INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Robust Audio Anti-Spoofing with Fusion-Reconstruction Learning on Multi-Order Spectrograms

Penghui Wen[1], *Kun Hu*[1,*], *Wenxi Yue*[1], *Sen Zhang*[1], *Wanlei Zhou*[2], *Zhiyong Wang*[1]

[1]School of Computer Science, The University of Sydney, NSW, Australia
[2]Faculty of Data Science, City University of Macau, Macao SAR, China

pwen5103@uni.sydney.edu.au, kun.hu@sydney.edu.au, wenxi.yue@sydney.edu.au,
szha2609@uni.sydney.edu.au, wlzhou@cityu.mo, zhiyong.wang@sydney.edu.au

Abstract

Robust audio anti-spoofing has been increasingly challenging due to the recent advancements on deepfake techniques. While spectrograms have demonstrated their capability for anti-spoofing, complementary information presented in multi-order spectral patterns have not been well explored, which limits their effectiveness for varying spoofing attacks. Therefore, we propose a novel deep learning method with a spectral fusion-reconstruction strategy, namely S^{2}pecNet, to utilise multi-order spectral patterns for robust audio anti-spoofing representations. Specifically, spectral patterns up to second-order are fused in a coarse-to-fine manner and two branches are designed for the fine-level fusion from the spectral and temporal contexts. A reconstruction from the fused representation to the input spectrograms further reduces the potential fused information loss. Our method achieved the state-of-the-art performance with an EER of 0.77% on a widely used dataset - ASVspoof2019 LA Challenge.

Index Terms: audio spoofing detection, anti-spoofing, audio feature fusion, deep learning

1. Introduction

Audio based automatic speaker verification (ASV) has a wide range of applications due to the biometric authentication property of voice [1], including multi-speaker speech recognition [2], speech authentication-based telephone banking and voice-based forensics [3]. To improve the robustness of ASV systems against increasingly complex deepfake techniques, various methods have been devised [4, 5, 6] to identify audio spoofing.

The identification of audio spoofing is generally treated as a binary classification task which classifies an audio recording as genuine or spoofed. Early studies primarily focused on devising hand-crafted features to anti-spoofing, such as Cochlear Filter Cepstral Coefficient Instantaneous Frequency (CFCCIF) [7], Linear Frequency Cepstral Coefficients (LFCC) [8] and Constant-Q Cepstral Coefficients (CQCC) [9]. Recently, various deep learning methods have been proposed for audio deepfake recognition. For example, a convolutional neural network (CNN) was first adopted with 2D audio spectrograms [10], then the deep residual network (ResNet) methods (e.g., [11]) were proposed to formulate the anti-spoofing problem as one-class feature learning [12] to improve the generalisability. Based on the effectiveness of sub-band spectrogram features in anti-spoofing, a dual-band fusion algorithm was proposed [13]. To characterise temporal relations in audio signals, a recurrent neural network (RNN) based method was proposed in [14]. More recently, a spectro-temporal graph attention network method AASIST [15] was proposed to formulate spoofing patterns with 1st-order spectrograms (i.e., raw spectrograms).

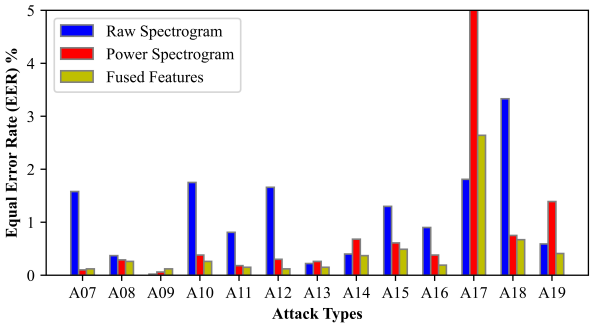

Figure 1: *Illustration of the performance for anti-spoofing on ASVspoof2019 LA Challenge, which is highly sensitive with the order of the spectral features used.*

These existing methods are often based on a specific category of audio features. However, as shown in Figure 1, different types of features exhibit varying effectiveness in detecting different types of attacks. Thus, instead of using a single source of spectral features as in AASIST, we suggest that diverse orders of audio spectral patterns can benefit the speech anti-spoofing in a complementary manner. For example, the 2nd-order spectrograms (i.e., power spectrograms) are suggested to be more sensitive to the noise patterns in real-world speech [16]. As shown in Figure 3, the power spectrogram can detect subtle variations regarding spoofing clues in high frequency regions with low amplitude values, compared with the raw spectrogram.

Therefore, in this study, a novel deep learning architecture, namely S^{2}pecNet, is proposed for robust anti-spoofing by using multi-order spectrogram patterns, which are up to the second-order including both raw and power spectrograms. Specifically, raw and power spectral patterns are fused in a coarse-to-fine manner and two branches are involved for the fine-level fusion from the spectral and temporal contexts. To minimize the information loss during the feature learning and fusion, a reconstruction mechanism is devised to reconstruct the fused representation to its associated input spectrograms. Comprehensive experiments on a commonly used dataset - ASVspoof 2019 LA, demonstrates the effectiveness of our proposed method, S^{2}pecNet, which achieves the state-of-the-art performance regarding the metrics minimum tandem detection cost function (min t-DCF) [17] and equal error rate (EER) [18].

In summary, the key contributions of this work are: (i) a novel deep learning based fusion architecture for audio anti-

*Corresponding Author. This study was partially supported by Australian Research Council (ARC) grant #DP210102674. Our code is available at https://github.com/ph-w2000/S2pecNet.

10.21437/Interspeech.2023-563

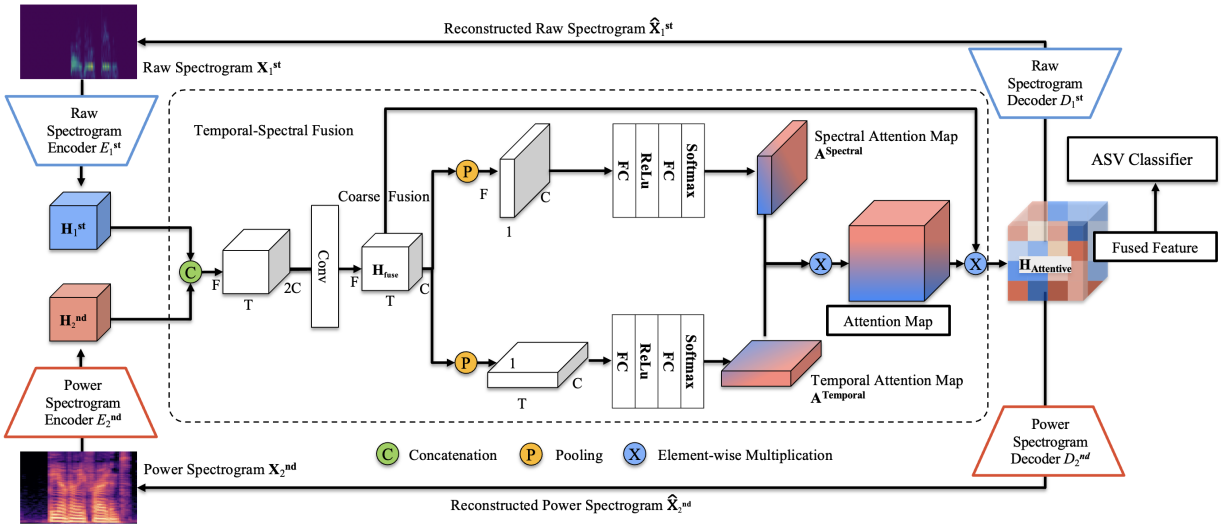

Figure 2: *Illustration of the overall architecture for the proposed S^{2}pecNet.*

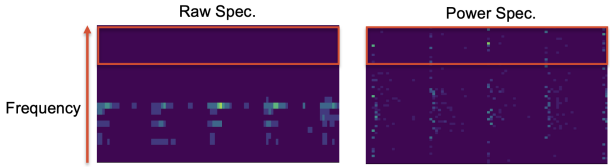

Figure 3: *Illustration of raw and power spectrograms, where the area in red bounding boxes represents high frequency regions.*

spoofing with multi-order spectrograms; (ii) a coarse-to-fine fusion mechanism with two branches that are involved for the fine-level fusion from the spectral and temporal contexts; and (iii) a reconstruction strategy to maintain the information in the fused speech representations.

2. Proposed Method

Figure 2 illustrates the overall architecture of the S^{2}pecNet method. The 1st-order raw spectrogram and the 2nd-order power spectrogram of an input audio are first fed into their encoders, respectively. Next, the two encoded features are concatenated and fed into a temporal-spectral fusion module in pursuit of a spoofing-sensitive representation by exploiting the supplementary information between the two spectrograms. Additionally, to minimise the potential information loss of the fused representations, a reconstruction mechanism is introduced with two decoders to reconstruct the fused features back to the original raw and power spectrograms.

2.1. Raw Spectrogram and Power Spectrogram Encoding

S^{2}pecNet takes an input audio waveform $\mathbf{X}$ from its first-order and second-order spectral characteristics as input in pursuit of a comprehensive representation of various spoofing patterns. For the first-order patterns, the input audio's raw spectrogram $\mathbf{X}_{1st}$ is fed into a CNN based encoder E_{1st} to formulate an audio feature map $\mathbf{H}_{1st} \in \mathbb{R}^{C\times F\times T}$, where C, F, and T denote the number of channels, the number of spectral bins, and sequence length, respectively. In terms of the second-order patterns, the input audio's power spectrogram $\mathbf{X}_{2nd}$ is encoded by another CNN encoder E_{2nd} and a feature map can be obtained as $\mathbf{H}_{2nd} \in \mathbb{R}^{C\times F\times T}$. Note that E_{1st} and E_{2nd} are set to generate their output feature maps with the same dimension.

2.2. Temporal-Spectral Fusion

The output feature maps $\mathbf{H}_{1st}$ and $\mathbf{H}_{2nd}$ of the two encoders are with different spectral orders. Therefore, a temporal-spectral fusion (TSF) module is devised to refine and fuse the two feature maps. TSF formulates the dependencies between the two spectral domains and explores their complementary spoofing-related patterns in a coarse-to-fine manner. Initially, a coarse fusion step is performed by concatenating $\mathbf{H}_{1st}$ and $\mathbf{H}_{2nd}$ in a channel-wise manner and applying a set of convolution filters on the concatenated feature map to obtain a coarse fused representation $\mathbf{H}_{\text{fuse}}$. Then, to characterise finer spoofing-sensitive features from $\mathbf{H}_{\text{fuse}}$, an attention map $\mathbf{A}$ is obtained to highlight its patterns that are more susceptible to spoofing by formulating the long-term temporal dependencies and the spectral patterns.

To obtain $\mathbf{A}$, two sub-attention maps $\mathbf{A}^{\text{spectral}}$ and $\mathbf{A}^{\text{temporal}}$ are derived from two different contexts: one explores the temporal context while the other explores the spectral context. In detail, $\mathbf{H}_{\text{fuse}}$ is pooled along its temporal and spectral dimensions, respectively, and we have:

$$\mathbf{H}^{\text{spectral}}_{\text{fuse}} = \max_t(|\mathbf{H}_{\text{fused}}|), \mathbf{H}^{\text{temporal}}_{\text{fuse}} = \max_s(|\mathbf{H}_{\text{fused}}|), \quad (1)$$

where $\mathbf{H}^{\text{spectral}}_{\text{fuse}} \in \mathbb{R}^{C\times F\times 1}$, $\mathbf{H}^{\text{temporal}}_{\text{fuse}} \in \mathbb{R}^{C\times 1\times T}$, $|\cdot|$ refers to an element-wise absolute operator, $\max_s$ is a global spectral pooling operator and $\max_t$ indicates a global temporal pooling operator. To this end, $\mathbf{H}^{\text{spectral}}_{\text{fuse}}$ contains global temporal information across frequency bins, and $\mathbf{H}^{\text{temporal}}_{\text{fuse}}$ contains global spectral information across time. Next, the two attention maps $\mathbf{A}^{\text{spectral}} \in \mathbb{R}^{C\times F\times 1}$ and $\mathbf{A}^{\text{temporal}} \in \mathbb{R}^{C\times 1\times T}$ are obtained as:

$$\mathbf{A}^{\text{spectral}} = \text{Conv}_s(\mathbf{H}^{\text{spectral}}_{\text{fuse}}), \mathbf{A}^{\text{temporal}} = \text{Conv}_t(\mathbf{H}^{\text{temporal}}_{\text{fuse}}), \quad (2)$$

where Conv_t and Conv_s denote the convolution layers for obtaining the two attention maps, respectively. To this end, the final attention map $\mathbf{A}$ is obtained by: $\mathbf{A} = \mathbf{A}^{\text{spectral}} \times \mathbf{A}^{\text{temporal}}$, and the final fused representation can be derived as:

$$\mathbf{H}_{\text{attentive}} = \mathbf{A} \times \mathbf{H}_{\text{fused}}. \quad (3)$$

2.3. Raw Spectrogram and Power Spectrogram Decoding

To prevent information loss during the encoding procedure and the feature fusion, a raw spectrogram decoder D_{1st} and a power spectrogram decoder D_{2nd} are devised to reconstruct the input raw spectrograms and power spectrograms, respectively. Specifically, D_{1st} consists of a series of deconvolution layers, which takes the fused feature $\mathbf{H}_{\text{attentive}}$ to reconstruct the raw spectrogram $\hat{\mathbf{X}}_{1st}$. Similarly, D_{2nd} applies deconvolution layers to $\mathbf{H}_{\text{attentive}}$ and produce the reconstructed power spectrogram $\hat{\mathbf{X}}_{2nd}$. A raw spectrogram reconstruction loss $\mathcal{L}_{1st}$ and a power spectrogram reconstruction loss $\mathcal{L}_{2nd}$ are introduced as:

$$\mathcal{L}_{1st} = \|\hat{\mathbf{X}}_{1st} - \mathbf{X}_{1st}\|, \mathcal{L}_{2nd} = \|\hat{\mathbf{X}}_{2nd} - \mathbf{X}_{2nd}\|. \quad (4)$$

Minimizing the two losses aims to best reconstruct the original raw and power spectrograms using the fused representation.

2.4. Spoofing Detection

A classifier further takes the fused representation $\mathbf{H}_{\text{attentive}}$ as input to produce a binary classification prediction $\hat{y}$. The classification loss $\mathcal{L}_{\text{cls}}$ is with a weighted binary cross-entropy (WACE) to quantify the difference between the prediction $\hat{y}$ and the ground truth y, which can be formulated as:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N} y \cdot \log(\hat{y}) + (1-y) \cdot \log(1-\hat{y}). \quad (5)$$

2.5. Model Training

The overall loss $\mathcal{L}$ of the proposed method S^{2}pecNet is with the three loss terms mentioned above, including the raw spectrogram reconstruction loss $\mathcal{L}_{1st}$, the power spectrogram reconstruction loss $\mathcal{L}_{2nd}$, and the classification loss $\mathcal{L}_{\text{cls}}$. The relative importance of them is controlled by a hyper-parameter α:

$$\mathcal{L} = \alpha(\mathcal{L}_{1st} + \mathcal{L}_{2nd}) + \mathcal{L}_{\text{cls}}. \quad (6)$$

3. Experiments & Discussions

3.1. Dataset and Evaluation Metrics

A widely used dataset, ASVspoof2019 LA [19], was adopted for evaluation. It consists of both bona fide audio recordings and 19 different types of spoofing attacks generated through text-to-speech (TTS) [20] and voice conversion (VC) [21]. We followed the same partitions of training, development, and evaluation as in AASIST. The training and development partitions include 6 different spoofing attacks (A01-A06), while the evaluation partition includes 13 different attacks (A07-A19). The overview of this dataset is listed in Table 1. For evaluation metrics, the minimum tandem detection cost function (min t-DCF) and the equal error rate (EER) were adopted. Moreover, it has been demonstrated that the performance of spoofing detection algorithms can vary greatly depending on initial random seeds [5]. Hence, we report the average metrics across a number of random seeds.

Table 1: *Overview of the ASVspoof2019 LA dataset*

Partition	Bona fide	Spoofed	
	# utterance	# utterance	attacks
Training	2,580	22,800	A01 - A06
Development	2,548	22,296	A01 - A06
Evaluation	7,355	63,882	A07 - A19

3.2. Implementation Details

A raw waveform was obtained with 64,600 frames (approximately 4 seconds). A sinc-convolution filter was used to obtain the raw spectrogram of an input audio, which was further encoded through a ResNet encoder with 6 residual blocks. For the power spectrogram, we formulated a 60-dimensional LFCC features for a frame, of which the size is 20 ms with a hop size of 10 ms. A ResNet-18 was utilised as the power spectrogram encoder. The classifier for spoofing detection followed the setting as in AASIST. Both CONV_s and CONV_t of the TSF module consisted of two fully-connected layers, a batch normalization layer, a SiLU function and a sigmoid function. Our S^{2}pecNet was implemented and trained using PyTorch for 100 epochs on an NVIDIA RTX A6000 GPU with a batch size 48. An Adam optimizer was adopted with a learning rate of 3×10^{-4} and a cosine annealing learning rate decay.

Table 2: *Comparison with the state-of-the-art methods. min t-DCF and EER values were from AASIST, while # of parameters and inference time were based on official implementations, indicating the average inference duration for one-second audio.*

Method	#Param	Runtime	min t-DCF ↓	EER ↓
AASIST	297k	0.0052	0.028	0.83
RawGAT-ST [22]	437k	0.0049	0.034	1.06
MCG-Res2Net [23]	960k	-	0.052	1.78
OC-Softmax [11]	12450k	0.0018	0.059	2.19
SE-Res2Net [24]	920k	-	0.074	2.50
S^{2}pecNet (Ours)	1284k	0.0072	**0.024**	**0.77**

3.3. Performance Comparison

Table 2 lists a performance comparison between our proposed S^{2}pecNet and the state-of-the-art methods. It can be observed that S^{2}pecNet outperforms these existing methods and shows strong capability for robust audio spoofing detection. Specifically, Table 3 lists the comparison between our S^{2}pecNet and the state-of-the-art AASIST method regarding the best metrics. Our S^{2}pecNet demonstrates superior performance in terms of both min t-DCF and EER, where our method improves the average EER by 25%. In addition, S^{2}pecNet has superior or comparable performance on most attacks, except for the A17 attack where the AASIST model outperforms ours with a significant gap. A17 employed an acoustic model VAE-GAN [25], which generate spectral shapes that are more realistic and detailed in the high-frequency patterns. The power spectrogram can be fooled by A17 since its modelling relies on observing subtle variations in high frequency regions.

3.4. Ablation Study

3.4.1. Spectral Complementary Patterns

An ablation study was conducted to investigate the impact of the complementary information between the two spectrograms. Three settings were investigated under the same condition with raw spectrograms, power spectrograms and fused spectrograms (i.e., the representations obtained from S^{2}pecNet), respectively.

As listed in Table 4, raw spectrogram is able to achieve significantly better performance than power spectrogram on A17 and A19, while the power spectrogram demonstrates largely better performance on A07, A10, A11, A12, A15, A16, and A18. Since each setting demonstrated its advantages on different attacks, the fused spectrograms can effectively exploit the

Table 3: *Comparison with AASIST. Results are based on EER and values in parentheses show the best readout.*

System	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	min t-DCF	EER (%)
AASIST	0.80	0.44	**0.00**	1.06	0.31	0.91	**0.10**	**0.14**	0.65	0.72	**1.52**	3.40	0.62	0.035(0.028)	1.13(0.83)
Ours	**0.12**	**0.26**	0.12	**0.26**	**0.15**	**0.12**	0.15	0.37	**0.49**	**0.19**	2.64	**0.67**	**0.41**	**0.027(0.024)**	**0.84(0.77)**

complementary information of them to achieve the best performance in most attacks. Grad-CAM [26] was adopted to visualise the attention maps on the two spectrograms. As shown in Figure 4, S^{2}pecNet is able to identify high frequency information from the power spectrogram, while focusing on low frequency information from the raw spectrogram, which confirms that utilising two spectrograms can explore the complementary information together for better spoofing detection.

Table 4: *Performance comparisons regarding EER on raw spectrograms, power spectrograms, and fused spectrograms.*

Attacks ↓	Raw spec.	Power spec.	Fused spec.
A07	1.58	**0.10**	0.12
A08	0.37	0.29	**0.26**
A09	**0.02**	0.06	0.12
A10	1.75	0.48	**0.26**
A11	0.81	0.18	**0.15**
A12	1.66	0.30	**0.12**
A13	0.22	0.26	**0.15**
A14	0.40	0.68	**0.37**
A15	1.30	0.61	**0.49**
A16	0.90	0.38	**0.19**
A17	**1.81**	31.46	2.64
A18	3.33	0.75	**0.67**
A19	0.59	1.39	**0.41**

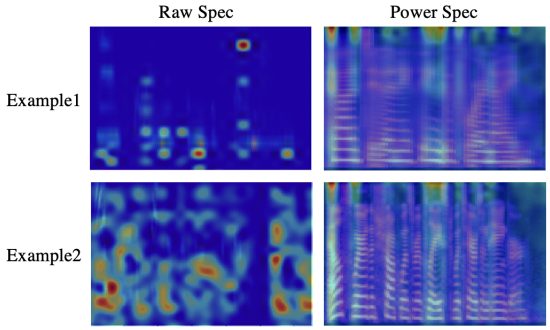

Figure 4: *Grad-CAM on spectrograms of two examples.*

3.4.2. Impact of TSF Module

To evaluate the effectiveness of the TSF module, comparisons were conducted with other fusion methods, including concatenation, early fusion, late fusion, and other state-of-the-art fusion methods. As shown in Table 5, concatenation can achieve reasonably good performance by avoiding information loss, whereas many complex fusion methods lead to worse performance, which suggest greater information loss. Our TSF module can effectively exploit the complementary information between two spectral domains embedded with temporal dependencies, and achieved the best performance.

Table 5: *Comparisons with different fusion methods.*

Method	EER(%) ↓	min t-DCF ↓
Early-Fusion [27]	7.90	0.1800
Late-Fusion [28]	3.53	0.0848
MFFN [29]	3.24	0.0645
CEFNet(ACM) [30]	1.48	0.0465
SA-Fuser(w) [31]	1.32	0.0320
Concatenation	1.03	0.0340
TSF (Ours)	**0.84**	**0.0271**

3.4.3. Impact of Reconstruction Decoders

We further investigate the impact of the reconstruction decoders which aim to retain the original information in the final fused representation. As shown in Table 6, the reconstruction decoders can improve detection performance by retaining more useful complementary information. Additionally, the settings of hyper-parameter α are explored, which is used to balance the detection and the reconstruction. The results in Table 7 indicate that the best performance is achieved with $\alpha = 0.1$.

Table 6: *Ablation study on reconstruction decoders.*

Method	EER(%) ↓	min t-DCF ↓
w/o reconstruction decoders	0.93	0.0295
w/ reconstruction decoders	**0.84**	**0.0271**

Table 7: *Hyper-parameter selection in terms of α.*

α	EER(%) ↓	min t-DCF ↓
1	1.02	0.0295
0.1	**0.84**	**0.0271**
0.01	0.96	0.0283

4. Conclusion

We present a novel method S^{2}pecNet for audio spoofing detection by exploiting complementary information from multi-order spectrograms. Specifically, a TSF module is devised to fuse the two spectral representations in a coarse-to-fine manner. To minimize information loss, a fused representation is reconstructed to its input spectrograms. Comprehensive experiments demonstrate the superiority of S^{2}pecNet over the state-of-the-art methods. As S^{2}pecNet does not work well for some specific spoofing attacks, utilising higher-order and flexible spectral patterns in a data-driven scheme could be worth studying in future research. Additionally, considerations should be given to advanced techniques for spurious synthesis (e.g., [32]), to enhance the robustness of audio spoofing detection.

5. References

[1] P. Zhang, P. Hu, and X. Zhang, "Norm-constrained score-level ensemble for spoofing aware speaker verification," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2022.

[2] J. Xu, K. Hu, C. Xu, T. D. Chung, and Z. Wang, "Speaker-aware monaural speech separation." in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020.

[3] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[4] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of audio deepfake detection." in *The Speaker and Language Recognition Workshop (Odyssey)*, 2020, pp. 132–137.

[5] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2021.

[6] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, D. Erro, and T. Raitio, "Toward a universal synthetic speech spoofing detection using phase information," vol. 10, no. 4. IEEE, 2015, pp. 810–820.

[7] 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 *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015.

[8] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015.

[9] M. Todisco, H. Delgado, and N. W. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients." in *The Speaker and Language Recognition Workshop (Odyssey)*, 2016.

[10] C. Zhang, C. Yu, and J. H. Hansen, "An investigation of deep-learning frameworks for speaker verification antispoofing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 684–694, 2017.

[11] 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.

[12] S. S. Khan and M. G. Madden, "A survey of recent trends in one class classification," in *Conference on Artificial Intelligence and Cognitive Science (AICS)*, 2010, pp. 188–197.

[13] Y. Zhang, W. Wang, and P. Zhang, "The effect of silence and dual-band fusion in anti-spoofing system," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2021.

[14] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019.

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

[16] V. Tyagi and C. Wellekens, "On desensitizing the mel-cepstrum to spurious spectral components for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. IEEE, 2005, pp. I–529.

[17] T. Kinnunen, H. Delgado, N. Evans, K. A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, and J. Yamagishi, "Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2195–2210, 2020.

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

[19] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1008–1012, 2019.

[20] V. Shchemelinin and K. Simonchik, "Examining vulnerability of voice verification systems to spoofing attacks by means of a tts system," in *The International Conference on Speech and Computer (SPECOM)*. Springer, 2013, pp. 132–137.

[21] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4401–4404.

[22] H. Tak, J.-W. 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 *ASVspoof Workshop*, 2021.

[23] X. Li, X. Wu, H. Lu, X. Liu, and H. Meng, "Channel-wise gated Res2Net: Towards robust detection of synthetic speech attacks," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2021.

[24] X. Li, N. Li, C. Weng, X. Liu, D. Su, D. Yu, and H. Meng, "Replay and synthetic speech detection with Res2Net architecture," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6354–6358.

[25] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017.

[26] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[27] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[28] Y. Zhang, L. Gong, L. Fan, P. Ren, Q. Huang, H. Bao, and W. Xu, "A late fusion CNN for digital matting," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[29] D. Zheng, X. Zheng, L. T. Yang, Y. Gao, C. Zhu, and Y. Ruan, "MFFN: Multi-view feature fusion network for camouflaged object detection," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[30] G. Feng, Z. Hu, L. Zhang, and H. Lu, "Encoder fusion network with co-attention embedding for referring image segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[31] Z. Zhong, D. Schneider, M. Voit, R. Stiefelhagen, and J. Beyerer, "Anticipative feature fusion transformer for multi-modal action anticipation," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[32] C. Liu, H. Chen, T. Zhu, J. Zhang, and W. Zhou, "Making deepfakes more spurious: Evading deep face forgery detection via trace removal attack," *IEEE Transactions on Dependable and Secure Computing*, 2023.