ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096278

LEVERAGING POSITIONAL-RELATED LOCAL-GLOBAL DEPENDENCY FOR SYNTHETIC SPEECH DETECTION

Xiaohui Liu[1]*, Meng Liu*[1]*, Longbiao Wang*[1,*] *, Kong Aik Lee*[2,*] *, Hanyi Zhang*[1]*, Jianwu Dang*[1]

[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Institute for Infocomm Research, A*STAR, Singapore

ABSTRACT

Automatic speaker verification (ASV) systems are vulnerable to spoofing attacks. As synthetic speech exhibits local and global artifacts compared to natural speech, incorporating local-global dependency would lead to better anti-spoofing performance. To this end, we propose the Rawformer that leverages positional-related local-global dependency for synthetic speech detection. The two-dimensional convolution and Transformer are used in our method to capture local and global dependency, respectively. Specifically, we design a novel *positional aggregator* that integrates local-global dependency by adding positional information and flattening strategy with less information loss. Furthermore, we propose the squeeze-and-excitation Rawformer (SE-Rawformer), which introduces squeeze-and-excitation operation to acquire local dependency better. The results demonstrate that our proposed SE-Rawformer leads to 37% relative improvement compared to the single state-of-the-art system on ASVspoof 2019 LA and generalizes well on ASVspoof 2021 LA. Especially, using the *positional aggregator* in the SE-Rawformer brings a 43% improvement on average.

Index Terms— anti-spoofing, local-global, transformer, positional encoding

1. INTRODUCTION

Advanced text-to-speech synthesis (TTS) and voice conversion (VC) technologies capable of generating realistic fake human voices have posed a significant threat to automatic speaker verification (ASV) systems [1–3]. Given the vulnerability of ASV to spoofing attacks, various studies and initiatives have been devoted to the defection of synthetic speech. Among others, the ASVspoof challenge series [1–4] has attracted a lot of attention recently with the provision of common evaluation rules, performance metrics, and datasets. The datasets include both bonafide and synthetic speech with a wide range of state-of-the-art TTS and VC technologies.

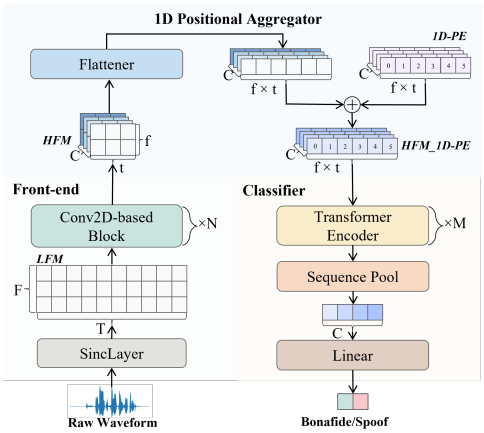
The discriminative information of synthesized and converted speech exists locally (e.g., unnatural stress and intonation) and globally (e.g., excessive smoothing). Integrating the local-global dependency would therefore promote a better anti-spoofing performance. Commonly used models for anti-spoofing countermeasure (CM) include *convolutional neural network* (CNN) [5–13] and *graph attention network* (GAT) [14–17]. Due to the locality of the convolutional and pooling layers, the CNN-based systems are suitable for obtaining local dependency and achieving relatively satisfying results, while they are inefficient in obtaining global dependency by stacking multiple convolutional layers. GAT-based systems introduced in [14–17] further improve the performance of anti-spoofing by leveraging graph attention [18] to capture discriminative cues spanning across frequency bins and temporal locations. Specifically, GAT captures local-global dependency by modeling non-Euclidean relationships between graph nodes, where each graph node aggregates local dependency in the feature map. However, in order to reduce the amount of computation, the constructed graph nodes are limited, which leads to a loss of information. Even in the best GAT-based system AASIST proposed in [16], a part of the spectro-temporal information is lost in pooling feature maps along the frequency and temporal dimensions, respectively, to extract a temporal graph and a spectral graph, which affects the acquisition of global dependency.

Compared with sequence model like *long short-term memory* (LSTM) [19], the use of Transformer [20] is less common in anti-spoofing. It captures global dependency efficiently through the self-attention mechanism, which is crucial for anti-spoofing [21]. To this end, we aim to improve the AASIST by using the Transformer instead of the GAT to capture local-global dependency more efficiently. Since the use of Transformers breaks through the computational constraints, we can directly reshape the feature map along the time and frequency dimensions into a longer sequence, which reduces the loss of time-frequency information. Due to the temporal nature of speech, we hypothesize that adding positional encoding to the sequence benefits for preserving spectro-temporal positional information. Thus, We propose a countermeasure for synthetic speech detection leveraging positional-related local-global dependency via Transformer with positional encoding, denoted as Rawformer. The Rawformer uses a RawNet2 [13] front-end to obtain a higher-level feature map (HFM) for modeling local dependency. Then, global dependency is required by our proposed *positional aggregator* and Transformer-based classifier. We design the *positional aggregator* to add positional information to the HFM and re-construct it in a two-dimensional sequence suitable for Transformer. In this regard, we investigate the appropriate feature map configurations in conjunction with 1D and 2D positional encoding to integrate the local-global dependency. Also, We explore the use of squeeze-and-excitation operation to model local dependency by strengthening the connections between channels.

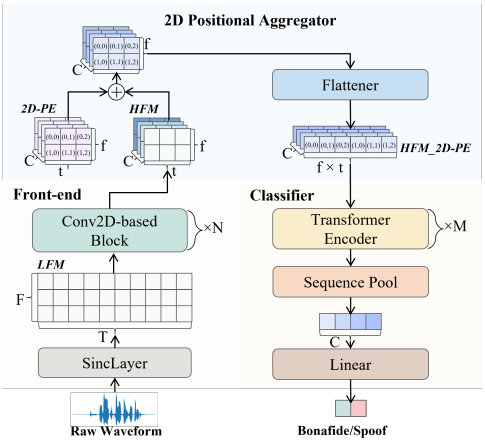
2. BASELINE SYSTEM

In [16], a RawNet2 front-end is used to learn higher-level feature maps (HFM) $S_{HFM} \in \mathbb{R}^{C \times f \times t}$, where C, f, and t stand for the numbers of channels, frequency and temporal bins, respectively. The feature maps are pooled (i.e., *max* pooling) along the frequency dimension, and the rescaling features of $\mathbb{R}^{C \times t}$ is modeled as a temporal graph $\mathcal{G}_t$. The feature maps are pooled along temporal dimension to construct a spectral graph $\mathcal{G}_s$. Let N_t be the the number of

*Corresponding author



(a) Rawformer with 1D positional encoding.



(b) Rawformer with 2D positional encoding.

Fig. 1. Illustration of the overall structure of our proposed Rawformer.

nodes in the temporal graphs $\mathcal{G}_t$, and N_s be the number of nodes in the spectral graphs $\mathcal{G}_s$. To further model the spectro-temporal relationship, $\mathcal{G}_t$ and $\mathcal{G}_s$ are concatenated to obtain a combined heterogeneous graph, which contains $N_t + N_s$ nodes. The combined graph is processed with a heterogeneous attention network (MGO) leading to the final graph $\mathcal{O}$. The graph is read out through the readout module using the max and $mean$ operations together with a stack node. By using element-wise max and $mean$ read operations, the AASIST can handle a variable number of nodes in both temporal and spectral axes. Therefore it could take the variable-length speech as an input.

3. POSITIONAL-RELATED LOCAL-GLOBAL DEPENDENCY FOR ANTI-SPOOFING

To mitigate the information loss and extra dependency caused by separating frequency and temporal pooling in the baseline, we propose to capture the global dependency using Transformer encoders instead of GATs. Furthermore, since speech has temporal information, we hypothesize that positional information is critical for anti-spoofing. Thus, we propose the *positional aggregator* to integrate the local-global dependency better. To explore the optimal method for adding positional information, we design Rawformer with 1D and 2D positional encoding, respectively.

Figure 1 shows the overall structure of the proposed Rawformer. The Rawformer with 1D or 2D positional encoding consists of three main modules: a Rawnet2 front-end, a *positional aggregator*, and a classifier equipped with Transformer encoders. Firstly, the Rawnet2 front-end is used to obtain a higher-level feature map (HFM). Next, the *positional aggregator* introduces positional information to the HFM and maps the three-dimensional HFM to a two-dimensional sequence. Then, the two-dimensional sequence is taken as the input to the classifier with Transformer encoders for scoring.

3.1. Local dependency representation

Rawformer uses a Rawnet2 front-end similar to the AASIST [16] to obtain the higher-level feature map (HFM). First, a sinc layer [22] consisting of a series of sinc functions is used to extract spectro-temporal features from the raw waveform to form the lower-level feature map (LFM) $S_{LFM} \in \mathbb{R}^{F \times T}$, where F and T are the numbers of frequency and the temporal bins, respectively. Next, the HFM $S_{HFM} \in \mathbb{R}^{C \times f \times t}$ capturing the local dependency is obtained by processing the LFM with N two-dimensional convolution blocks. Here, C, f, t denotes the number of channels, frequency bins, and temporal locations after dimensionality reduction. In addition to using the same ResNet blocks [23] as in AASIST, we propose a Res-SERes2Net block, which integrates squeeze-and-excitation operation [7] into the ResNet block to enhance the connection between convolutional channels.

3.2. Positional aggregator

Following the hypothesis that positional information is crucial to capturing global dependency, we propose the *positional aggregator* to add positional information to the HFM. We develop the Rawformer with 1D and 2D *positional aggregators*. The 1D *positional aggregator* adds 1D positional information after flattening, while the 2D *positional aggregator* adds 2D positional information before flattening as shown in Figure 1. We aim to explore the impact of different positional encodings on integrating speech local-global dependency.

1D positional aggregator. Figure1 (a) shows the structure of the 1D *positional aggregator*. Firstly, the HFM is flattened along the time and then the frequency axes, where the HFM is reshaped into a two-dimensional sequence with length $f \times t$ and channels C. We then apply 1D Sinusoidal Positional Encoding [20] to assign positional information to the sequence. In particular, alternate sine and cosine functions are used to add positional information to the two-dimensional sequence, as follows:

$$P\left(x, 2i\right) = sin\left[\frac{x}{10000^{\left(\frac{2i}{C}\right)}}\right] \tag{1}$$

$$P\left(x, 2i+1\right) = cos\left[\frac{x}{10000^{\left(\frac{2i}{C}\right)}}\right] \tag{2}$$

Here, $x \in (0, f \times t)$ is the index of the combined frequency and temporal dimensions (after flattening), $i \in \left[0, \frac{C}{2}\right)$ is the channel index.

2D postional aggregator. Figure1 (b) shows the structure of the 2D *positional aggregator*. The HFM is positionally encoded with the 2D Sinusoidal Positional Encoding [24]. In particular, two alternate groups of sine functions and cosine functions are used as follows:

$$P\left(x, y, 2i\right) = sin\left[\frac{x}{10000^{\left(\frac{4i}{C}\right)}}\right] \tag{3}$$

$$P\left(x, y, 2i+1\right) = cos\left[\frac{x}{10000^{\left(\frac{4i}{C}\right)}}\right] \tag{4}$$

$$P\left(x, y, 2j + \frac{C}{2}\right) = sin\left[\frac{y}{10000^{\left(\frac{4j}{C}\right)}}\right] \quad (5)$$

$$P\left(x, y, 2j + 1 + \frac{C}{2}\right) = cos\left[\frac{y}{10000^{\left(\frac{4j}{C}\right)}}\right] \quad (6)$$

Here, (x, y) is a point in spectro-temporal space, $x \in (0, f)$, $y \in (0, t)$, are the frequency and time indices, and $i, j \in \left[0, \frac{C}{4}\right)$ indicate the channels.

The new HFM contains positional information in both the time and frequency dimensions. Then, it is reshaped into a two-dimensional sequence with length $f \times t$ and channels C. Though the 2D *positional aggregator* reflects the 2D positional relationship of HFM, it introduces stronger positional information, which may weaken identifiable information. Thus, it is hard to predict which *positional aggregator* is more recommended in the Rawformer. The comparative experiment between 1D and 2D *positional aggregators* is carried out in Section 4.5.

3.3. Global dependency modeling

The Rawformer proposed in this paper uses M Transformer encoders to process the HFM with added positional information and further excavate the global dependency. The structure of the Transformer encoder is the same as that in [20], including multi-head attention and feed-forward modules. It processes the sequence as a whole and needn't rely on hidden layers that record previous states like the LSTM [19] to capture global dependency. In addition, the layer norm and residual structure are used between two modules in a Transformer encoder to prevent the gradient from disappearing. Finally, embeddings are input into Sequence Pool [21] layer and Linear layer successively.

4. EXPERIMENTS AND ANALYSIS

4.1. Datasets and metrics

We conduct experiments on the ASVspoof 2019 LA [19] and ASVspoof 2021 LA [3] to validate the effectiveness and generalization ability of the systems. The ASVspoof 2019 LA dataset is based upon a standard multi-speaker speech synthesis database with no significant channel or background noise effects. The ASVspoof 2021 LA dataset contains spoofed and bonafide speech communicated across telephony and VoIP networks with various coding and transmission effects. The minimum tandem detection cost function (min t-DCF) [25] and the equal error rate (EER) [2] are two metrics we used in this paper.

4.2. Implementation details

During the training stage, speech is cropped or concatenated to fix a length of 4 seconds. We set the sinc layer with 70 filters and use fixed cut-in and cut-off in each filter. The AdamW [26] optimizer with a learning rate of 8×10^{-4} is used. We use the BCELoss and combine the training and development set of ASVspoof 2019 LA to train the models for 300 epochs. The training batch size is 32. We adopt two methods to preprocess the test speech in the test phase. One is to fix speech to 4 seconds, the same strategy as in the training stage (Fix); the other is directly inputting speech with different durations (Var).

4.3. Results of Rawformer with different configurations

This section explores the appropriate combination of N two-dimensional convolution blocks and M Transformer encoders to capture the local-global dependency for anti-spoofing. All the experiments are performed on the Rawformer with 1D positional encoding (Rawformer_1D-PE) and ResNet blocks. As shown in Table 1, we first fix M as 2 and then change the value of N. The results show that if N equals 4, 5, and 6, the EERs are similar when testing with a fixed-length setting, while it has different degrees of performance improvement when testing with variable lengths. Among them, 4 ResNet blocks with 2 Transformer encoders obtain the best performance in the variable-length setting. Next, we add the number of Transformer encoders to 3 and fix the ResNet blocks as 4 and 6, respectively. According to the experimental results, the system with 6 ResNet blocks and 3 Transformer encoders gets a pretty good result in both fixed and variable length settings. Thus, we perform the following experiments based on these two configurations: $N = 4$, $M = 2$ (Rawformer-S), and $N = 6$, $M = 3$ (Rawformer-L). Furthermore, we investigate the use of Res-SERes2Net block replacing ResNet blocks of Rawformer-S which we refer to as squeeze-and-excitation Rawformer (SE-Rawformer). We use the ResNet block as the first block to enlarge the number of channels to 32, and three Res-SERes2Net blocks are used as the following blocks.

Table 1. Comparison of Rawformer with different configurations on ASVspoof 2019 LA eval set with fixed-length (Fix) and variable-length (Var) settings. The results shown are the best results from one run.

N	M	Fix		Var	
		EER(%)	**t-DCF**	**EER(%)**	**t-DCF**
2	2	3.23	0.0928	4.31	0.1245
3		2.81	0.0823	1.54	0.0429
4		1.16	0.0361	**0.61**	**0.0181**
5		**0.99**	**0.0328**	0.65	0.0188
6		1.06	0.0328	0.75	0.0224
7		1.25	0.0421	0.87	0.0273
4	3	1.10	0.0366	0.60	0.0176
6		**0.84**	**0.0271**	**0.56**	**0.0164**

4.4. Comparison with the baseline system

We compare the performance of the Rawfomer with the AASIST baseline on the ASVspoof 19 LA and ASVspoof 2021 LA in Table2[1]. It is worth noting that the AASIST and Rawformer-L share the same HFM with the shape of $64 \times 23 \times 29$. The AASIST-4Block and Rawformer-S also have HFM with the shape of $64 \times 23 \times 16$. The Rawformer-based systems have fewer or similar parameters and flops using the same front-end compared with the AASIST-based systems. According to the experimental results on the ASVspoof 2019 LA, the proposed Rawformer systems exhibit good performances, which are close to or better than the AASIST-based systems with a fixed-length setting (Fix). When testing with variable lengths (Var), the performance of most Rawformer-based systems improves substantially, while the AASIST performs worse than the fixed-length setting. From this, we conclude that Rawformer can extract positional-related local-global dependency with less information loss than GAT-based AASIST. Taking advantage of Transformer's more efficient acquisition of global dependency, our proposed Rawformer reorganizes frequency bins and temporal locations into a longer ordered sequence with the help of positional encod-

[1]We obtain AASIST's results in the combined dataset of training and development of ASVspoof 2019 LA using our environment.

Table 2. Comparison between Rawformer systems and AASIST [16] on ASVspoof 2019 LA eval set and ASVspoof 2021 LA eval set with fixed-length (Fix) and variable-length (Var) settings. LE and PE are abbreviations for learnable positional encoding and positional encoding, respectively. The results shown are the average of the best results from three runs.

Model	PE	Param(M)	Flops(G)	19LA(Fix)		19LA(Var)		21LA(Fix)		21LA(Var)	
				EER(%)	t-DCF	EER(%)	t-DCF	EER(%)	t-DCF	EER(%)	t-DCF
AASIST [16]	LE	0.30	9.13	**0.93**	**0.0285**	1.58	0.0495	10.51	0.4884	12.68	0.5380
AASIST-4Block	LE	0.21	3.14	1.20	0.0341	2.56	0.0811	9.15	0.4370	11.81	0.5052
Rawformer-S	-	0.18	3.15	1.40	0.0465	0.86	0.0275	6.44	0.3515	6.42	0.3482
	1D	0.18	3.15	1.24	0.0405	0.71	0.0227	5.11	**0.3142**	5.03	0.3122
	2D	0.18	3.15	1.27	0.0426	0.88	0.0276	8.21	0.3991	8.26	0.4012
Rawformer-L	-	0.29	9.17	1.05	0.0340	0.86	0.0253	8.72	0.3971	10.01	0.4254
	1D	0.29	9.17	0.95	0.0308	0.69	0.0204	6.83	0.3609	8.07	0.3850
	2D	0.29	9.17	1.01	0.0334	0.63	0.0198	7.90	0.3859	8.87	0.4095
SE-Rawformer	-	0.37	6.10	1.49	0.0471	1.01	0.0293	9.57	0.4180	9.50	0.4231
	1D	0.37	6.10	1.05	0.0344	**0.59**	**0.0184**	**4.98**	0.3186	**4.53**	**0.3088**
	2D	0.37	6.10	1.47	0.0456	0.97	0.0296	5.40	0.3309	5.16	0.3253

ings. Therefore, the sequence contains the spectro-temporal positional relationship in the original speech improving anti-spoofing performance. Furthermore, the SE-Rawformer achieves an EER of 0.59% and a min t-DCF of 0.0184 on ASVspoof 2019 LA, which are the best reported in the literature.

To verify the generalization of the Rawformer, we conduct a series of experiments on ASVspoof 2021 LA without any data augmentation strategy. When testing with a variable-length setting, the performance of Rawformer-S and SE-Rawformer with 4 blocks improves, while the Rawformer-L with 6 blocks degrades, compared with the fixed-length setting. As such, a deeper front-end may overfit the training data and reduce generalization ability. The Rawformer systems generally perform better than the AASIST. Thus, the Rawformer is more robust to out-of-domain data. In particular, the SE-Rawformer obtains a significant boost in EER of 4.53% and a min t-DCF of 0.3088 on ASVspoof 2021 LA.

4.5. Results of Rawformer with positional aggregator

In this section, we investigate the effects of *positional aggregator* on modeling speech local-global dependency with three Rawformer systems. As shown in Table 2, adding positional information by 1D and 2D *positional aggregators* to the HFM benefits the anti-spoofing performance by 26% and 12% on average, respectively. Therefore, *positional aggregator* is beneficial to preserving speech's spectro-temporal positional information, which is crucial for integrating local-global dependency.

To better understand the differences between 1D and 2D *positional aggregators*, the visualization method is used to analyze the output positional encoding of 1D and 2D *positional aggregators*. Figure 2 shows the 1D and 2D positional encoding with the same HFM. The 2D Positional Encoding has a periodicity with t of $\frac{C}{2}$, and the rest $\frac{C}{2}$ at every position in each period are the same, which means that all positions in a period have the same frequency domain. Though 2D *positional aggregator* reflects the spectro-temporal positional relationship of HFM, it reduces the difference between positions and weakens the long-term dependency of speech. The experiment results indicate that 1D *positional aggregator* is more conducive to modeling the local-global dependency than 2D *positional aggregator* which leads to a 43% improvement on average in our best model SE-Rawformer.

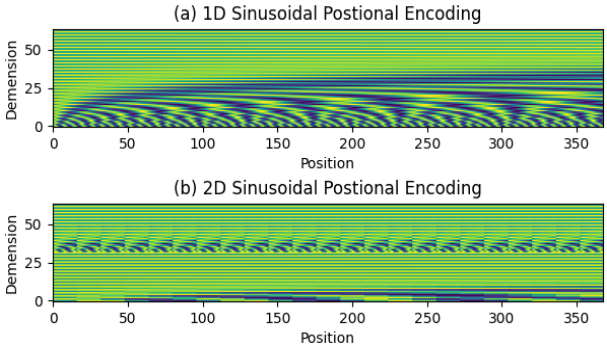


Fig. 2. Visualization of the output positional encoding of 1D and 2D *positional aggregators*. Note that the shape of HFM is $64 \times 23 \times 16$, where 64, 23, and 16 stand for channel numbers (C), spectral bins (f), and temporal bins (t), respectively.

5. CONCLUSIONS

We proposed an anti-spoofing countermeasure focusing on mining positional-related local-global dependency, referred to as the Rawformer. Our proposed *positional aggregator* contributes to a better integration of local-global dependency with less information loss. Moreover, the Rawformer has excellent capability to handle variable-length input speech compared to the GAT-based baseline. Experimental results show that 1D *positional aggregator* is more recommended in the Rawformer. The SE-Rawformer using our proposed Res-SERes2Net blocks to enhance the connection between convolutional channels achieved an EER of 0.59% and a min t-DCF of 0.0184 on ASVspoof 2019 LA, which outperformed all the single state-of-the-art systems. Furthermore, the Rawformer is relatively robust to unseen domains. In future work, we will introduce the one-class learning method to further enhance the generalization of the Rawformer to unknown attacks.

6. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grant 62176182. The work of Kong Aik Lee is supported by the Agency for Science, Technology and Research (A*STAR), Singapore, through its Council Research Fund (Project No. CR-2021-005).

7. REFERENCES

[1] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov, "ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge," in *Proc. Interspeech*, 2015, pp. 2037–2041.

[2] Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi H Kinnunen, Ville Vestman, et al., "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.

[3] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas 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.

[4] Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee, "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," in *Proc. Interspeech*, 2017, pp. 2–6.

[5] Weicheng Cai, Haiwei Wu, Danwei Cai, and Ming Li, "The DKU Replay Detection System for the ASVspoof 2019 Challenge: On Data Augmentation, Feature Representation, Classification, and Fusion," in *Proc. Interspeech 2019*, 2019, pp. 1023–1027.

[6] Yexin Yang, Hongji Wang, Heinrich Dinkel, Zhengyang Chen, Shuai Wang, Yanmin Qian, and Kai Yu, "The sjtu robust anti-spoofing system for the asvspoof 2019 challenge.," in *Interspeech*, 2019, pp. 1038–1042.

[7] Xu Li, Na Li, Chao Weng, Xunying Liu, Dan Su, Dong Yu, and Helen Meng, "Replay and synthetic speech detection with res2net architecture," in *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2021, pp. 6354–6358.

[8] Xu Li, Xixin Wu, Hui Lu, Xunying Liu, and Helen Meng, "Channel-Wise Gated Res2Net: Towards Robust Detection of Synthetic Speech Attacks," in *Proc. Interspeech 2021*, 2021, pp. 4314–4318.

[9] You Zhang, Fei Jiang, and Zhiyao Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[10] Xinhui Chen, You Zhang, Ge Zhu, and Zhiyao Duan, "UR Channel-Robust Synthetic Speech Detection System for ASVspoof 2021," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 75–82.

[11] Xinyue Ma, Tianyu Liang, Shanshan Zhang, Shen Huang, and Liang He, "Improved lightcnn with attention modules for asv spoofing detection," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[12] Zhenchun Lei, Hui Yan, Changhong Liu, Minglei Ma, and Yingen Yang, "Two-path gmm-resnet and gmm-senet for asv spoofing detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6377–6381.

[13] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

[14] Hemlata Tak, Jee weon Jung, Jose Patino, Massimiliano Todisco, and Nicholas W. D. Evans, "Graph attention networks for anti-spoofing," in *Interspeech*, 2021.

[15] Hemlata Tak, Jee weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 1–8.

[16] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas 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.

[17] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee weon Jung, Junichi Yamagishi, and Nicholas W. D. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *Odyssey*, 2022.

[18] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio, "Graph attention networks," *Proc.ICLR*, 2018.

[19] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, et al., "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, pp. 101114, 2020.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] Changtao Li, Feiran Yang, and Jun Yang, "The role of long-term dependency in synthetic speech detection," *IEEE Signal Processing Letters*, vol. 29, pp. 1142–1146, 2022.

[22] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016.

[24] Zobeir Raisi, Mohamed A Naiel, Paul Fieguth, Steven Wardell, and John Zelek, "2d positional embedding-based transformer for scene text recognition," *Journal of Computational Vision and Imaging Systems*, vol. 6, no. 1, pp. 1–4, 2020.

[25] Tomi Kinnunen, Kong Aik Lee, Héctor Delgado, Nicholas Evans, Massimiliano Todisco, Md Sahidullah, Junichi Yamagishi, and Douglas A Reynolds, "t-dcf: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Speaker Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018.

[26] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.