2023 International Conference on Computer Applications Technology (CCAT) | 979-8-3503-0154-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/CCAT59108.2023.00026

Enhancing Synthesized Speech Detection with Dual Attention Using Features Fusion

Bo Wang
School of Information and Communication Engineering
Dalian University of Technology
Dalian, China
bowang@dlut.edu.cn

Yanyan Ma*
School of Information and Communication Engineering
Dalian University of Technology
Dalian, China
mayany@mail.dlut.edu.cn

Yeling Tang
School of Information and Communication Engineering
Dalian University of Technology
Dalian, China
tyl_@mail.dlut.edu.cn

Rui Wang
School of Information and Communication Engineering
Dalian University of Technology
Dalian, China
wangrui11291224@163.com

Maozhen Zhang
School of Information and Communication Engineering
Dalian University of Technology
Dalian, China
maozhenzhang@mail.dlut.edu.cn

***Abstract*—Automatic Speaker Verification (ASV) is a system based on speech recognition used for identity verification. However, with the continuous improvement of synthetic speech generation technology, the quality of generated speech is also getting higher and higher. This allows some malicious actors to use synthetic speech to deceive, which poses a serious threat to ASV. Therefore, we need to continuously improve the authenticity identification to cope with this challenge. Although many synthetic speech detection algorithms have been proposed, their generalization performance is still not ideal. This means that the performance of ASV systems may suffer in the face of previously unseen synthetic speech attacks. Therefore, researchers need to further explore new methods for detecting synthetic speech attacks and strengthen the robustness of ASV systems against synthetic speech attacks. In order to extract more abundant and reliable speech features, this paper proposes a dual-attention network, specifically combining the features extracted by Wav2vec with traditional speech features Logmel through self-attention, and inputting them into a Resnet network with convolutional block attention module (CBAM). The results have shown that our method has achieved a very competitive performance on the Asvspoof 2021 LA and DF datasets, with a *t-DCF* of 0.3008 on LA and an *EER* of 3.9% on DF, indicating good generalization performance.**

***Keywords*—Wav2vec, Logmel, synthetic speech, resnet, CBAM**

I. Introduction

Speech is the primary way of communication between people and the interaction between people and machines. It not only conveys the content we want to express, but also contains unique identity information for each individual. Currently, automatic speaker verification (ASV) technology has become a form of speaker biometric recognition. However, with the continuous development of spoofing attacks, the reliability of ASV systems is seriously threatened. Spoofing attacks include replay, text-to-speech (TTS), speech conversion (VC), impersonation, and other adversarial attacks. So ASV systems may have to deal with many types of attacks.

To address spoofing attacks, the speech community has held four ASVspoof challenges since 2015, focusing on people developing spoofing detection for reply, TTS and VC. Specifically, ASVspoof 2019 [1] includes all of the previous attacks, breaking them down into logical access (LA) and physical access (PA). LA is typically utilized as a means to counter TTS and VC attacks, whereas PA is more commonly employed to address reply attacks. Our main focus is on the LA mission. To get closer to the actual application scenario, the latest ASVspoof 2021 [2] competition focuses on detecting low-quality fake speech. The evaluation data of LA in ASVspoof 2019 has been transformed into two separate tasks in ASVspoof 2021: Logical Access (LA) and DeepFake (DF) tasks. The two tasks use different encoding modes. The LA task involves the transmission of speech through different types of telephones, while the DF task involves the compression and decompression of speech using various methods.

In recent years, in order to improve the reliability of ASV spoofing strategies, feature extraction and model design are generally optimized. Hand-crafted feature extraction is indeed very important and performs well in the field of speech, such as discrete wavelet transform [3, 4], mel-frequency cepstral coefficients, and other frequency domain features [5]. Hand-crafted features for deception detection include fundamental frequency, power spectrum, linear frequency cepstral coefficients (LFCC), mel frequency cepstral coefficients (MFCC), modified group delay (MGD), relative phase shift (RPS), constant-Q cepstral coefficients (CQCC), and many variations and combinations of them [6-9]. However, the hand-crafted features are strongly dependent on the corresponding features

979-8-3503-0154-0/23/$31.00 ©2023 IEEE
DOI 10.1109/CCAT59108.2023.00026

of known attacks, and the performance of detecting unknown attacks is poor. To solve this problem, some existing works [10, 11] operate directly on raw speech waveforms using different networks. For the back-end network, convolutional neural networks are becoming more and more popular in speech field because of their remarkable effect in image field. Alzantot [12] et al. proposed a detection scheme based on deep ResNet, and conducted fractional fusion for three different front-end features (MFCC, spectral graph, and CQCC). The detection scheme based on Res2Net proposed by Li [13] et al. introduced the squeeze-excitation module into the Res2Net network to further improve the detection performance. Subsequently, various attention mechanisms are used to improve the robustness of forged speech detection. Zheng [14] et al. proposed an attentional mechanism combining graph convolution with capsule to generate capsule nodes, and a graph adaptive attentional mechanism to study context information in different global graph convolution, so as to effectively improve next dynamic routing connection and final classification.

Although the previous work performed decently in spoofing detection speech, the generalization performance still has great room for improvement. Based on these works, it is proved that the self-supervised pre-training model [15, 16] has a significant effect on improving the generalization performance of synthetic speech detection. Since Wav2vec learns high-level semantic information of audio signals and traditional manual feature extraction is acoustic information, this paper explores whether the combination of the two can get more representative and richer feature representation, thus improving the performance of the model. In addition, in synthetic speech detection, the attention mechanism can make the model focus more on the features related to speech synthesis, such as acoustic features and prosodic features. So this paper adds attention mechanism to network model to improve detection performance.

Our main contribution through this work is to combine self-supervised features with traditional features for self-attention, input them into Resnet network with attention mechanism, and extract more representative synthetic speech detection features. The rest of this article is organized as follows. Section 2 describes the relevant work in this paper, Section 3 describes the overall framework of our proposed system, and Section 4 gives the experimental setup and results. Finally, we summarize the conclusions of this study in Section 5.

II. Related Work

A. Wav2vec 2.0

Self-supervised learning (SSL) [17] used to learn speech representations from unlabeled speech data has attracted wide attention in the past few years. In general, self-supervised presentation learning trains models with large amounts of unlabeled data and performs well with only small amounts of labeled data by learning that there are acoustic feature representations for downstream tasks. Among many self-supervised speech models, this study focuses on Wav2vec 2.0. Wav2vec 2.0 is a self-supervised pre-training model of multi-layer convolutional neural networks that takes audio signals as input and encodes them into feature representations that can be input to downstream tasks. Wav2vec 2.0 model is composed of convolutional feature encoder network and context network. Encoder $f : X \rightarrow Z$ is a series of convolutional operation, which changes input audio X into potential speech representation into $z_1, \ldots, z_T$. Potential features are then input into a context network composed of transformer models $g : Z \rightarrow C$ to get context representation $c_1, \ldots, c_T$.Transformer architecture follows BERT model, as shown in Fig. 1.

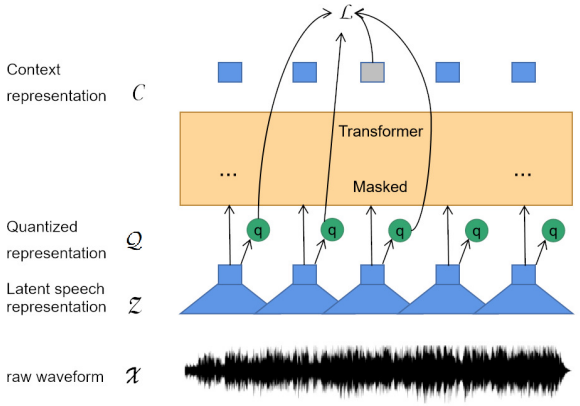

Fig. 1: Wav2vec 2.0 model framework

In the pre-training phase, latent representation $z_1, \ldots, z_T$ was quantified into representation $q_1, \ldots, q_T$. The potential representation $z_1, \ldots, z_T$ is masked and fed to the context network g to obtain the context representation $c_1, \ldots, c_T$. In order to the model to more accurately predict the acoustic feature q_n corresponding to the masked time step n given the corresponding context vector $c_1, \ldots, c_T$, we need to calculate a contrast loss. This loss function will compare the similarity between the target acoustic features of masked time step n and a set of interference items (negative sample) $q_{n_0}(n_0 \neq n)$ randomly sampled from other time steps, so that the model can learn a more accurate and robust acoustic representation. In this paper, the pre-training model Wav2vec 2.0 XLS-R is used for feature extraction.

B. Resnet

Convolutional neural network (CNN) has made breakthroughs in the field of forged speech detection, but when the network is deepened continuously, the problem of gradient disappearing or gradient explosion will appear. The ResNet was originally proposed to address the problem of degradation and gradient vanishing in deep neural networks. The core of the network is to use skip connections to reduce the training parameters of the deep network and to propagate the parameter updates to the lower layers faster during training. Today, ResNet has become one of the most widely used networks in the field of forged speech detection. Subramani [18] et al. used the EfficientCNN network combined with

the residual module to establish the RES-EfficientCNN model with fewer parameters and high detection accuracy. Lai [19] et al. proposed the ASSERT detection system based on SENet and ResNet, which integrated five different systems such as SENet, mean-standard deviation ResNet and extended ResNet. SENet can also be combined with network architectures such as ResNet for better detection results.

III. Proposed Method

We propose a dual attention network that combines self-supervised features with hand-crafted features to combine self-attention and input them into a Resnet network with an attention mechanism. For convenience, classification network is represented as ATT-Resnet. The specific block diagram is shown in Fig. 2.

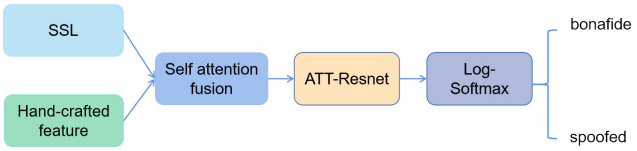

Fig. 2: Structure of the proposed system.

The following describes the self-attention fusion module and the ATT-Resnet module respectively.

A. Self attention fusion

Inspired by [20], this paper utilizes a front feature for detecting forged speech that combines self-attentional features from Wav2vec and manual features, resulting in a more comprehensive and enriched representation. Since Logmel uses log-scale Mel filter, the frequency resolution of low frequency region and high frequency region can be improved, so that the information in different frequency ranges can be represented evenly. In addition, when calculating Logmel spectrum, the short-time Fourier transform (STFT) is usually used to retain the time information. This can provide better temporal resolution than MFCC while preserving the frequency information in the audio signal. Therefore, using Logmel as a handcrafted feature combined with self-supervised features through self-attention, the framework is shown in Fig. 3. For convenience, Logmel is represented as Lm in the figure.

We use the form of self-attention in [21] to combine Logmel with self-supervised features for self-attention. First, f_{Logmel} and f_{SSL} are linearly mapped through six learnable matrices W^Q_{Logmel}, W^Q_{SSL}, W^K_{Logmel}, W^K_{SSL}, W^V_{Logmel}, W^V_{SSL} to get Q_{Logmel}, Q_{SSL}, K_{Logmel}, K_{SSL}, V_{Logmel}, V_{SSL}. The specific formula is shown in (1) - (6).

$$Q_{Logmel} = f_{Logmel}(S)W^Q_{Logmel} \tag{1}$$

$$Q_{SSL} = f_{SSL}(S)W^Q_{SSL} \tag{2}$$

$$K_{Logmel} = f_{Logmel}(S)W^K_{Logmel} \tag{3}$$

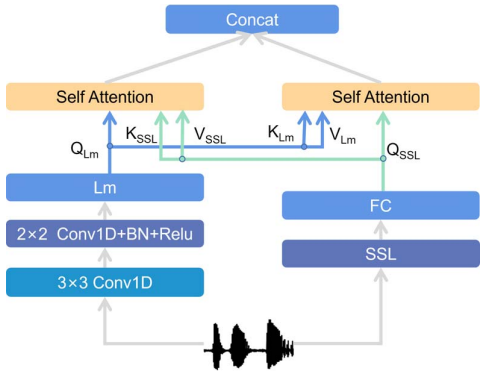

Fig. 3: Logmel combined with self-supervised features through self-attention framework

$$K_{SSL} = f_{SSL}(S)W^K_{SSL} \tag{4}$$

$$V_{Logmel} = f_{Logmel}(S)W^V_{Logmel} \tag{5}$$

$$V_{SSL} = f_{SSL}(S)W^V_{SSL} \tag{6}$$

Logmel is matched dimensionally with self-supervised feature through convolution operation. The specific model parameters are shown in Fig. 3. Then, Logmel is used as query and self-supervised feature as key and value to obtain the context vector C_{Logmel} of Logmel. Conversely, self-supervised feature is used as query and logmel as key and value. The self-supervised context vector C_{SSL}. Equations (7) and (8) describe those self-attention mechanisms.

$$C_{Logmel} = Softmax(\frac{Q_{Logmel} * K_{SSL}}{\sqrt{D}})V_{SSL} \tag{7}$$

$$C_{SSL} = Softmax(\frac{Q_{SSL} * K_{Logmel}}{\sqrt{D}})V_{Logmel} \tag{8}$$

Where D is constant. In accordance with paper [21], we set D to 128 in the experiment. Finally, the sum will be merged in the feature dimension and then passed through the classification network.

B. ATT-Resnet

We adopted the network architecture changed from [22], and added the tail of each residual module to the attention module Convolutional Block Attention Module (CBAM) [23] on the basis of Resnet. The overall network structure is shown in Fig. 4.

CBAM is composed of channel attention module and space attention module in series. The channel attention module can dynamically adjust the feature weights of different channels, highlighting the importance of extracting a certain level of feature information from each input feature mapping channel, so as to improve the feature expression effect. Spatial attention module is weighted based on the importance of the features of each spatial location. Because there are different information

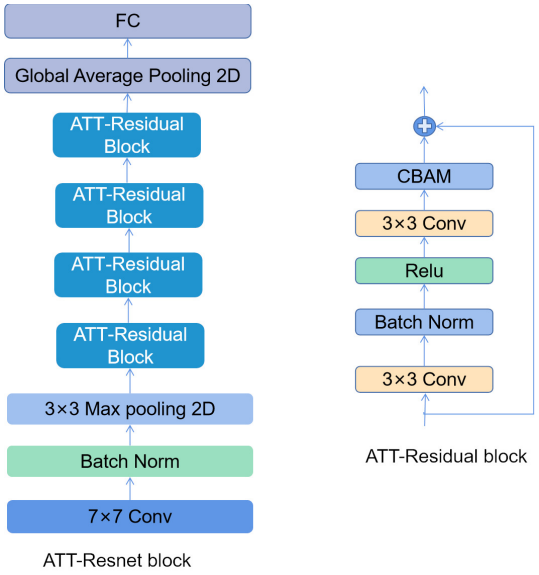

Fig. 4: Logmel combined with self-supervised features through self-attention framework

contributions in different regions of the image, the local features of the image can be better extracted by weighting the spatial position. Fig. 5-a shows the overall block diagram. Channel attention module is compressed in space dimension to highlight the importance of extracting a certain level of feature information from each input feature mapping channel. The overall network structure is shown in Fig. 5-b. The channel pays attention to the feature map F first through average pooling and maximum pooling. Compared with the single pooling method, the double pooling method of average pooling and maximum pooling has stronger feature expression ability and can learn more rich and diverse feature information from the feature mapping. Then, the pooled feature map and a hidden layer are respectively input into a shared network containing multi-layer perceptron MLP and a hidden layer to generate the channel attention diagram. Then, the normalized attention weight is obtained by the activating function. The overall calculation formula is as follows.

$$M_c(F) = \sigma(MLP(Avgpool(F)) + MLP(Maxpool(F))) \tag{9}$$

Where is the sigmoid function.

Spatial attention module generates spatial attention map by using spatial relations between features. The general block diagram is depicted in Fig. 5-c. Similar to channel attention module, average pooling and maximum pooling operations are applied to aggregate channel information of feature map along the channel axis to generate two 2D maps: $F^S_{avg} \in R^{1\times H\times W}$ and $F^S_{max} \in R^{1\times H\times W}$, and join them to generate an efficient feature map, and then join and convolve them through the standard convolution layer to generate two-dimensional spatial attention diagrams.

$$M_s(F) = \sigma(f^{7\times 7}([Avgpool(F); Maxpool(F)])) \tag{10}$$

Where is sigmoid function, the convolution operation of filter kernel is 7×7.

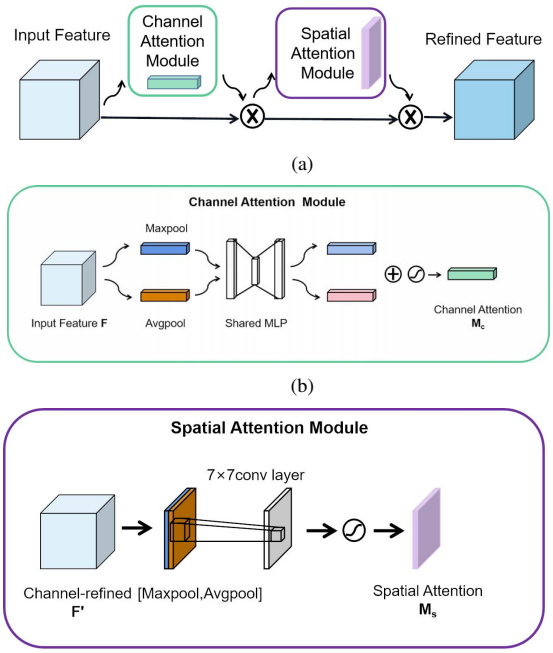

Fig. 5: CBAM overall architecture. a.b.c represents the block diagram of CBAM, channel attention module and spatial attention module respectively

IV. Experiment

A. Datasets and Evaluation Metrics

The ASVspoof 2021 dataset includes LA and DF evaluation datasets. TTS, VC, and mixed spoofing attacks in the LA evaluation task are consistent with those in the ASVspoof 2019 LA dataset. But on top of that, each speech is transmitted over a variety of telephone systems, including Voice over ip (VoIP) and the Public switched Telephone Network (PSTN). The DF evaluation task is similar to the LA task, where each speech signal is encoded and then decoded to recover the uncompressed audio, as detailed in [2]. However, since no new training and development data set has been published for the ASVspoof 2021 data set, the training and development partition of the ASVspoof 2019 database based on the speech derived from the VCTK basic corpus is used. ASVspoof 2019 LA train part is a combination of 2580 bonafide files and 22800 spoofed files to train our model. The development section consisted of 2,548 bonafide utterances and 22,926 spoof utterances for verification, as shown in Table I.

TABLE I: ASVspoof 2019 LA Dataset Details

Subset	Bonafide	Spoofed
Development	2548	22296
Training	2580	22800

For the LA dataset, the system was evaluated using the tandem detection cost function (*t-DCF*) and equal error rate

(*EER*), while the DF dataset did not involve the ASV system. Therefore, the primary metric for the DF task is the *EER*. The assumption is that scores greater than a threshold are labeled as α, while scores less than or equal to the threshold are labeled as β. Suppose a is the number of spoofed speeches in α, b is the total number of spoofed speeches, c is the number of bonafide speeches in β, and d is the total number of bonafide speeches. The specific calculation formula is :

$$P_{fa}(\theta) = \frac{a}{b} \tag{11}$$

$$P_{miss}(\theta) = \frac{c}{d} \tag{12}$$

$$EER = P_{fa}(\theta) = P_{miss}(\theta) \tag{13}$$

$P_{fa}(\theta)$ refers to the ratio of the number of forged speech samples judged as genuine by the detection system to the total number of forged speech samples. $P_{miss}(\theta)$ refers to the ratio of the number of genuine speech samples judged as forged by the detection system to the total number of genuine speech samples. Equal error rate equals the value when $P_{fa}(\theta)$ and $P_{miss}(\theta)$ are equal, The lower the equal error rate, the better the performance of the forged speech detection model.

$$t\text{-}DCF = min(\gamma P_{miss}(\theta) + P_{fa}(\theta)) \tag{14}$$

γ depends on the performance of ASV system and the priority of forgery attack. The smaller the *t-DCF*, the better the performance of the forged speech detection system.

B. Experimental Setting

In the training process, we randomly cut the input utterance into a 4-second segment. The input manual feature of the model is 128-dimension log mel-filterbank, which is used to combine the self-attention with the self-supervised feature after fine tuning. The hamming window length is 25ms, the skip size is 10ms, and there are 512 size FFT bins. We apply mean and variance normalization to logarithmic mel filter banks. All models were trained with a batch size of 2 and optimized with an Adam optimizer with a weight attenuation of 1e-4. The loss function is cross entropy loss, and the learning rate is set to 1e-6.

C. Results and Analysis

1) Main Experiment: We compared it with other systems on the ASVspoof 2021 LA dataset, and the experimental results are shown in Table II. As can be seen from the table, the *t-DCF* and *EER* of our proposed system can be 0.3008 and 4.12%. Performance indicators are better than all other single systems (without model fusion). It has good robustness in synthetic speech detection.

2) Ablation Experiment: In order to verify the performance of our proposed system, we conducted experiments on each module separately. Table III shows the experimental results of different modules. As shown in the table, combining self-supervised features with Logmel using self-attention as input to the ATT-Resnet network achieved an *EER* of 4.12%, which is 0.73% lower than inputting them into the Resnet network, demonstrating the effectiveness of self-attention. Combining self-supervised features with Logmel using self-attention as input to the ATT-Resnet network also outperformed using only self-supervised features by 0.33%, proving that CBAM can improve the performance of synthetic speech detection. Moreover, the experimental results showed that combining self-supervised features with Logmel using self-attention as input to the ATT-Resnet network produced the best results.

TABLE II: PERFORMANCE COMPARISON WITH OTHER SINGLE SYSTEMS ON THE EVALUATION SET OF THE ASVSPOOF 2021 LA

System	*EER*(%)	*t-DCF*
CQCC-GMM [24]	15.62	0.4974
LFCC-GMM [25]	19.3	0.5758
LFCC-LCNN [26]	9.26	0.3445
RawNet2 [11]	9.5	0.4257
LFCC-ECAPA-TDNN [27]	5.46	0.3094
SSL-AASIST [16]	4.48	0.3094
Our	**4.12**	**0.3008**

TABLE III: RESULT OF DIFFERENT MODULES IN ASVSPOOF 2021 LA SCENARIOS

System	*EER*(%)	*t-DCF*
SSL-ATT-Resnet18	4.52	0.3105
Self attention fusion-Resnet18	4.85	0.3134
Self attention fusion-ATT-Resnet18	**4.12**	**0.3008**

3) Experiment across data sets: In order to verify the generalization performance of the model used, the trained model was evaluated on ASVspoof 2021 DF, and the experimental results were shown in Table IV. It can be seen that the experimental results of self-attention combination are better than the existing system performance, but when CBAM module is added, *EER* increases slightly. The reason may be that some high-frequency information may be lost or signal noise is more serious in the low-quality compressed synthesized speech of MP3 or MP4, so more special and detailed audio preprocessing and feature extraction methods are needed. CBAM adapts attention to adjust channel responses and spatial locations for better performance. High-frequency information and noise are one of the important features of speech signals. If this information is missing or interfered, it will affect the correct extraction and analysis of speech features by CBAM. Therefore, the performance of CBAM will be reduced in the experiment.

V. CONCLUSION

In this paper, we have implemented a self-attention-based combination of self-supervised features and Logmel features to better capture complex patterns and contextual information in audio signals, thereby achieving generalization performance for synthetic speech detection. The experimental results show that compared to the SSL baseline, we achieved a certain improvement in performance on the ASVspoof 2021 LA and

TABLE IV: COMPARISON OF THE RESULTS OF EACH SYSTEM ON THE ASVSPOOF 2021 DF DATASET

System	*EER*(%)
CQCC-GMM [24]	25.56
LFCC-GMM [25]	25.25
LFCC-LCNN [26]	23.48
RawNet2 [11]	22.38
LFCC-ECAPA-TDNN [27]	20.33
SSL-AASIST [16]	4.57
Self attention fusion-Resnet18	**3.9**
Self attention fusion-ATT-Resnet18	5.34

DF datasets. With the help of CBAM, the performance of self-attentional combination features on LA is further improved.

Due to the relatively large number of parameters in the pre-training of the self-supervised model, the experimental efficiency is low. In the future work, we plan to explore how to achieve a certain reduction in the size of the model without significant changes in the experimental results.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No. 62106037, No. 62076052), the Science and Technology Innovation Foundation of Dalian (No. 2021JJ12GX018), the Application Fundamental Research Project of Liaoning Province (2022JH2/101300262), and the Major Program of the National Social Science Foundation of China (No.19ZDA127).

REFERENCES

[1] Andreas Nautsch,Xin Wang, Nicholas Evans, Tomi H. Kinnunen, Ville Vestman, Massimiliano Todisco, et al. ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech, in IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 3, no. 2, pp. 252-265, April 2021.

[2] Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., et al. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 47-54, 2021.

[3] Firoz Shah. A, Raji Sukumar A, Babu Anto P, Discrete Wavelet Transforms and Artificial Neural Networks for Speech Emotion Recognition, International Journal of Computer Theory and Engineering vol. 2, no. 3, pp. 319-322, 2010.

[4] Shivesh Ranjan, Exploring the Discrete Wavelet Transform as a Tool for Hindi Speech Recognition, International Journal of Computer Theory and Engineering vol. 2, no. 4, pp. 642-646, 2010.

[5] Isao Nakanishi, Hironori Namba, and Shigang Li, Speech Enhancement Based on Frequency Domain ALE with Adaptive De-Correlation Parameters, International Journal of Computer Theory and Engineering vol. 5, no. 2, pp. 292-297, 2013.

[6] Todisco, M., Delgado, H., Lee, K.A., Sahidullah, M., Evans, N., Kinnunen, T., Yamagishi, J. (2018) Integrated Presentation Attack Detection and Automatic Speaker Verification: Common Features and Gaussian Back-end Fusion. Proc. Interspeech 2018, 77-81.

[7] Tian X, Wu Z, Xiao X, Chng E, Li H, Spoofing detection from a feature representation perspective. ICASSP 2016 - 2016 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016: 2119-2123.

[8] HUANG Lian, PUN C M. Audio replay spoof attack detection using segment-based hybrid feature and DenseNet-LSTM network. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).Brighton, UK. IEEE, 2019: 2567-2571.

[9] DAS R K, YANG Jichen, LI Haizhou. Assessing the scope of generalized countermeasures for anti-spoofing. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain. IEEE, 2020: 6589-6593.

[10] Tak, H., Jung, J.-w., Patino, J., Todisco, M., Evans, N. Graph Attention Networks for Anti-Spoofing. Proc. Interspeech 2021, 2356-2360.

[11] Tak H, Patino J, Todisco M, Andreas N, Nicholas E, Anthony L, End-to-end anti-spoofing with rawnet2. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6369-6373.

[12] Alzantot, M., Wang, Z., Srivastava, M.B. Deep Residual Neural Networks for Audio Spoofing Detection. Proc. Interspeech 2019, 1078-1082.

[13] LI Xu, LI Na, WENG Chao, Liu Xunying, Su Dan, Yu Dong, et al. Replay and synthetic speech detection with Res2Net architecture. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON,Canada. IEEE, 2021: 6354-6358.

[14] Jung J, Heo H S, Tak H,Shim H, Chung J, Lee B, et al. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 6367-6371.

[15] Wang, X., Yamagishi, J. (2022) Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures. Proc. The Speaker and Language Recognition Workshop (Odyssey 2022), 100-106.

[16] Tak, H., Todisco, M., Wang, X., Jung, J.-w., Yamagishi, J., Evans, N. (2022) Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation. Proc. The Speaker and Language Recognition Workshop (Odyssey 2022), 112-119.

[17] A. Baevski, Y. Zhou, A. Mohamed,M. Auli, Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, in Proc NIPS, vol. 33, pp. 12449–12460, 2020.

[18] SUBRAMANI N, RAO D. Learning efficient representations for fake speech detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 4, pp. 5859-5866, 2020.

[19] LAI C I, CHEN Nanxin, VILLALBA J, Dehak N, ASSERT: anti-spoofing with squeeze-excitation and residual networks. Interspeech 2019. ISCA: ISCA, 2019: 1013-1017.

[20] Dan Berrebbi, Jiatong Shi, Brian Yan, Osbel Lopez-Francisco, Jonathan D. Amith, Shinji Watanabe, Combining spectral and selfsupervised features for low resource speech recognition and translation, unpublished.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, et al. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, vol. 30, pp. 6000–6010, 2017.

[22] Ling, H., Huang, L., Huang, J., Zhang, B., Li, P. (2021) Attention-Based Convolutional Neural Network for ASV Spoofing Detection. Proc. Interspeech 2021, 4289-4293.

[23] Woo S, Park J, Lee J Y, Kweon I S, Cbam: Convolutional block attention module. Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[24] Massimiliano Todisco, Hector Delgado, and Nicholas ´ Evans, Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification, Computer Speech & Language, vol. 45, pp. 516 – 535, 2017.

[25] Sahidullah, M., Kinnunen, T., Hanilçi, C. (2015) A comparison of features for synthetic speech detection. Proc. Interspeech 2015, 2087-2091, 2015.

[26] Wang, X., Yamagishi, J. (2021) A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection. Proc. Interspeech 2021, 4259-4263, 2021.

[27] X Chen, Y Zhang, G Zhu, Z Duan,UR Channel-Robust Synthetic Speech Detection System for ASVspoof 2021. Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 75-82, 2021.