fake audio detection through frequency feature masking[J]. arXiv preprint arXiv:2202.04328, 2022.

[14] Lavrentyeva G, Novoselov S, Malykh E, et a1. Audio replay attack detection with deep learning frameworks [C] // Proc of Interspeech 2017. Grenoble, France: ISCA, 2017: 82-86

[15] Alzantot M, Wang Z, Srivastava M B. Deep residual neural networks for audio spoofing detection[J]. arXiv preprint arXiv:1907.00501, 2019.

[16] Lai C I, Chen N, Villalba J, et al. ASSERT: Anti-spoofing with squeeze-excitation and residual networks[J]. arXiv preprint arXiv:1904.01120, 2019.

[17] Gao S H, Cheng M M, Zhao K, et al. Res2net: A new multi-scale backbone architecture[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 43(2): 652-662.

[18] Li X, Li N, Weng C, et al. Replay and synthetic speech detection with res2net architecture[C]//ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2021: 6354-6358.

[19] X. Li, X. Wu, H. Lu, X. Liu, and H. Meng, "Channel-wise gated res2net:Towards robust detection of synthetic speech attacks," Proc. Interspeech 2021, 2021.

[20] Kim G, Han D K, Ko H. Specmix: A mixed sample data augmentation method for training withtime-frequency domain features[J]. arXiv preprint arXiv:2108.03020, 2021.

[21] Zhong Z, Zheng L, Kang G, et al. Random erasing data augmentation[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 13001-13008.

[22] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev,and G. Lavrentyeva, "STC Antispoofing Systems for the ASVspoof2021 Challenge," in Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 61–67.

[23] T. Chen, E. Khoury, K. Phatak, and G. Sivaraman, "Pindrop Labs' Submission to the ASVspoof 2021 Challenge," in Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 89–93.

[24] J. C´aceres, R. Font, T. Grau, and J. Molina, "The Biometric Vox System for the ASVspoof 2021 Challenge," in Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 68–74.

[25] W. H. Kang, J. Alam, and A. Fathan, "CRIM's System Description for the ASVSpoof2021 Challenge," in Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021 pp. 100–106.

[26] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kin- nunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch et al., "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," arXiv preprint arXiv:2210.02437, 2022.

[27] A. Cohen, I. Rimon, E. Aflalo, and H. H. Permuter, "A study on data augmentation in voice anti-spoofing," Speech Communication, vol. 141, pp. 56–67, 2022.

[28] 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 Spoof- ing Coutermeasures Challenge, 2021.

[29] X. Wang, X. Qin, T. Zhu, C. Wang, S. Zhang, and M. Li, "The dku-cmri system for the asvspoof 2021 challenge: vocoder based replay channel response estimation," Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, pp. 16–21, 2021.

[30] Xue J, Fan C, Lv Z, et al. Audio Deepfake Detection Based on a Combination of F0 Information and Real Plus Imaginary Spectrogram Features[C]//Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia. 2022: 19-26.

[31] 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.

[32] Xue J, Fan C, Yi J, et al. Learning from yourself: A self-distillation method for fake speech detection[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[33] Liu X, Wang X, Sahidullah M, et al. ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild[J]. arXiv preprint arXiv:2210.02437, 2022.