Detection of Speech Spoofing Based on Dense Convolutional Network

Yong Wang[1], Xiaozong Chen[1(✉)], Yifang Chen[1], and Shunsi Zhang[2]

[1] Guangdong Polytechnic Normal University, Guangzhou 510000, China
xiaozong_chan@163.com
[2] Guangzhou Quwan Network Technology Co., Ltd., Guangzhou 510627, China

Abstract. In recent years, the rapid development of voice synthesis technologies has led to an increasing concern about the abuse of fake human voices for malicious purposes, such as deepfake audio, spam calls and social engineering attacks. This paper proposes a novel deep learning-based model to effectively identify counterfeit human voices generated by various voice synthesis algorithms. The proposed model employs a combination of Dense-Style Network to capture both spectral and temporal features of human speech. The model is extensively evaluated on ASVspoof 2019 datasets. The experimental results indicate that our model achieves competitive performance compared to existing methods and has a certain degree of anti-compression ability. In addition, anti-compression research was conducted to investigate the recognition performance of the model in response to compressed speech. Our findings pave the way for further research in combating against the misuse of artificially generated human voices and sound authenticity verification in general.

Keywords: Dense-Style Network · ASVspoof 2019 · anti-compression

1 Introduction

Speech spoofing [1] refers to the use of specific methods to alter the sound of one's voice, leading individuals or systems to misjudge the speaker's identity and achieve a deceptive effect. Speech spoofing can be categorized into two types: mechanical spoofing and electronic spoofing [2]. The first type involves physical means such as manipulating organs of the human body, covering the mouth, or pinching the nose to alter the voice. This type is known as mechanical speech spoofing [3]. Although this form of spoofing can achieve certain results, its effectiveness is limited by the language skills of the speaker.

Electronic spoofing refers to the manipulation of a speaker's original voice using electronic devices or voice processing software. In comparison to mechanical spoofing, electronic spoofing utilizes electronic devices and built-in algorithms to modify the temporal or spectral characteristics of the voice, resulting in a more natural-sounding spoofing. Through various algorithms, digital tools are used to process speech with the aim of disguising the voice, ensuring that the altered voice does not raise suspicion.

J. Cai et al. (Eds.): SPNCE 2023, LNICST 525, pp. 258–267, 2025.
https://doi.org/10.1007/978-3-031-73699-5_18

Compared to mechanical speech spoofing, electronic spoofing poses greater risks, thus in this study we focus on researching the detection of electronic spoofing, collectively referred to as speech spoofing. Currently, there are several typical forms of speech spoofing, including Voice Conversion (VC) [4, 5], Speech Synthesis (SS) [6, 7], Voice Redubbing [8, 9], and Voice Transformation (VT) [10, 11].

Existing research has shown that speech spoofing poses a threat to Automatic Speaker Verification (ASV) systems [12–14]. Attackers can utilize speech spoofing techniques to forge others' voiceprints or speech samples, thereby deceiving identity authentication through ASV systems. Currently, research on ASV systems primarily focuses on two aspects: feature parameter extraction and pattern recognition. Feature parameter extraction involves analyzing and processing the speech signal to remove irrelevant information for speaker identification, obtaining essential characteristics representing individuals within the speech signal. Pattern recognition, on the other hand, classifies the extracted feature vectors to determine whether the current speaker is a known individual. Pattern recognition is one of the core technologies in ASV systems and includes designing and training classifiers. Commonly used classifiers include Gaussian Mixture Models (GMM), Support Vector Machines (SVM), and Deep Neural Networks (DNN). As a gateway safeguarding people's privacy, the security of ASV systems is crucial for ensuring privacy protection.

There are various methods related to speech spoofing. In terms of detection models, Korsh et al. [17] designed a DNN-deep neural network structure that enables the learning of speech features and classification models together. Dinkel et al. [18] proposed a deep model for spoofing detection based on raw waveforms, eliminating the need for any preprocessing or post-processing of data, making training and evaluation a streamlined process that consumes less time compared to other neural network-based methods. Liu et al. [19] introduced an end-to-end anti-spoofing model composed entirely of one-dimensional convolutional neural networks, specifically for detecting speech spoofing under noisy conditions. Huang et al. [20] proposed a novel model based on segment-wise linear filterbank features, combining the advantages of CNN and RNN, which outperforms traditional GMM models and exhibits better resistance to overfitting. Gong et al. [21] presented a new neural network-based model for replay attack detection, utilizing both spectral and spatial information from multi-channel audio, significantly improving the performance of replay attack detection. In terms of speech features, Chen et al. [22] discovered through experiments that the number of minimum value MDCT coefficients in fake audio is fewer than in genuine audio. Therefore, Renza [23] and Ghobadi [24] applied watermarks on audio using MDCT characteristics to aid in identifying the forged sections within the audio. Sathya et al. [25] used Cosine-Normalized Phase based Cepstral Coefficient (CNPCC) features to enhance the detection of deceptive speech. Das et al. [26] proposed an improved version of γ-frequency cepstral coefficients to enhance the performance of spoofing detection. Alzantot et al. [27] designed a network model based on deep residual networks that combines three different speech features. Zhan et al. [28] introduced a fragment-based approach for detecting deceptive speech, combining a constant Q cepstral coefficient-based method with a constant Q cepstral coefficient-based speech segment extraction method to improve the robustness of embedded systems based on speech authentication.

In terms of classifiers, Tian et al. [29] proposed a time-domain CNN-based classifier and investigated spoofing detection based on unit selection. Zhe et al. [30] proposed the use of an ensemble classifier set, including multiple Gaussian Mixture Model (GMM)-based classifiers, as well as two new GMM average super-vector Gradient Boosting Decision Tree (GSV-GBDT) and GSV-Random Forest (GSV-RF) classifiers. La et al. [31] employed a linear kernel SVM classifier to classify the extracted i-vector advanced feature representations. Cui et al. [32] introduced a backend classifier based on dense convolution and short connections, combining popular features such as constant Q cepstral coefficients with Linear Frequency Cepstral Coefficients (LFCC), thereby improving fusion performance. Sun et al. [33] constructed a novel joint voice detector based on gamma-tone frequency cepstral coefficient features, combining self-attention residual networks and light gradient boosting machines. This classifier avoids overfitting, has low computational complexity, replaces traditional fully connected layer classifiers, and effectively discriminates between genuine and deceptive sounds.

In this paper, we don't use any front-end feature. The raw voice data been processed will feed to the network directly. For back-end classification, we refer to the excellent DenseNet and design a DNN with a DenseNet-style architecture.

2 Proposed Model

This paper aims to address security issues caused by speech spoofing. The goal is to establish a model for detecting speech spoofing and classify whether the speech is original or manipulated. It is generally believed that deeper networks perform better, but in the case of speech spoofing detection, spoofed speech often only undergoes minor modifications on genuine speech. If the network is too deep, the features of spoofed speech may be lost [34]. Therefore, we should avoid using excessively deep networks.

2.1 Model Structure

We constructed a dense convolutional neural network. The input of this model is all processed 6 s speech data. The network structure is shown in Fig. 1.

The network has a total of 49 layers, with 971,890 model parameters. The input tensor size of the speech data is 1×96000 (which means channels $\times$ sample rates). The DenseNet-style blocks are used, as shown in Fig. 2.

Each block consists of 10 layers, including a 1×3 convolutional layer, batch normalization layer, and ReLu activation layer, each of which is repeated twice for a total of 6 layers. This is followed by a 1×3 convolutional layer, batch normalization layer, and finally a 1×1 convolutional layer for the input data of the block. The outputs of these two layers are concatenated together and passed through the final ReLu activation layer. The specific network parameters in the entire network structure are shown in Table 1.

2.2 Model Training Strategy

The raw speech data of ASVspoof 2019 vary in length, and preprocessing of the dataset is necessary to ensure a uniform tensor input size for the model. Additionally, the ratio

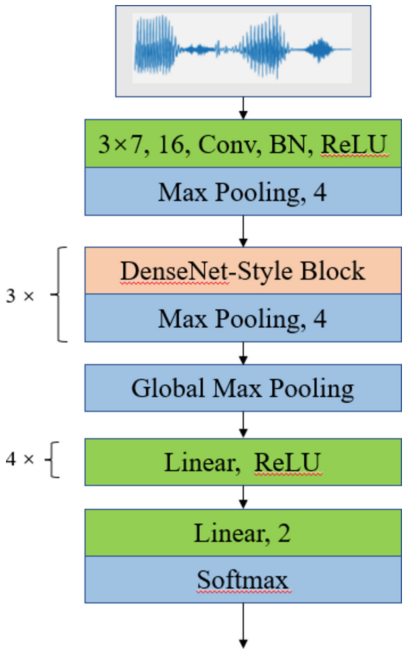

Fig. 1. Model Struct

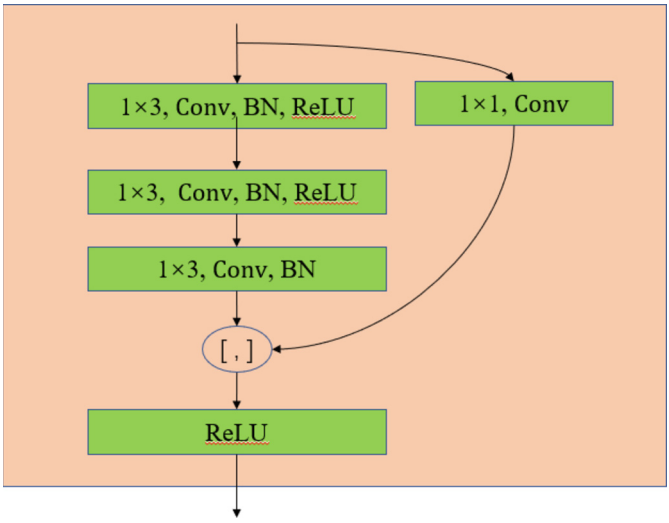

Fig. 2. DenseNet-Style Block

of genuine speech data to spoofed speech data in the original dataset is unbalanced, and measures should be taken to address this imbalance. The model training includes the following strategies.

Table 1. Detailed parameters of each layer of the network model.

Layer	Number of layers	Input Size	Output Size
3×7 Conv, BN, ReLU	3	1 × 96000	16 × 96000
Max Pooling 4	1	16 × 96000	16 × 24000
DenseNet-Style Block	10	16 × 24000	64 × 24000
Max Pooling 4	1	64 × 24000	64 × 6000
DenseNet-Style Block	10	64 × 6000	256 × 6000
Max Pooling 4	1	256 × 6000	256 × 1500
DenseNet-Style Block	10	256 × 1500	512 × 1500
Max Pooling 4	1	512 × 1500	512 × 375
Global Max Pooling	1	512 × 375	512 × 1
flatten	1	512 × 1	512
Linear, ReLU	2	512	256
Linear, ReLU	2	256	128
Linear, ReLU	2	128	64
Linear, ReLU	2	64	32
Linear	1	32	2
Softmax	1	2	2

Data Preprocessing. In this paper, we used the ASVspoof 2019 dataset. Since the duration of speech in the dataset varies, we adopted the method proposed in [35], which involves truncating or duplicating speech data to ensure that all speech data have a duration of 6 s, with a default sampling rate of 16 kHz. The input precision of the feature vectors is 9.6×10^4, and the batch size is set to 32.

Weighted Cross Entropy. WCE was used to deal with the imbalance of training data caused by the different number of true and false voices. Assuming we have two categories (Category 1 and Category 2), the probability of the predicted categories by the model is p_1 and p_2. The true category probabilities are y_1 and y_2 ($y_1 + y_2 = 1$), the weighted Cross entropy loss can be expressed as:

$$\mathrm{WCE} = [w_1 \times y_1 \times \log(p_1) + w_2 \times y_2 \times \log(p_2)] \tag{1}$$

Among them, w1 and w_2 represents the weights of category 1 and category 2, respectively, y_1 and y_2 represents the true labels (0 or 1) of category 1 and category 2, respectively, i.e. the first and second elements of the one hot encoding vector, p_1 and p_2 represents the probabilities of category 1 and category 2 predicted by the model, which are transformed by the SoftMax function, Log represents the Natural logarithm

The Adam optimizer with a Learning rate of 0.95 is selected for model training, and the model parameters with the lowest Equal Error Rate (EER) are selected from 100 epochs in the validation set to test the test set each time.

3 Result

3.1 Corpus

The speech data in the LA subset of the ASVspoof 2019 database is entirely derived from the VCTK corpus. The VCTK corpus consists of authentic speech from 107 speakers, including 46 males and 61 females. All the authentic speech samples were recorded using the same recording configuration without any channel or background noise interference. The authentic speech samples in the LA subset are directly selected from the VCTK corpus. The spoofed speech samples in the dataset are generated by applying various speech synthesis and voice conversion techniques to these authentic speech samples. The sampling rate for all the speech data is 16 kHz.

The LA dataset is divided into three subsets: the training set, the development set, and the evaluation set. Both the training set and the development set include spoofed speech samples generated using six identical speech synthesis and voice conversion techniques. These six techniques represent known attack types and can be used to train and adjust the synthetic speech detection system. In contrast, the evaluation set includes spoofed speech samples generated using two of the aforementioned known attacks and eleven additional speech synthesis and voice conversion techniques that differ from the six known attacks. These eleven techniques represent unknown attack types that the system may encounter.

3.2 Equal Error Rate

The Equal Error Rate (EER) can be used to evaluate the performance of an authentication speaker recognition system. False acceptance rate (FAR) represents the probability of incorrectly accepting a non-matching sample as a match without proper authorization. False rejection rate (FRR) represents the probability of incorrectly rejecting a matching sample as a non-match. Calculate the FAR and FRR for a range of possible thresholds and plot the Detection Error Tradeoff (DET) curve, which shows the relationship between them. EER corresponds to the point on the DET curve where the false acceptance rate equals the false rejection rate in terms of error rate.

EER is the threshold at which the false acceptance rate equals the false rejection rate. In other words, the EER is the error rate achieved when the false acceptance rate and false rejection rate are equal. The EER provides an intuitive and comparable measure of performance for authentication or speaker recognition systems in balancing error rates. A lower EER corresponds to better performance.

3.3 Main Result

After preprocessing the data, it was fed into the network. After 100 epochs of training, the model's loss curve, as shown in Fig. 3, was obtained. From the loss curve, it can be observed that the loss gradually decreases smoothly with an increasing number of training epochs. The decreasing slope indicates that the training effectiveness of the model is improving gradually. This smooth decreasing trend suggests that as the number

of training epochs increases, the model reduces prediction errors and captures the patterns and features within the data. The decreasing slope also suggests that the model is approaching convergence. We achieved the optimal results, as shown in Table 2 finally.

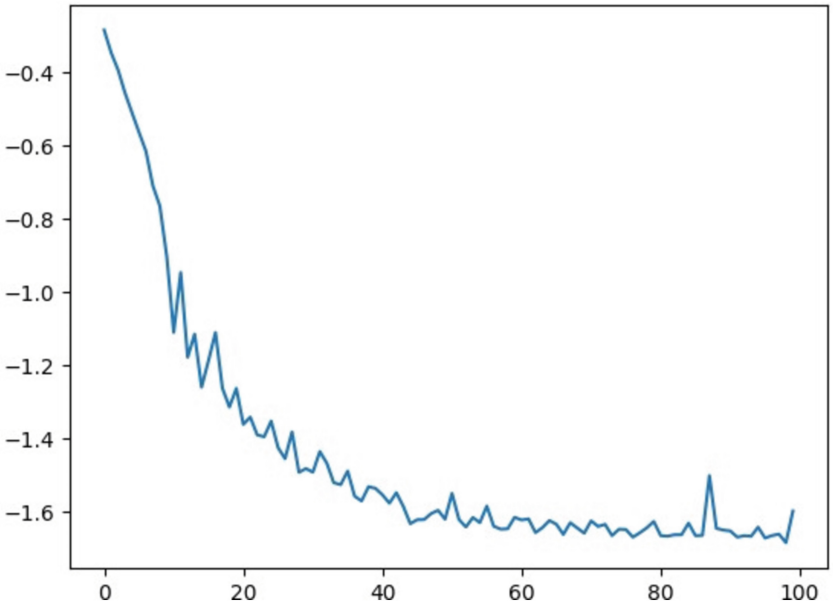

Fig. 3. Loss (Logarithm Base 10 Scale) Curve of 100 epochs

From the table, it can be seen that the EER of our model's results are significantly better than the baseline, as well as outperforming results [37–39], and approaching the results of [40]. However, our model's parameters are much smaller than theirs. Even though the model parameters of [35] are fewer than ours, our EER is lower than theirs.

Table 2. EER (%) between the proposed model and state-of-art method

Method	Params	Dev	Eval
Baseline CQCC+GMM [36]	-	2.71	8.09
8 Features+MLP [37]	-	0	4.13
Spec+CQCC+ResNet+SE [38]	5.80 M	0	6.7
Spec+VGG+SincNet [39]	>4.32 M	0	8.01
3 Features+CNN [40]	30.6 M	0	1.86
CQT+Res2Net+SE [35]	0.92 M	0.43	2.5
Dense-Style Net	0.97 M	0.71	1.98

3.4 Anti-compression Research

Compressing audio is widely used in everyday internet connections, making it necessary for models to have good anti-compression performance. In this study, we compressed the dataset with 16:1 MP3 compression and repeated the experimental steps mentioned above. The resulting EER results are shown in Table 3.

Table 3. EER (%) between the uncompressed and compressed models.

	Dev	Eval
No Compressed	0.71	1.98
Compressed	1.66	3.34

4 Conclusion

In this paper, we addressed the problem of spoof speech detection using deep learning. Our research aimed to lower the EER while reducing the computational complexity.We made several significant contributions in this study. Firstly, we proposed a Dense-Style convolutional neural network architecture that incorporates attention mechanisms to enhance feature selection and improve classification performance. Furthermore, we conducted Anti-Compression studies to analyze the impact of compression in our model.

In conclusion, the results indicate that our model has a lower EER compared to the existing state-of-the-art model. Moreover, our model performs better even though it has fewer parameters and lower complexity than the models that outperform ours. In terms of compression resistance, the research findings demonstrate that our model exhibits a certain level of resistance to 16:1 MP3 compression.

Acknowledgement. I would like to express my heartfelt gratitude to all those who have supported and assisted me throughout the process of completing this research. First and foremost, I would like to thank my advisor, Professor Yong Wang, for providing me with invaluable guidance and advice throughout the entire research process. Additionally, I would like to extend my thanks to the resources and support provided by the university, which have given me the necessary conditions and environment to conduct this research. I sincerely appreciate each and every person who has contributed to my research work. Without your support and assistance, I would not have been able to accomplish this study. Thank you very much!

References

1. Perrot, P., Aversano, G., Chollet, G.: Voice disguise and automatic detection: review and perspectives. Prog. Nonlinear Speech Process., 101–117 (2007)
2. Lau, Y.W., Wagner, M., Tran, D.: Vulnerability of speaker verification to voice mimicking. In: Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004, pp. 145–148. IEEE (2004)

3. Zhang, C., Li, B., Chen, S., et al.: Acoustic analysis of whispery voice disguise in Mandarin Chinese. In: Proceedings of the Interspeech, pp. 1413–1416 (2018)
4. Stylianou, Y., Cappé, O., Moulines, E.: Continuous probabilistic transform for voice conversion. IEEE Trans. Speech Audio Process. **6**(2), 131–142 (1998)
5. Erro, D., Navas, E., Hernaez, I.: Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. IEEE Trans. Audio Speech Lang. Process. **21**(3), 556–566 (2012)
6. Tokuda, K., Nankaku, Y., Toda, T., et al.: Speech synthesis based on hidden markov models. Proc. IEEE **101**(5), 1234–1252 (2013)
7. Yamagishi, J., Kobayashi, T., Nakano, Y., et al.: Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Audio Speech Lang. Process. **17**(1), 66–83 (2009)
8. Shang, W, Stevenson, M.: Score normalization in playback attack detection. In: Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1678–1681. IEEE (2010)
9. Villalba, J., Lleida, E.: Detecting replay attacks from far-field recordings on speaker verification systems. In: Vielhauer, C., Dittmann, J., Drygajlo, A., Juul, N.C., Fairhurst, M.C. (eds.) Biometrics and ID Management. BioID 2011. Lecture Notes in Computer Science, vol. 6583, pp. 274–285. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19530-3_25
10. Perrot, P., Chollet, G.: The question of disguised voice. J. Acoust. Soc. Am. **123**(5), 3878 (2008)
11. Perrot, P., Aversano, G., Chollet, G.: Voice disguise and automatic detection: review and perspectives. Prog. Nonlinear Speech Process., 101–117 (2007)
12. Evans, N.W.D., Kinnunen, T., Yamagishi, J.: Spoofing and countermeasures for automatic speaker verification. In: Proceedings of the Interspeech, pp. 925–929 (2013)
13. Kinnunen, T., Wu, Z.Z., Lee, K.A., et al.: Vulnerability of speaker verification systems against voice conversion spoofing attacks: the case of telephone speech. In: Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4401–4404. IEEE (2012)
14. Leon, P.D., Pucher, M., Yamagishi, J., et al.: Evaluation of speaker verification security and detection of HMM-based synthetic speech. IEEE Trans. Audio Speech Lang. Process. **20**(8), 2280–2290 (2012)
15. Black, A.W., Zen, H., Tokuda, K.: Statistical parametric speech synthesis. In: Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP 2007, vol. 4, IV-1229–IV-1232. IEEE (2007)
16. Ma, Y., Ren, Z., Xu, S.: RW-Resnet: a novel speech anti-spoofing model using raw waveform. arXiv preprint arXiv:2108.05684 (2021)
17. Korshunov, P., Goncalves, A.R., Violato, R., et al.: On the use of convolutional neural networks for speech presentation attack detection. In: IEEE International Conference on Identity, pp. 1–8. IEEE (2018)
18. Dinkel, H., Chen, N., Qian, Y., et al.: End-to-end spoofing detection with raw waveform CLDNNS. In: Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4860–4864. IEEE (2017)
19. Liu, P., Zhang, Z., Yang, Y.: End-to-end spoofing speech detection and knowledge distillation under noisy conditions. In: Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE (2021)
20. Huang, L., Pun, C.M.: Audio replay spoof attack detection by joint segment-based linear filter bank feature extraction and attention-enhanced DenseNet-BiLSTM network. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 1813–1825 (2020)
21. Gong, Y., Yang, J., Poellabauer, C.: Detecting replay attacks using multi-channel audio: a neural network-based method. IEEE Sign. Process. Lett. **27**, 920–924 (2020)

22. Chen, B., Luo, W., Luo, D.: Identification of audio processing operations based on convolutional neural network. In: Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security, pp. 73–77 (2018)
23. Renza, D., Lemus, C.: Authenticity verification of audio signals based on fragile watermarking for audio forensics. Expert Syst. Appl. **91**, 211–222 (2018)
24. Ghobadi, A., Boroujerdizadeh, A., Yaribakht, A.H., et al.: Blind audio watermarking for tamper detection based on LSB. In: Proceedings of the 2013 15th International Conference on Advanced Communications Technology (ICACT), pp. 1077–1082. IEEE (2013)
25. Sathya, A., Swetha, J., Das, K.A., et al.: Robust features for spoofing detection. In: Proceedings of the 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2410–2414. IEEE (2016)
26. Das, K.A., George, K.K., Kumar, C.S., et al.: Modified gammatone frequency cepstral coefficients to improve spoofing detection. In: Proceedings of the 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 50–55. IEEE (2016)
27. Alzantot, M., Wang, Z., Srivastava, M.B.: Deep residual neural networks for audio spoofing detection. In: Proceedings of the Interspeech 2019, pp. 1078–1082 (2019)
28. Zhan, J., Pu, Z., Jiang, W., et al.: Detecting spoofed speeches via segment-based word CQCC and average ZCR for embedded systems. IEEE Trans. Comput. Aided Des. Integr. Circ. Syst. **41**(11), 3862–3873 (2022)
29. Tian, X., Xiao, X., Chng, E.S., et al.: Spoofing speech detection using temporal convolutional neural network. In: Proceedings of the 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 1–6. IEEE (2016)
30. Ji, Z., Li, Z.Y., Li, P., et al.: Ensemble learning for countermeasure of audio replay spoofing attack in ASVspoof2017. In: Proceedings of the Interspeech, pp. 87–91 (2017)
31. Lavrentyeva, G., Novoselov, S., Malykh, E., et al.: Audio replay attack detection with deep learning frameworks. In: Interspeech, pp. 82–86 (2017)
32. Cui, S., Huang, B., Huang, J., et al.: Synthetic speech detection based on local autoregression and variance statistics. IEEE Sig. Process. Lett. **29**, 1462–1466 (2022)
33. Sun, X., Fu, J., Wei, B., et al.: A self-attentional ResNet-LightGBM model for IoT-enabled voice liveness detection. IEEE Internet Things J. **10**(9), 8257–8270 (2022)
34. Hua, G., Teoh, A., Zhang, H.: Towards end-to-end synthetic speech detection. IEEE Sig. Process. Lett. **28**, 1265–1269 (2021)
35. Li, X., Li, N., Weng, C., Liu, X., Su, D., Yu, D., Meng, H.: Replay and synthetic speech detection with Res2Net architecture. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021) (2021)
36. Todisco, M., et al.: ASVspoof 2019: future horizons in spoofed and fake audio detection. In: Proceedings of the Interspeech, pp. 1008–1012 (2019)
37. Das, R.K., Yang, J., Li, H.: Long range acoustic features for spoofed speech detection. In: Proceedings of the Interspeech, pp. 1058–1062 (2019)
38. Lai, C.-I., Chen, N., Villalba, J., Dehak, N.: ASSERT: anti-spoofing with squeeze-excitation and residual networks. In: Proceedings of the Interspeech, vol. 2019, pp. 1013–1017 (2019)
39. Zeinali, H., Stafylakis, T., Athanasopoulou, G., Rohdin, J., Gkinis, I., Burget, L., Černocký, J.: Detecting spoofing attacks using VGG and SincNet: BUT-omilia submission to ASVspoof 2019 challenge. In: Proceedings of the Interspeech, pp. 1073–1077 (2019)
40. Lavrentyeva, G., Novoselov, S., Tseren, A., Volkova, M., Gorlanov, A., Kozlov, A.: STC antispoofing systems for the ASVspoof2019 challenge. In: Proceedings of the Interspeech, pp. 1033–1037 (2019)