INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

TO-Rawnet: Improving RawNet with TCN and Orthogonal Regularization for Fake Audio Detection

Chenglong Wang[1,2], Jiangyan Yi[2,3], Jianhua Tao[3,4], Chu Yuan Zhang[2,3], Shuai Zhang[4], Ruibo Fu[2], Xun Chen[1]

[1]University of Science and Technology of China, Hefei, China
[2]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijng, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[4]Department of Automation, Tsinghua University

chenglong.wang@nlpr.ia.ac.cn

Abstract

Current fake audio detection relies on hand-crafted features, which lose information during extraction. To overcome this, recent studies use direct feature extraction from raw audio signals. For example, RawNet is one of the representative works in end-to-end fake audio detection. However, existing work on RawNet does not optimize the parameters of the Sinc-conv during training, which limited its performance. In this paper, we propose to incorporate orthogonal convolution into RawNet, which reduces the correlation between filters when optimizing the parameters of Sinc-conv, thus improving discriminability. Additionally, we introduce temporal convolutional networks (TCN) to capture long-term dependencies in speech signals. Experiments on the ASVspoof 2019 show that the Our TO-RawNet system can relatively reduce EER by 66.09% on logical access scenario compared with the RawNet, demonstrating its effectiveness in detecting fake audio attacks.

Index Terms: ASVspoof, fake audio detection, end-to-end, orthogonal convolution

1. Introduction

In the field of fake audio detection, the use of standard hand-crafted feature is common [1, 2, 3, 4]. Linear frequency cepstrum coefficients (LFCC) are a benchmark feature for anti-spoofing tasks [5, 6]. LFCC uses linear filters instead of Mel filters, emphasizing high-frequency features over Mel frequency cepstral coefficients (MFCC). Constant Q cepstral coefficients (CQCC) are derived from a constant-Q transform (CQT) [7, 8], capturing frequency domain features effectively. Additional features, such as group delay gram [9], log power spectrum (LPS) [10], and cochlear filter cepstral coefficients instantaneous frequency (CFCCIF) [11], also perform well. However, standard features may smooth the speech spectrum, hindering the extraction of vital narrow-band speaker traits like pitch [12] and formants. In contrast, direct processing of raw waveforms enables the network to learn task-specific low-level embeddings.

Recently, researchers have shown increased interest in utilizing raw waveforms as system inputs [13, 14, 15, 16]. For example, Dinkel [17] proposes a raw waveform convolutional long short-term neural network (CLDNN) to enhance defense against unknown attacks. Another study [18] improves end-to-end fake audio detection by constructing ResWavegram from one-dimensional convolutions applied to the raw waveform. Moreover, SincNet [14] is a novel neural network architecture that enhances feature extraction using adjustable cutoff frequency parameters in bandpass filters based on Sinc functions, outperforming fixed-parameter Mel filters. Additionally, models like RawNet2 [15] and AASIST [19] encode raw waveforms using Sinc functions. However, the flexibility of parameter settings may sometimes result in suboptimal local minima, and prior studies have not iteratively optimized Sinc-conv parameters [15, 19, 20].

Studies in image processing have demonstrated that the utilization of orthogonal learned filters can enhance model capacity, leading to improved feature expression and intra-class feature representation [21, 22, 23]. Building upon the work of [24], we enhance RawNet2 [25] by incorporating an orthogonal constraint. The convolution kernels are initialized as linear-scale Sinc filters using band-pass filtered Sinc functions. By enforcing the orthogonality property on the convolution kernels, we reduce correlation between filters. Additionally, convolutional neural networks (CNNs) face challenges in capturing long-term dependencies due to the constraint of kernel size. Drawing inspiration from [26], we employ a temporal convolution network (TCN) instead of Conv1d to expand the receptive field of the convolutional kernel. Our proposed method improves the discriminative power of RawNet2 by extracting more robust features from raw audio signals and capturing the complex temporal dynamics of speech signals. We evaluate the effectiveness of our approach on two benchmark datasets, demonstrating its superiority over other state-of-the-art systems for fake audio detection. Our contributions can be summarized as follows:

- We proposed a novel deep neural network architecture called TO-RawNet for fake audio detection. The model combines the advantages of orthogonal convolution and TCN to improve upon RawNet2. To our best knowledge, this is the first application of the combination of orthogonal convolution and TCN in the field of fake audio detection.
- Compared to RawNet, experiments conducted on the ASVspoof 2019 dataset demonstrate that our TO-RawNet system can significantly reduce EER by 66.09% in the logical access scenario.

The structure of this paper is as follows: Section 2 provides an overview of related work. Section 3 details our proposed method. Experiments, results and discussions are reported in Section 4 and 5, respectively. Finally, we conclude the paper in Section 6.

2. Related Work

Recently, more and more researchers have been using raw waveform inputs directly in the field of fake audio detection.

 10.21437/Interspeech.2023-1206

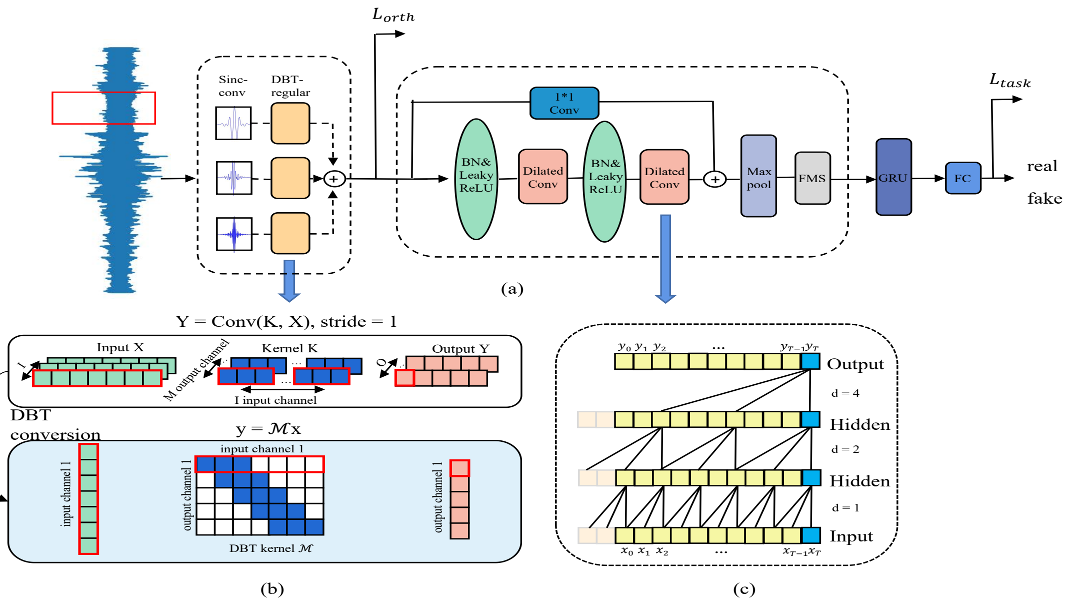

Figure 1: *(a) Overall framework of the TO-RawNet based fake audio detection model. (b) The convolution expression; $Conv(K, X)$ is converted into a faster DBT vector representation; $y = \mathcal{M}x$. (c) dilated convolution module.*

SincNet [14] is a neural network architecture that is designed to operate directly on the raw waveform of audio signals. The first layer of SincNet consists of a bank of band-pass filters that are parametrized as Sinc functions, allowing for the extraction of useful features directly from the raw waveform. By using a constrained first layer with fewer learnable parameters, SincNet is able to learn a more meaningful filterbank structure, resulting in more meaningful output. RawNet2 [25], another neural network architecture, also employs a bank of band-pass filters parametrized as Sinc functions to extract features from the raw waveform. The upper layers of RawNet2 consist of residual blocks and gated recurrent units (GRUs) [27], with the addition of filter-wise feature map scaling (FMS). By applying a sigmoid function to the residual block outputs, FMS serves as an attention mechanism to obtain more distinct representations, resulting in improved discriminative power. Studies have shown that using end-to-end architectures based on learned features rather than knowledge-based and hand-crafted features has the potential to improve the performance of fake audio detection.

3. Proposed Methods

3.1. Orthogonal Convolution

This paper is based on the differentiable frontend of Sinc-conv, and aims to improve feature expressiveness and intra-class feature representation by using orthogonal convolutions and regularization constraints. The specific operational steps are as follows: as shown in Figure 1 (b), we view the convolution operation as a matrix-vector multiplication, where the kernel matrix $\mathcal{M}$ is generated by the convolution kernel K. Using the linear property of the convolution operation, we adopt the Doubly Block-Toeplitz (DBT) matrix construction method to transform the convolution expression $Conv(K, X)$ into a faster DBT matrix-vector representation, as shown below:

$$Y = Conv(K, X) \Leftrightarrow y = \mathcal{M}x \tag{1}$$

where $\mathcal{M}$ is the DBT matrix, and x and y represent the input and output tensors, respectively. The shape of the DBT matrix $\mathcal{M}$ is $K \in R^{(OT_2)\times(IT_1)}$, Where O and I are the output and input channels, and T_2 and T_1 are the feature map lengths of the output and input. And its rows need to be orthogonalized to reduce the correlation between filters. The orthogonality condition of the DBT matrix $\mathcal{M}$ is shown in equation (2):

$$\left\langle \mathcal{M}_{it_1'}, \mathcal{M}_{jt_2'} \right\rangle = \begin{cases} 1 & (i, t_1') = (j, t_2') \\ 0 & \text{else} \end{cases} \tag{2}$$

where t_1' and t_2' represent two different filter positions, and i and j represent the corresponding row positions of these filters in the matrix. However, since $\mathcal{M}$ is highly structured and sparse, a more efficient method for orthogonal calculation was proposed in [24], as shown in equation (3):

$$Y = Conv(K, K, padding = P, stride = S) = I_{r0} \tag{3}$$

where K represents the size of the convolution kernel, S represents the stride, and $P = \lfloor \frac{K-1}{S} \rfloor \cdot S$ represents padding. I_{r0} is a tensor, where the center is an $n \times n$ identity matrix, and the rest is padded with zeros. By minimizing the difference between $Z = Conv(K, K, padding = P, stride = S)$ and I_{r0}, a roughly orthogonal convolution can be obtained. The loss function of the orthogonal convolution can be expressed as:

$$min_K L_{orth} = \|Z - I_{r0}\|_F^2 \tag{4}$$

The final training loss is as follows:

$$L = L_{task} + \lambda L_{orth} \tag{5}$$

Where L_{task} represents the loss of the classification task, and λ is the weight of the orthogonal regularization loss, and we set three different values in the experiment. Please refer to Algorithm 1 for the specific pseudocode implementation, where o_c represents output channels and i_c represents input channels.

Algorithm 1 Detailed Procedure of Orthogonal Convolution.

```
function deconv_orth_dist(kernel, stride, padding):
    [o_c, i_c, kernel_size] = kernel.shape
    output = conv_1d(kernel, kernel, stride, padding)
    target = zeros((o_c, o_c, output.shape[-1]))
    center = floor_divide(output.shape[-1], 2)
    target[:, :, center] = eye(o_c)
    return norm(subtract(output, target))
```

3.2. Temporal Convolution Network

Inspired by TCN [26], we propose the dilated convolution block to extract features, as shown in Figure 1 (c). The residual block first employs batch normalization and the leaky ReLU activation function, followed by dilated convolution. Next, a 1x1 convolution is used to adjust the output channels to match the input channels. To expedite convergence and facilitate the training of deeper models, we incorporate a residual [28] path. Each block's output serves as the input for the subsequent block. The dilation factor is doubled for each block up to a certain limit and then repeated ($e.g.$, $1, 2, 4, ..., 2^n$). This exponential increase in dilation factor ensures that the model captures sufficient temporal contextual information for detecting fake audio. It enlarges the network's receptive field and captures forgery traces in the entire speech with fewer stacked layers.

3.3. TO-Rawnet

Figure 1 (a) illustrates the architecture of our proposed TO-Rawnet system. First, the raw waveform is fed into the Sinc-conv layer with orthogonal regularization to produce a high-level speech representation. The orthogonal regularization helps reduce redundancy by enabling each filter to focus on distinct frequency components. Next, the high-level representations are fed into the residual module, which includes dilation convolutions with exponentially increasing receptive fields. This enables the network to effectively increase its perception range, allowing it to capture more global information from the input audio. Subsequently, we connect a GRU to extract an utterance-level representation, which is then fed into a softmax activation function to perform real/fake classification.

4. Experiments

4.1. Dataset

4.1.1. ASVspoof 2019 Challenge Dataset

ASVspoof 2019 LA [5] mainly has 19 spoofing attack algorithms (A01-A19), with two types of spoofing attacks: text to speech (TTS) and voice conversion (VC). The LA data set contains three subsets: the training set, the development set, and the evaluation set. Table 1 details the number of real and fake audio of the ASVspoof2019 LA dataset. The attack algorithms in the training and development sets overlap, while the evaluation set includes unseen spoofing attacks.

4.1.2. ASVspoof 2021 Challenge Dataset

ASVspoof 2021 LA [6] poses greater challenges than the previous versions. Although the training and development sets remain the same as those of ASVspoof 2019 LA database, the evaluation set is distinct. Specifically, the evaluation data for 2021 LA contains encoding and transmission artifacts that stem from actual telephony systems.

Table 1: *The detailed information of ASVspoof2019 LA dataset and ASVspoof2021 LA dataset.*

Set	**Genuine**	**Spoofed**	**Total**
	# utterance	# utterance	# utterance
Train	2,580	22,800	25,380
Dev	2,548	22,296	24,844
Eval(2019 LA)	7,355	64,578	71,933
Eval(2021 LA)	18,452	163,114	181,566

4.2. Experimental Setup

The audio sampling rate is 16k. To form batches, we standardized the duration of the raw waveform input to approximately 4 seconds (64600 samples) by either truncating longer utterances or concatenating shorter ones. The Sinc-conv layers have a filter length of l = 129, a stride of d = 1, and utilize n = 128 filters. We used fixed linear-scale Sinc filters. To enhance the performance of our model, we utilized six residual-blocks architecture that consists of 12 dilated convolution blocks with varying dilation factors, where the highest dilation factor is 32. In order to prevent over-fitting and under-fitting during the training process, we experimented with different channel combination configurations to determine the optimal combination. The number of channels in the first two residual blocks and the last four residual blocks are set to (32, 64), (128, 256), and (256, 512), respectively. We named them small (S), medium (M), and large (L), in that order. To further improve the discriminative power of our model, we employed FMS independently for each residual-block output. This technique enhances the most informative filter outputs and improves the overall accuracy of the model. To aggregate frame-level representations into an utterance-level representation, we utilized a GRU layer with 1024 hidden nodes. The output of the GRU layer is passed through a softmax activation function, which produces two-class predictions, i.e., real or fake. We propose an orthogonal regularization loss in this paper and set three different weights λ for this loss: 0.05, 0.1, and 0.2.

To train the model, we use the Adam optimizer with a learning rate of $5 * 10^{-5}$. We set the batch size to 32. The model is trained for 150 epochs. The training set is used to train the model, the development set is used to select the model with the best performance, and finally, the evaluation set is used for evaluation. The results of the ASVspoof2021 competition suggest that data augmentation can reduce overfitting and improve generalization [6, 29, 30]. To this end, in our experiments on ASVspoof2021, we utilized data augmentation techniques. Specifically, we employed the open-source tool RawBoost[1] for performing data augmentation in the LA task. We added linear and nonlinear convolutional noise and impulsive signal-dependent additive noise in the LA database.

In this work, in order to evaluate the results of different fake audio detection systems, the equal error rate (EER) [31] is used as the evaluation metric.

5. Results and Discussion

5.1. Ablation Experiments

Table 2 reveals that the hyperparameters perform better when λ is set to 0.1. As a result, for subsequent experiments, we have kept λ fixed at 0.1. The Orth-RawNet-M based model consis-

[1] https://github.com/TakHemlata/RawBoost-antispoofing

Table 2: *Orth-RawNet based models tested on the ASVspoof2019 LA evaluation set. Orth-RawNet-S, Orth-RawNet-M and Orth-RawNet-L with different values of λ. Results are the average (best) obtained from three runs of each experiment with different random seeds.*

Methods	λ	EER
Orth-RawNet-S	0.05	4.39 (4.15)
	0.1	3.82 (3.63)
	0.2	4.51 (4.43)
Orth-RawNet-M	0.05	3.86 (3.57)
	0.1	**3.19 (3.06)**
	0.2	3.52 (3.36)
Orth-RawNet-L	0.05	3.78 (3.65)
	0.1	3.66 (3.59)
	0.2	3.94 (3.73)

tently outperforms the other models when the same λ is used, indicating that the channel combination of (128, 256) prevents over-fitting or under-fitting issues.

Table 3 shows results for ablation experiments for which one of the components in the TO-RawNet model is removed. The results indicate that both orthogonal regularization and TCN have positive effects on the processing of speech signals. Experimental comparisons demonstrate that the TO-RawNet model, which combines these two techniques, performs the best. This is because orthogonal regularization can reduce the correlation between filters, thereby improving the model's generalization ability, while TCN can capture the long-term dependencies in speech signals. Furthermore, all of the models we tested outperformed the baseline system RawNet.

Based on the third column of Table 3, we can draw the conclusion that while the improvement in performance on ASVspoof2021 is less significant than that on ASVspoof2019, the model's performance has still been enhanced through the combination of orthogonal regularization and TCN.

5.2. Compared with Other Systems

Compared with the two baseline systems (CQCC-GMM and LFCC-GMM), our method has shown a significant improvement, with EER decreasing from 9.57% and 8.09% to 1.58%, respectively. Compared with neural networks using traditional feature extraction methods [25, 32, 33], the proposed model has better performance. Compared with RawNet2, which also uses raw waveform as input, the proposed method in this paper achieved a relative improvement of 66.09%. Considering that our method is an improvement upon RawNet2, this demonstrates the effectiveness of our approach. However, compared with the current state-of-the-art (SOTA) single-system AASIST model, the performance of our TO-RawNet model is slightly inferior. The AASIST model also uses raw waveform as input and employs a well-designed graph neural network after Sinc-conv encoding. To verify the effectiveness of our method, we applied orthogonal regularization in the Sinc-conv stage of the AASIST model (since TCN cannot be added to graph neural networks). The results showed that the performance of Orth-AASIST, after orthogonal regularization, was further improved compared to the original AASIST model, with EER decreasing from 1.13% to 1.02%. This demonstrates the universal applicability of the orthogonal regularization proposed in this paper.

Table 3: *The ablation experiments on the ASVspoof2019 LA and ASVspoof2021 LA evaluation sets are represented by EER1 and EER2, respectively. Results are the average (best) obtained from three runs of each experiment with different random seeds.*

Methods	EER1	EER2
RawNet	4.66	5.31
Orth-RawNet-S	3.82 (3.63)	5.02 (4.86)
Orth-RawNet-M	3.19 (3.06)	4.81 (4.73)
Orth-RawNet-L	3.66 (3.59)	4.62 (4.55)
TCN-RawNet-S	3.43 (3.37)	5.26 (5.13)
TCN-RawNet-M	2.86 (2.62)	5.08 (4.87)
TCN-RawNet-L	3.25 (3.14)	5.12 (4.96)
TO-RawNet-S	1.97 (1.86)	4.05 (3.84)
TO-RawNet-M	**1.58 (1.23)**	**3.70 (3.58)**
TO-RawNet-L	2.56 (2.37)	3.93 (3.78)

Table 4: *Performance comparison of the proposed methods to some known single systems on the ASVspoof2019 LA evaluation set.*

Methods	Front-end	EER
CQCC-GMM (Baseline1) [34]	CQCC	9.57
LFCC-GMM (Baseline2) [34]	LFCC	8.09
S1-RawNet2 [25]	Raw waveform	5.64
S2-RawNet2 [25]	Raw waveform	5.13
S3-RawNet2 [25]	Raw waveform	4.66
Resnet18-OC-softmax [32]	LFCC	2.19
MCG-Res2Net50 [33]	CQT	1.78
AASIST [19]	Raw waveform	1.13
TO-RawNet (ours)	Raw waveform	**1.58**
Orth-AASIST (ours)	Raw waveform	**1.02**

6. Conclusions

We propose a new end-to-end fake speech detection system named TO-RawNet, which has two new contributions: (i) using orthogonal regularization to constrain the learning process of filters, thereby improving the ability of feature expression and intra-class feature representation; (ii) introducing TCN to capture long-term dependencies in time-series data. Compared to RawNet, our TO-RawNet system reduces the EER by 66.09% in logical access scenarios. Furthermore, we apply the orthogonal regularization technique to the SOTA single-system AASIST and observe performance improvement, verifying the generalizability of orthogonal regularization. In the future, we will verify the performance of TO-RawNet across datasets and further improve its performance on backend models.

7. Acknowledgments

This work is supported by the National Key Research and Development Plan of China (No.2020AAA0140003), the National Natural Science Foundation of China (NSFC) (No.61831022, No.U21B2010, No.62101553, No.61971419, No.62006223, No.62276259, No.62201572, No. 62206278), Beijing Municipal Science&Technology Commission,Administrative Commission of Zhongguancun Science Park No.Z211100004821013, Open Research Projects of Zhejiang Lab (NO. 2021KH0AB06).

8. References

[1] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Galka, "Audio replay attack detection using high-frequency features." in *Interspeech*, 2017, pp. 27–31.

[2] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection-results on the asvspoof 2017 challenge." in *Interspeech*, 2017, pp. 7–11.

[3] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin, "Stc anti-spoofing systems for the asvspoof 2015 challenge," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5475–5479.

[4] P. Korshunov and S. Marcel, "Cross-database evaluation of audio-based spoofing detection systems," Tech. Rep., 2016.

[5] 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*, 2019, pp. 1008–1012.

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

[7] J. C. Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[8] 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.

[9] F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention." in *Interspeech*, 2018, pp. 681–685.

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

[11] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Sixteenth annual conference of the international speech communication association*, 2015.

[12] X. Xiao, X. Tian, S. Du, H. Xu, E. Chng, and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: the ntu approach for asvspoof 2015 challenge." in *Interspeech*, 2015, pp. 2052–2056.

[13] J.-w. Jung, H.-S. Heo, J.-h. Kim, H.-j. Shim, and H.-J. Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *Proc. Interspeech 2019*, pp. 1268–1272, 2019.

[14] M. Ravanelli and Y. Bengio, "Interpretable convolutional filters with sincnet," *arXiv preprint arXiv:1811.09725*, 2018.

[15] J.-w. Jung, S.-b. Kim, H.-j. Shim, J.-h. Kim, and H.-J. Yu, "Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms," *Proc. Interspeech 2020*, pp. 1496–1500, 2020.

[16] J.-w. Jung, Y. J. Kim, H.-S. Heo, B.-J. Lee, Y. Kwon, and J. S. Chung, "Pushing the limits of raw waveform speaker recognition," *arXiv preprint*, vol. 2203, 2022.

[17] H. Dinkel, N. Chen, Y. Qian, and K. Yu, "End-to-end spoofing detection with raw waveform cldnns," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4860–4864.

[18] Y. Ma, Z. Ren, and S. Xu, "RW-Resnet: A Novel Speech Anti-Spoofing Model Using Raw Waveform," in *Proc. Interspeech 2021*, 2021, pp. 4144–4148.

[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 *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6367–6371.

[20] H. Zeinali, T. Stafylakis, G. Athanasopoulou, J. Rohdin, I. Gkinis, L. Burget, J. Černockỳ *et al.*, "Detecting spoofing attacks using vgg and sincnet: but-omilia submission to asvspoof 2019 challenge," *arXiv preprint arXiv:1907.12908*, 2019.

[21] P. Aghdaie, B. Chaudhary, S. Soleymani, J. Dawson, and N. M. Nasrabadi, "Attention aware wavelet-based detection of morphed face images," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2021, pp. 1–8.

[22] A. Araujo, B. Negrevergne, Y. Chevaleyre, and J. Atif, "On lipschitz regularization of convolutional layers using toeplitz matrix theory," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 6661–6669.

[23] A. Thomas, A. Gu, T. Dao, A. Rudra, and C. Ré, "Learning compressed transforms with low displacement rank," *Advances in neural information processing systems*, vol. 31, 2018.

[24] J. Wang, Y. Chen, R. Chakraborty, and S. X. Yu, "Orthogonal convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 505–11 515.

[25] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6369–6373.

[26] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[27] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] 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.

[30] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6382–6386.

[31] J.-M. Cheng and H.-C. Wang, "A method of estimating the equal error rate for automatic speaker verification," in *2004 International Symposium on Chinese Spoken Language Processing*. IEEE, 2004, pp. 285–288.

[32] 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.

[33] X. Li, X. Wu, H. Lu, X. Liu, and H. Meng, "Channel-wise gated res2net: Towards robust detection of synthetic speech attacks," *arXiv preprint arXiv:2107.08803*, 2021.

[34] 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.