End-to-End Dual-Branch Network Towards Synthetic Speech Detection

Kaijie Ma, Yifan Feng, Beijing Chen, and Guoying Zhao, *Fellow, IEEE*

***Abstract*—Synthetic speech attacks bring more threats to Automatic Speaker Verification (ASV) systems, thus many synthetic speech detection (SSD) systems have been proposed to help the ASV system resist synthetic speech attacks. However, existing SSD systems still lack the generalization ability for the attacks generated by unknown synthesis algorithms. This letter proposes an end-to-end ensemble system, namely Dual-Branch Network, in which linear frequency cepstral coefficients (LFCC) and constant Q transform (CQT) are used as the input of two branches respectively. In addition, four fusion strategies are compared for the fusion of two branches to obtain an optimal one; multi-task learning and convolutional block attention module (CBAM) are introduced into the Dual-Branch Network to help the network learn the common forgery features from different forgery types of speech and enhance the representation power of learned features. Experimental results on the ASVspoof 2019 logical access (LA) dataset demonstrate that the proposed system outperforms existing state-of-the-art systems on both t-DCF and EER scores and has good generalization for unknown forgery types of synthetic speech.**

***Index Terms*—ASVspoof 2019 LA, attention mechanism, generalization ability, multi-task learning, synthetic speech detection.**

I. Introduction

AUTOMATIC Speaker Verification (ASV) is an important part of biometric authentication system. It can confirm a speaker's identity by analyzing the acoustic characteristics of a spoken utterance [1]. However, the spoofing attacks which consist of text-to-speech (TTS), voice conversion (VC) [2] and replay may pose threats to the stability of ASV systems. Especially, the rapid development of TTS and VC brings new challenges to ASV systems. For a nearly decade, several ASVspoof challenges [3], [4], [5], [6] have been held to develop countermeasures against spoofing attacks. The logical access (LA) of the ASVspoof 2019 aims to develop synthetic speech detection (SSD) systems against synthetic speech generated by TTS and VC algorithms [5].

The aim of SSD systems is to distinguish synthetic speech from real speech. To construct efficient SSD systems, efforts may be divided into two aspects: feature engineering and model design. Many works pay attention to feature engineering because of the competitive performance of hand-crafted features such as linear frequency cepstral coefficients (LFCC), constant Q transform (CQT), constant Q cepstral coefficients (CQCC), as well as their variants and combinations [7], [8], [9], [10], [11], [12], [13]. As for the model design, Deep Neural Networks (DNNs) including Residual Neural Networks (ResNets) [14], [15], [16], [17], Light Convolutional Neural Networks (LCNNs) [18], [19], as well as their variants [20], [21], [22], and some representative networks such as Capsule Network [23], Densely Connected Convolutional Network [24], [25], Res-TSSDNet [26], One-dimensional Convolutional Transformer [27], Graph Attention Network [28], have been widely used to develop powerful SSD systems because of the advantage of DNNs in classification tasks.

It is critical for SSD systems to generalize for spoofing attacks generated by unknown synthesis algorithms, thus some existing works [18], [22], [23], [24], [25], [29] fuse the results of different single feature systems to obtain high-performance ensemble systems by taking advantages of the complementarity of different features [22]. However, these ensemble systems in which all sub-systems are optimized separately are not end-to-end. It limits the optimization of the performance. Therefore, we propose an end-to-end ensemble system in which all sub-systems are optimized at the same time to further improve the system performance. In addition, in [30], [31], the multi-task learning reduces the differences of speech characteristics from various speakers to improve the generalization of the network, thus here the multi-task learning is also introduced into SSD. The contributions of this letter are as follows:

1) We present an end-to-end Dual-Branch Network for SSD. In addition, we compare four dual-branch fusion strategies on the network performance and obtain an optimal one to guide the training of the whole network.
2) We utilize the multi-task learning to enable the network to detect different forgery types of synthetic speech. Specifically, a forgery type classifier is added to each branch as an additional task. Then, the feature learned by the network would not contain the specific forgery information associated to a certain forgery type by the adversarial training between the network and the forgery type classifier. Besides, convolutional block attention module (CBAM) is used to improve the representation power of learned features.

Manuscript received 18 December 2022; revised 27 February 2023; accepted 17 March 2023. Date of publication 27 March 2023; date of current version 13 April 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62072251, in part by the Academy of Finland for ICT 2023 project TrustFace under Grant 345948, and in part by Infotech Oulu. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiangui Kang. *(Corresponding author: Beijing Chen.)*

Kaijie Ma, Yifan Feng, and Beijing Chen are with the Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science & Technology, Nanjing 210044, China, and with the School of Computer Science, Nanjing University of Information Science &Technology, Nanjing 210044, China, and also with the Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science & Technology, Nanjing 210044, China (e-mail: ma_kaijie@126.com; fyf200613@qq.com; nbutimage@126.com).

Guoying Zhao is with the Center for Machine Vision and Signal Analysis, University of Oulu, FI-90014 Oulu, Finland (e-mail: guoying.zhao@oulu.fi).

The source code is available at https://github.com/imagecbj/End-to-End-Dual-Branch-Network-Towards-Synthetic-Speech-Detection.

Digital Object Identifier 10.1109/LSP.2023.3262419

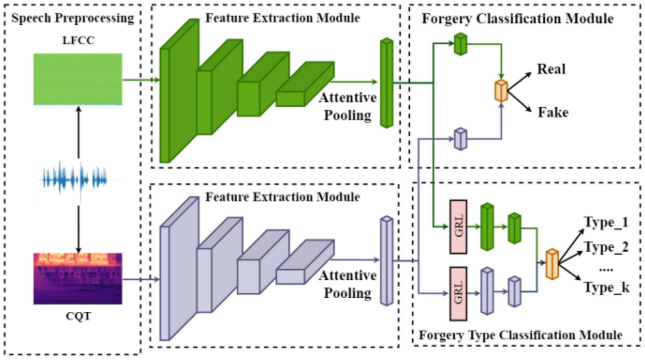

Fig. 1. Illustration of the overall structure of our dual-branch network.

TABLE I
THE DETAILED ARCHITECTURE OF THE SINGLE BRANCH NETWORK

Module	Layer	Filter size / Filters / Stride	Output
FEM	Input	-	(1, 60, 750)
	Conv1	9×3, 16, Stride(3×1)	(16, 18, 750)
	Residual Block 1	3×3, 64, Stride(1,1)	(64, 18, 750)
		[3×3, 64, Stride(1,1)] ×3	
	Residual Block 2	3×3, 128, Stride(2,2)	(128, 9, 375)
		[3×3, 128, Stride(1,1)] ×3	
	Residual Block 3	3×3, 256, Stride(2,2)	(256, 5, 188)
		[3×3, 256, Stride(1,1)] ×3	
	Residual Block 4	3×3, 512, Stride(2,2)	(512, 3, 94)
		[3×3, 512, Stride(1,1)] ×3	
	Conv2	3×3, 256, Stride(1,1)	(256, 1, 94)
	Attentive Pooling	-	(512, 1)
	Linear 1	-	(256, 1)
FTCM	Linear 2	-	(128, 1)
	Linear 3	-	(6, 1)
FCM	Linear 4	-	(2, 1)

II. THE PROPOSED METHOD

Typically, the performance of the existing ensemble system improves with the optimization of sub-systems. The sub-systems are trained respectively and the purpose of training is to optimize each sub-system rather than the whole ensemble system, which may limit the optimal performance of the ensemble system. Therefore, an end-to-end Dual-Branch Network is proposed.

A. Network Structure

The overall structure of our Dual-Branch Network is shown in Fig. 1 and the detailed architecture of the single branch network can be found in Table I. The Dual-Branch Network consists of four modules: Speech Preprocessing, Feature Extraction Module (FEM), Forgery Classification Module (FCM) and Forgery Type Classification Module (FTCM).

1) Speech Preprocessing: The LFCC and the log power spectrum of the CQT are used as the inputs of two branches, because the LFCC and CQT are extensively adopted in SSD as the inputs of single feature systems and both features have achieved better performance than other acoustic features on the unknown forgery types of attacks [16], [17], [18], [19], [20], [21], [22], [23], [24], [25].

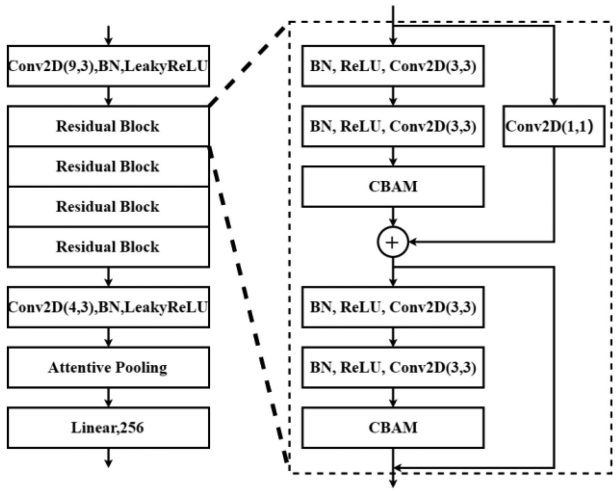

Fig. 2. Structure of the feature extraction module.

2) Feature Extraction Module (FEM): Each branch of the Dual-branch Network has an independent FEM which is used to extract features from the corresponding input. As shown in Fig. 2, the FEM is adapted from the ResNet18 [16] in which the CBAM is added into the Residual Block. The CBAM takes into account the significance of features on different channels and positions. It includes two tandem modules as channel attention module and spatial attention module, which has been experimentally found to improve the representation power of learned features [33].

3) Forgery Classification Module (FCM): The FCM aims to distinguish synthetic speech from real speech by using the extracted features of two branches.

4) Forgery Type Classification Module (FTCM): The FTCM aims to detect the forgery type of the fake speech, which is an additional task. Here, the Gradient Reversal Layer (GRL) [34] is placed in the front of the FTCM. The sign of the gradient is reversed when the gradient is backpropagated through the GRL. The training objectives of the FEM and FTCM are opposite, which forms the adversarial training to help the FEM learn common forgery features from different forgery types of speech. If the feature learned by the FEM only contains forgery information owned by some particular forgery types of speech, the FCM may be unable to distinguish the authenticity of unknown spoofing attacks. Therefore, the FTCM can benefit the FCM by helping the FEM learn common forgery features from different forgery types of speech.

B. Training Strategy

The objective of our Dual-Branch Network is to learn the discriminative features which can be used for the forgery classification task. In addition, the network should pay attention to common forgery features of different forgery types of speech to resist unknown synthetic attacks. The parameters of the network can be denoted as $\theta = \{\theta_{fe}, \theta_{fc}, \theta_{ftc}\}$, where $\theta_{fe}, \theta_{fc}, \theta_{ftc}$ are the parameters of the FEMs, FCM and FTCM, respectively. The whole network is then trained with two-part loss, namely forgery classification loss L_{fc} and forgery type classification loss L_{ftc},

as follows:

$$L\left(\theta_{fe},\theta_{fc},\theta_{ftc}\right)=L_{fc}\left(\theta_{fe},\theta_{fc}\right)-\lambda L_{ftc}\left(\theta_{fe},\theta_{ftc}\right) \quad (1)$$

where λ is a hyperparameter. The GRL takes the gradient from the subsequent layer, multiplies it by $-\lambda$ and passes it to the preceding layer during the loss backpropagation. The parameters of the network, $\hat{\theta}=\{\hat{\theta}_{fe},\hat{\theta}_{fc},\hat{\theta}_{ftc}\}$, are updated as:

$$\left(\hat{\theta}_{fe},\hat{\theta}_{fc}\right)=arg\min_{\theta_{fe},\theta_{fc}} L\left(\theta_{fe},\theta_{fc},\hat{\theta}_{ftc}\right) \quad (2)$$

$$\left(\hat{\theta}_{ftc}\right)=arg\max_{\theta_{ftc}} L\left(\hat{\theta}_{fe},\hat{\theta}_{fc},\theta_{ftc}\right) \quad (3)$$

1) Forgery Type Classification Loss L_{ftc}: The features extracted by two branches have great difference because LFCC and CQT focus on different speech information. Thus, two forgery type classifiers are designed for different features. The extracted forgery speech features are fed into forgery type classifiers. Then, the L_{ftc} is obtained by:

$$L_{ftc}=-\frac{1}{N}\times\left(\sum_{i=0}^{N-1}\sum_{k=0}^{K-1}y_{i,k}\log P_{i,k}^{lfcc}+\sum_{i=0}^{N-1}\sum_{k=0}^{K-1}y_{i,k}\log P_{i,k}^{cqt}\right) \quad (4)$$

where N is the number of fake samples in a mini-batch, K is the number of forgery types. $P_{i,k}$ means the prediction that the i-th sample is predicted to be the k-th forgery type and $P_{i,k}\in[0,1]$, the forgery type label $y_{i,k}\in\{0,1\}$, $y_{i,k}=1$ when the i-th sample belongs to the k-th forgery type. The objectives of two forgery type classifiers are to minimize the L_{ftc}, which is opposite to the objective of FEM. The adversarial training between the FEM and forgery type classifiers can help the FEM extract the common features from different forgery types of speech.

2) Forgery Classification Loss L_{fc}: Four fusion strategies of forgery classification loss are compared to obtain an optimal one. The forgery classification loss L_{fc} is obtained by logit fusion (5), prediction fusion (6), loss fusion (7) and alternating loss (8), as follows:

$$L_{fc}^{logits}=-\frac{1}{N}\sum_{i=0}^{N-1}\sum_{c=0}^{C-1}y_{i,c}\log\left(\frac{\exp\left(S_{i,c}^{lfcc}+S_{i,c}^{cqt}\right)}{\sum_{j=0}^{C-1}\exp\left(S_{i,j}^{lfcc}+S_{i,j}^{cqt}\right)}\right) \quad (5)$$

$$L_{fc}^{pred}=-\frac{1}{N}\sum_{i=0}^{N-1}\sum_{c=0}^{C-1}y_{i,c}\log\left(\frac{P_{i,c}^{lfcc}+P_{i,c}^{cqt}}{2}\right) \quad (6)$$

$$L_{fc}^{loss}=-\frac{1}{N}\times\left(\sum_{i=0}^{N-1}\sum_{c=0}^{C-1}y_{i,c}\log P_{i,c}^{lfcc}+\sum_{i=0}^{N-1}\sum_{c=0}^{C-1}y_{i,c}\log P_{i,c}^{cqt}\right) \quad (7)$$

$$L_{fc}^{alter}=\begin{cases}-\frac{1}{N}\sum_{i=0}^{N-1}\sum_{c=0}^{C-1}y_{i,c}\log P_{i,c}^{lfcc} & \text{if } epoch\%2=0\\ -\frac{1}{N}\sum_{i=0}^{N-1}\sum_{c=0}^{C-1}y_{i,c}\log P_{i,c}^{cqt} & \text{if } epoch\%2=1\end{cases} \quad (8)$$

TABLE II
PERFORMANCE COMPARISON OF THE DUAL-BRANCH NETWORK TRAINED WITH FORGERY CLASSIFICATION LOSS BASED ON FOUR FUSION STRATEGIES

Fusion Strategies	Development Set		Evaluation Set	
	EER (%)	t-DCF	EER (%)	t-DCF
Logits	0.11	0.003	1.93	0.049
Predictions	0.07	0.002	1.25	0.033
Loss	0.03	0.001	0.80	**0.021**
Alternating	0.03	0.001	**0.72**	0.022

where N is the number of samples in a mini-batch, C is the number of samples categories, $S_{i,c}\in(-\infty,+\infty)$ means the logits that the i-th sample is predicted to be the c-th category, $P_{i,c}$ is the normalization of $S_{i,c}$, $y_{i,c}\in\{0,1\}$ denotes the sample category label, $y_{i,c}=1$ when the i-th sample belongs to the c-th category. Same as the objective of the FEM, the forgery classifiers aim to minimize the L_{fc}. The joint training between the FEM and forgery classifiers can help the network learn the discriminative features.

III. EXPERIMENTS

A. Dataset and Evaluation Metrics

All experiments are implemented on the ASVspoof 2019 logical access (LA) dataset. Equal error rate (EER) and tandem detection cost function (t-DCF) [35] are utilized as the evaluation metrics. While EER only evaluates the performance of the spoofing detection system, t-DCF assesses the reliability of the spoofing detection system deployed in tandem with an ASV system.

B. Training Details

Two acoustic features are used in the experiments: LFCC and the log power spectrum of CQT. We extract the 60-dimensional LFCCs from the speech with the official baseline [5]. The frame length is 20 ms and the hop length is 10 ms. The CQT is extracted with 32 ms hop length and Hanning window, and the number of frequency bins are set to 100 similar to that in [20]. To form batches, each feature is padded or cropped to 750 frames. Thus, the dimensions of LFCC and CQT are 60×750 and 100×750, respectively. In addition, the number of forgery types K in the FTCM is set to 6 because the training set consists of 6 attack types (A01-A06) and the FTCM is removed during the testing phase. The training set contains fake samples and real samples, thus the number of sample categories C is set to 2.

We implement the Dual-Branch Network with PyTorch and use Adam optimizer with learning rate 1e-4 to update the weights. The batch size is set to 32. We train the network for 100 epochs and the network with the lowest validation EER is selected for evaluation. All experimental results are generated by a single Nvidia RTX 3090 GPU.

C. Ablation Study

Firstly, different fusion strategies are considered to guide the training of our proposed end-to-end network to obtain an optimal fusion strategy. Table II shows the performance of the Dual-Branch Networks trained with four fusion strategies shown in (5)–(8) on the development and evaluation set. These Dual-Branch Networks perform well on known synthetic speech attacks. For the evaluation set, the Dual-Branch Networks trained

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED SYSTEM TO THE ENSEMBLE SYSTEM FUSING TWO SINGLE FEATURE SYSTEMS

Systems	Development set		Evaluation set	
	EER (%)	t-DCF	EER (%)	t-DCF
LFCC + ResNet18	0.07	0.002	2.93	0.063
CQT + ResNet18	0.39	0.013	2.67	0.081
Ensemble system (Equal-weights)	0.11	0.001	1.86	0.052
Ensemble system (Learnable-weights)	0.07	0.001	1.78	0.047
Dual-Branch Network (Proposed)	**0.03**	**0.001**	**0.80**	**0.021**

TABLE IV
EER (%) PERFORMANCE COMPARISON OF THE PROPOSED SYSTEM AND THE ENSEMBLE SYSTEMS USING EQUAL-WEIGHTS FUSION AND LEARNABLE-WEIGHT FUSION ON INDIVIDUAL ATTACKS OF THE EVALUATION SET

Attacks	Systems		
	Ensemble system (Equal-weights)	Ensemble system (Learnable-weights)	**Dual-Branch Network**
A07	0.06	0.06	**0.05**
A08	1.19	0.58	**0.26**
A09	**0.00**	**0.00**	**0.00**
A10	**0.22**	0.26	0.26
A11	0.12	**0.09**	0.13
A12	**0.09**	0.12	0.17
A13	**0.12**	0.16	0.18
A14	0.17	0.22	**0.10**
A15	**0.18**	0.22	**0.18**
A16	**0.09**	**0.09**	0.10
A17	5.63	5.55	**2.48**
A18	1.46	0.99	**0.40**
A19	0.38	0.26	**0.17**
Total	1.86	1.78	**0.80**

with forgery classification loss based on loss fusion and alternating loss perform similarly, and achieve the best performance in term of t-DCF score and EER, respectively. In this work, we choose the Dual-Branch Network with the lowest t-DCF which is trained with the forgery classification loss based on the loss fusion strategy, because the t-DCF is the primary metric of the ASVspoof 2019.

Next, to demonstrate the importance of our end-to-end system, we compare it with two baselines, where two branches trained separately and the output scores of two branches are fused with equal-weights or learnable-weights. As shown in Table III, the ensemble system using learnable-weights performs better than that using equal-weights, achieving an EER 1.78 and a t-DCF 0.047, while our proposed system achieves an EER 0.80 and a t-DCF 0.0215. It is clear that the end-to-end system is useful for further improving the performance of ensemble systems. Table IV also shows the EER for each unknown attack type in the evaluation set. In this set, the A08, A17 and A18 attack types are more difficult to be distinguished than other attack types. The proposed system performance improves substantially on these challenging attack types compared with other ensemble systems.

Finally, to show the effectiveness of CBAM and FTCM, we conduct ablation experiments. Table V presents the experimental results. The experimental results show that CBAM and FTCM

TABLE V
RESULTS OF ABLATION EXPERIMENTS FOR CBAM AND FTCM ON THE EVALUATION SET

Module settings		Performance	
CBAM	FTCM	EER (%)	t-DCF
		1.91	0.044
√		0.92	0.024
	√	1.14	0.029
√	√	**0.80**	**0.021**

TABLE VI
PERFORMANCE COMPARISON OF THE PROPOSED SYSTEM TO STATE-OF-THE-ART SYSTEMS ON THE EVALUATION SET

Systems		Performance	
Input Feature	Network	EER (%)	t-DCF
Spec, LFCC	DenseNet [24]	1.98	0.047
LFCC	LCNN-LSTM-sum [19]	1.92	0.052
Spec, LFCC, CQT	SE-Res2Net50 [22]	1.89	0.045
LFCC, CQT, FFT	LCNN [18]	1.84	0.051
LFB	ResNet18 [14]	1.81	0.052
CQT	MCG-Res2Net50 [21]	1.78	0.052
Waveform	Raw PC-DARTS [32]	1.77	0.051
Waveform	Res-TSSDNet [26]	1.64	0.048
LFCC, STFT-gram	Capsule [23]	1.07	0.033
LFCC	OCT [27]	1.06	0.034
Waveform	RawGAT-ST [28]	1.06	0.033
SpecL, LFCC, ARS	scDenseNet [25]	0.98	0.032
LFCC, CQT	**Ensemble system (Learnable-weights)**	1.78	0.047
LFCC, CQT	**Dual-Branch Network**	**0.80**	**0.021**

are useful for the system. The combination of the CBAM and FTCM achieves the best performance.

D. Comparison With the SOTA Systems

To demonstrate the superiority of the proposed system, we compare our system with other SOTA systems. As shown in Table VI, the SOTA ensemble system scDenseNet [25] combined with SpecL, LFCC and ARS achieves an EER score of 0.98 and a t-DCF score of 0.032, while our proposed Dual-Branch Network combined with LFCC and CQT achieves an EER score of 0.80 and a t-DCF score of 0.021, representing improvement of 18.36% and 34.37%, respectively. The experimental results show that our proposed system has better generalization performance for unknown forgery types of synthetic speech attacks than the SOTA systems.

IV. CONCLUSION

In this letter, we present an end-to-end Dual-Branch Network for SSD. Moreover, the proposed network can not only detect synthetic speech but also pay attention to the common features of different forgery types of synthetic speech by using multi-task learning. Experiments on the ASVspoof 2019 LA show that our system has the better generalization performance for unknown forgery types of synthetic attacks than the SOTA systems. In the future, we will focus on improving the robustness of SSD systems which can detect the synthetic speech that have undergone compression or post-processing.

REFERENCES

[1] K. Delac and M. Grgic, "A survey of biometric recognition methods," in *Proc. Elmar 46th Int. Symp. Electron.*, 2004, pp. 184–193.

[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun.*, vol. 66, pp. 130–153, 2015.

[3] Z. Wu et al., "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2037–2041.

[4] T. Kinnunen et al., "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 2–6.

[5] M. Todisco et al., "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 1008–1012.

[6] J. Yamagishi et al., "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," in *Proc. ASVspoof Workshop*, 2021, pp. 47–54.

[7] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2087–2091.

[8] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Comput. Speech Lang.*, vol. 45, pp. 516–535, 2017.

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

[10] M. Pal, D. Paul, and G. Saha, "Synthetic speech detection using fundamental frequency variation and spectral features," *Comput. Speech Lang.*, vol. 48, pp. 31–50, 2018.

[11] J. Yang, R. K. Das, and N. Zhou, "Extraction of octave spectra information for spoofing attack detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2373–2384, Dec. 2019.

[12] J. Yang, R. K. Das, and H. Li, "Significance of subband features for synthetic speech detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 2160–2170, 2020.

[13] J. Yang, H. Wang, R. K. Das, and Y. Qian, "Modified magnitude-phase spectrum information for spoofing detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1065–1078, 2021.

[14] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of audio deepfake detection," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2020, pp. 132–137.

[15] J. Monteiro, J. Alam, and T. H. Falk, "Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers," *Comput. Speech Lang.*, vol. 63, 2020, Art. no. 101096.

[16] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Process. Lett.*, vol. 28, pp. 937–941, 2021.

[17] Y. Zhang, G. Zhu, F. Jiang, and Z. Duan, "An empirical study on channel effects for synthetic voice spoofing countermeasure systems," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 4309–4313.

[18] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC antispoofing systems for the ASVspoof2019 challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 1033–1037.

[19] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 4259–4263.

[20] I.-Y. Kwak et al., "ResMax: Detecting voice spoofing attacks with residual network and max feature map," in *Proc. IEEE 25th Int. Conf. Pattern Recognit.*, 2021, pp. 4837–4844.

[21] X. Li, X. Wu, H. Lu, X. Liu, and H. Meng, "Channel-wise gated Res2Net: Towards robust detection of synthetic speech attacks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 4314–4318.

[22] X. Li et al., "Replay and synthetic speech detection with Res2Net architecture," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6354–6358.

[23] A. Luo, E. Li, Y. Liu, X. Kang, and Z. J. Wang, "A capsule network based approach for detection of audio spoofing attacks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6359–6363.

[24] Z. Wang, S. Cui, X. Kang, W. Sun, and Z. Li, "Densely connected convolutional network for audio spoofing detection," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 1352–1360.

[25] S. Cui, B. Huang, J. Huang, and X. Kang, "Synthetic speech detection based on local autoregression and variance statistics," *IEEE Signal Process. Lett.*, vol. 29, pp. 1462–1466, 2022.

[26] G. Hua, A. B. J. Teoh, and H. Zhang, "Towards end-to-end synthetic speech detection," *IEEE Signal Process. Lett.*, vol. 28, pp. 1265–1269, 2021.

[27] C. Li, F. Yang, and J. Yang, "The role of long-term dependency in synthetic speech detection," *IEEE Signal Process. Lett.*, vol. 29, pp. 1142–1146, 2022.

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

[29] B. Chen, W. Tan, Y. Wang, and G. Zhao, "Distinguishing between natural and GAN-generated face images by combining global and local features," *Chin. J. Electron.*, vol. 31, no. 1, pp. 59–67, 2022.

[30] J. Li, M. Sun, and X. Zhang, "Multi-task learning of deep neural networks for joint automatic speaker verification and spoofing detection," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2019, pp. 1517–1522.

[31] G. Suthokumar, V. Sethu, K. Sriskandaraja, and E. Ambikairajah, "Adversarial multi-task learning for speaker normalization in replay detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6609–6613.

[32] W. Ge, J. Patino, M. Todisco, and N. Evans, "Raw differentiable architecture search for speech deepfake and spoofing detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 4319–4323.

[33] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[34] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[35] T. Kinnunen et al., "Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2195–2210, 2020.