Neural Networks 175 (2024) 106320

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Full Length Article

Spatial reconstructed local attention Res2Net with F0 subband for fake speech detection

Cunhang Fan [a,*], Jun Xue [a], Jianhua Tao [b,*], Jiangyan Yi [c], Chenglong Wang [c], Chengshi Zheng [d], Zhao Lv [a,*]

[a] *Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China*
[b] *Department of Automation, Tsinghua University, Beijing, 100190, China*
[c] *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China*
[d] *Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190, China*

ARTICLE INFO

Keywords:
ASVspoof
Fake speech detection
Fundamental frequency
Res2Net

ABSTRACT

The rhythm of bonafide speech is often difficult to replicate, which causes that the fundamental frequency (F0) of synthetic speech is significantly different from that of real speech. It is expected that the F0 feature contains the discriminative information for the fake speech detection (FSD) task. In this paper, we propose a novel F0 subband for FSD. In addition, to effectively model the F0 subband so as to improve the performance of FSD, the spatial reconstructed local attention Res2Net (SR-LA Res2Net) is proposed. Specifically, Res2Net is used as a backbone network to obtain multiscale information, and enhanced with a spatial reconstruction mechanism to avoid losing important information when the channel group is constantly superimposed. In addition, local attention is designed to make the model focus on the local information of the F0 subband. Experimental results on the ASVspoof 2019 LA dataset show that our proposed method obtains an equal error rate (EER) of 0.47% and a minimum tandem detection cost function (min t-DCF) of 0.0159, achieving the state-of-the-art performance among all of the single systems.

1. Introduction

Automatic speaker verification (ASV) technology has become increasingly mature, but it remains vulnerable to attack by existing synthetic speech techniques. Generally, fake speech can be divided into three types: audio playback (Ali, Sabir, & Hassan, 2021; Fan, Ding, Yi, Li & and Lv, 2023; Fan, Zhang et al., 2023; Hajipour, Akhaee, & Toosi, 2021; Kinnunen, Sahidullah et al., 2017; Kinnunen et al., 2017; Paul, Das, Sinha, & Prasanna, 2016; Shang & Stevenson, 2008), text-to-speech (TTS) (Huang, Lin, Liu, Chen, & Lee, 2022; Shchemelinin, Vadim, & Simonchik, 2013; Zhang, Gu, Yi, & Zhao, 2022), and voice conversion (VC) (Chen, Kumar, Nagarsheth, Sivaraman, & Khoury, 2020; Kinnunen et al., 2012; Tian, Lee, Wu, Chng, & Li, 2017). To reduce the risk of spoofing attacks on ASV caused by fake audio, the ASVspoof challenges have been held successively in 2015 (Wu et al., 2015), 2017 (Kinnunen, Sahidullah et al., 2017), 2019 (Todisco et al., 2019), and 2021 (Yamagishi et al., 2021). In 2022, the Audio Deep Synthesis Detection (ADD 2022) (Yi et al., 2022) was also successfully held. The ASVspoof challenge has two sub-challenges, one is logical access (LA)[1] attacks using TTS and VC algorithms, and the other is physical access (PA) attacks using audio playback. The research in this paper focuses on LA attacks. Currently, the main focus of research in fake speech detection (FSD) lies in the design of front-end features and back-end models.

For front-end features, many acoustic features are investigated (Das, Yang, & Li, 2019; Doan, Nguyen-Vu, Jung, & Hong, 2023; Fan et al., 2023; Huang, Cui, Huang, & Kang, 2023; Li, Wang, He, Abdullahi, & Li, 2022; Paul, Pal, & Saha, 2017; Wei, Long, Wei, & Li, 2022; Williams & Rownicka, 2019; Yang & Das, 2020; Yang, Das, & Zhou, 2019b), such as Mel Frequency Cepstral Coefficients (MFCC), constant Q cepstral coefficients (CQCC), linear frequency cepstral coefficients (LFCC) and so on. In addition, in Witkowski et al. (2017), it is proposed to use Inverse MFCC (IMFCC), Linear Prediction Cepstral Coefficients (LPCC), and LPCCres[2] features. Then, the high-frequency components of these three features are fed to a classifier that classifies real samples and replayed samples. In Chettri, Kinnunen, and Benetos (2020), it is proposed to divide the whole frequency band into multiple disjoint subbands. A joint subband modeling architecture is designed to learn the

* Corresponding authors.
E-mail addresses: cunhang.fan@ahu.edu.cn (C. Fan), jhtao@tsinghua.edu.cn (J. Tao), kjlz@ahu.edu.cn (Z. Lv).

[1] Proposed by the ASVspoof challenge, it is a scenario for the study of fake speech detection.

[2] Proposed by the authors of Witkowski, Kacprzak, Zelasko, Kowalczyk, and Galka (2017), and derived by the LPCC.

https://doi.org/10.1016/j.neunet.2024.106320
Received 18 July 2023; Received in revised form 5 March 2024; Accepted 13 April 2024
Available online 16 April 2024

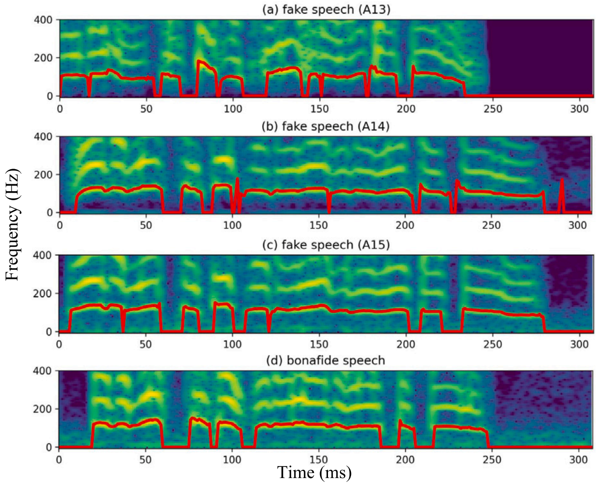


Fig. 1. The spectrum and F0 distribution of three different types fake speech and the corresponding bonafide speech. Where the red line means the distribution of F0. A13, A14 and A15 denote three different TTS algorithms drawn from the ASVspoof 2019 LA dataset, as described in Wang et al. (2020). The F0 distribution of these three different types of fake speech is distinctly different from the corresponding bonafide speech. This indicates that the F0 feature contains the discriminative information for the FSD task.

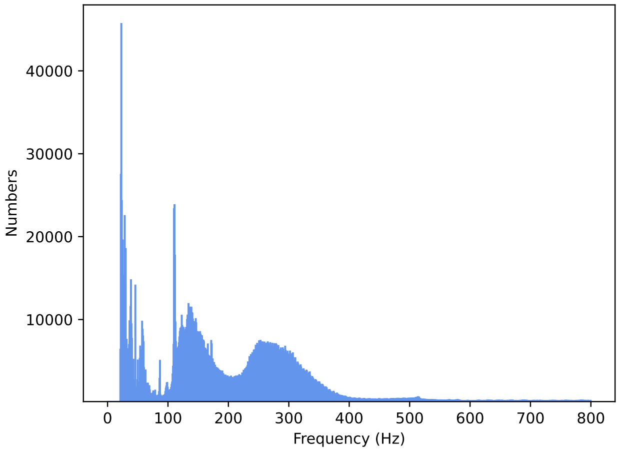


Fig. 2. The frequency of F0 distribution in the ASVspoof 2019 LA training dataset. Where the abscissa is the frequency corresponding to F0, and the ordinate is the number of F0 at this frequency.

specific features of subbands. In Zhang, Wang and Zhang (2021), it is proposed to divide the log power spectrogram (LPS) feature frequency band into two frequency bands, namely high frequency and low frequency. Based on the results of the experiments, low frequency has superior performance to high frequency. While these methods make significant advances in FSD and demonstrate that different frequency bands have different effects, they do not specify how the bands should be divided.

For back-end models, deep neural network (DNN) (Jung et al., 2022; Kim & Ban, 2023; Liu, Zhang, & Gao, 2024) classifiers can acquire impressive results for FSD task. The ResNet (He, Zhang, Ren, & Sun, 2016; Xue et al., 2023) architecture, which has been successful in the image field (He, Xu, Zhang, & Zhu, 2023; Sun, Ding, & Guo, 2022), has strong feature capture capabilities. Therefore, in Ling, Huang, Huang, Zhang, and Li (2021), Zhang, Jiang and Duan (2021), Zhang, Wang et al. (2021), authors propose a series of ResNet-based classifiers to detect fake speech. To further enhance the generalization capability of the model, Gao et al. (2019) propose the Res2Net structure, which partitions channels into multiple groups and enables interaction through residual connections within each group to extract multi-scale features. Consequently, researchers (Li et al., 2021; Li, Wu, Lu, Liu & Meng, 2021) have made many beneficial attempts to use the Res2Net architecture for the FSD task. However, the residual connections directly add information from the previous channel group to the next, and research (Li, Wu et al., 2021) has shown that the cross-channel information can generate redundant information. Therefore, after aggregating information from multiple channel groups in Res2Net, the salient discriminative features may be interfered with by redundant information. These issues limit the performance of the FSD system.

Recent research in text-to-speech (TTS) has indicated that the fundamental frequency (F0) is very important for the quality of synthetic speech. For example, in Łańcucki (2021), a new TTS model Fastpitch is proposed to predict the F0 contour during inference. Changed predictions make the generated speech more human-like. The field of Voice Conversion (VC) is also extensively exploring how to better model F0 when synthesizing speech. For example, in Qian, Jin, Hasegawa-Johnson, and Mysore (2020), the authors improve the autoencoder, which could better generate F0 contours consistent with the target speaker, as a way to significantly improve speech quality. However, it is worth noting that the rhythm of bonafide speech is often difficult to replicate, which leads to the F0 of synthetic speech being very different from the F0 of real speech. To compare the F0 distribution of fake speech and bonafide speech, Fig. 1 shows the spectrum and F0 distribution of three different fake voices and one bonafide speech, with the red line depicting the distribution of F0. From Fig. 1 we can find that the F0 distribution of these three different types of fake speech is distinctly different from the corresponding bonafide speech. This indicates that the F0 feature contains the discriminative information for the FSD task.

Unfortunately, F0 is difficult to model directly as a valid feature for FSD. To make full use of the discriminative information of F0, this paper proposes an F0 subband for FSD task, which is the subband of amplitude spectrum. Fig. 2 shows the F0 distribution in the ASVspoof 2019 LA training dataset. From Fig. 2 we can find that most of the F0 is distributed between 0–400 Hz. Therefore, the frequency band containing most of the F0 is used as the F0 subband. Overall, compared to other acoustic features, the F0 subband contains a priori discriminative information, which avoids interference from redundant information. In addition, to effectively model the F0 subband, we propose a novel spatial reconstructed local attention Res2Net (SR-LA Res2Net) for FSD. Specifically, the Res2Net is used as the backbone network, which can capture the multi-scale information of the input feature. However, the gradual superposition of cross-channel group information will cause more artifacts to the spatial structure of the feature, and the redundant information generated by aggregation may obscure some important information. To address these problems, we design a spatial reconstructed (SR) block at the residual connection in Res2Net, which is used to reconstruct the spatial structure. Finally, the local attention (LA)[3] block is integrated at the bottom of Res2Net to focus on local information and capture the discriminative information of the F0 subband.

The main contributions of this study can be summarized as two-fold. Firstly, we propose to use the F0 subband for the FSD task, which is a very discriminative feature. Secondly, a novel SR-LA Res2Net architecture is designed to model the F0 subband, which can effectively solve the shortcomings of Res2Net when expanding feature receptive fields. The experimental results on the ASVspoof 2019 LA dataset show that our proposed method is very effective for the FSD task, and it can acquire the state-of-the-art performance among all of the single systems.

The rest of this article is arranged as follows. Section 2 introduces the related works. The proposed method is introduced in Section 3.

[3] The module embedded in the Res2Net architecture proposed in this paper.

Experiments and results are given in Section 4. Section 5 shows the discussions. Section 6 draws conclusions.

2. Related works

For the FSD task, many studies (Chettri et al., 2020; Yang, Das, & Li, 2019a; Zhang, Yi & Zhao, 2021) have shown that different frequency bands have different effects. In Zhang, Wang et al. (2021), the authors focus on global channel attention using squeeze and extraction blocks and explore the impact of high frequency and low frequency subband for the FSD task. The low frequency subband achieves good performance. In Ling et al. (2021), the authors propose a frequency attention block and a channel attention block, which pay attention to the basic subband correlation and channel relationship, respectively.

In addition, many studies (Jung et al., 2022; Lv, Zhang, Tang, & Hu, 2022; Tak et al., 2021; Tak, weon Jung, Patino, Todisco & Evans, 2021) are based on the Res2Net for FSD and acquire quite good performances. In Li et al. (2021), the authors use Res2Net to enhance the system's generalization to unseen spoofing attacks and integrate squeeze and extract blocks to further improve performance. In Li, Wu et al. (2021), the authors propose a channel-wise gated Res2Net (CG-Res2Net), which dynamically adjusts the correlation between channels through a gating mechanism and suppresses channels with small correlations. It further enhances the generalization ability of the system against unseen spoofing attacks.

In this paper, we propose the F0 subband and SR-LA Res2Net for FSD. Compared with Res2Net and CG-Res2Net, the proposed SR-LA Res2Net has better generalization ability.

3. The proposed F0 subband with SR-LA Res2Net

In this paper, we propose an F0 subband with SR-LA Res2Net for FSD. Because the F0 of synthetic speech is very different from the real one. Therefore, we think the F0 contains the discriminative information and apply the F0 subband as the input feature for FSD. To further improve the performance of FSD, we propose the SR-LA Res2Net to model the F0 subband feature, which can effectively solve the shortcomings of Res2Net when expanding feature receptive fields.

3.1. F0 subband

To make full use of the discriminative information of F0, we extract the F0 subband based on the LPS. Specifically, the short-time Fourier transform (STFT) is used to convert the time domain raw waveform $\mathbf{x}[k]$ into the time–frequency (T–F) domain.

$$\boldsymbol{X}_{\mathrm{r}}[t,f] + j \cdot \boldsymbol{X}_{\mathrm{i}}[t,f] = STFT(\mathbf{x}[k]) \tag{1}$$

where k is the time index of raw waveform $\mathbf{x}[k]$. $STFT$ means the operation of STFT. $\boldsymbol{X}_{\mathrm{r}} \in \mathbb{R}^{F\times T}$ and $\boldsymbol{X}_{\mathrm{i}} \in \mathbb{R}^{F\times T}$ are the corresponding real and imaginary part of STFT, respectively. t is the index of time frame and f is the index of frequency bin. F and T are the number of frequency bins and time frames, respectively. For convenience, (t, f) is omitted from the following formulas in this paper.

The full frequency bands of $\boldsymbol{LPS}_{\mathrm{full}}$ can be acquired as follows:

$$\boldsymbol{LPS}_{\mathrm{full}} = \log\sqrt{(\boldsymbol{X}_{\mathrm{r}})^2 + (\boldsymbol{X}_{\mathrm{i}})^2} \in \mathbb{R}^{F\times T} \tag{2}$$

From Fig. 2 we can find that most of the F0 is distributed between 0–400 Hz. Therefore, the 0–400 Hz of LPS is applied as our F0 subband $\boldsymbol{LPS}_{\mathrm{F0}}$.

$$\boldsymbol{LPS}_{\mathrm{F0}} = \boldsymbol{LPS}_{0\text{–}400\ \mathrm{Hz}} \tag{3}$$

3.2. Model architecture

To effectively model the F0 subband and improve the performance of FSD, we propose the SR-LA Res2Net architecture. Fig. 3 shows the schematic diagram of the proposed SR-LA Res2Net architecture. Firstly, to extract the multi-scale information of the F0 subband, the Res2Net is used as the backbone. However, when the channel group is constantly superimposed, the Res2Net may generate redundant information so that much important information may be lost. To address this issue, the SR block is proposed at the residual connection between channel groups, which can restore the spatial structure. In addition, an LA block is designed at the bottom of Res2Net to pay attention to local information and remove the influence of redundant information. Therefore, our proposed SR-LA Res2Net can further remove spatial artifacts and redundant information while extracting multi-scale features, thereby improving the generalization ability of the model to unseen spoofing attacks.

3.2.1. The res2net architecture

The ResNet has been applied in various fields as soon as it was proposed and has achieved great performance. Even if ResNet's residual connections can reduce the impact of network depth, just increasing the network depth does not improve the performance of the model very well. So Gao et al. (2019) proposed the Res2Net architecture, which obtains multi-scale features through the information transfer of channel groups. Firstly, to expand the range of interaction between channel groups, the input features are divided into n subsets according to the channel dimension after 1 × 1 convolution, denoted as s_i, where $i \in \{1, 2, \dots, n\}$. As for s_1, it does not undergo any processing. As for s_2, it is directly output after a 3 × 3 convolution $K_2(\cdot)$. As for s_3 to s_n, each s_i needs to be added to the output of K_{i-1} before passing through $K_i(\cdot)$. This process can be formulated as follows:

$$y_i = \begin{cases} s_i, & i = 1 \\ K_i\left(s_i\right), & i = 2 \\ K_i\left(s_i + y_{i-1}\right), & 3 \le i \le n \end{cases} \tag{4}$$

where n is defined as the scale dimension, indicating the number of channel groups applied to split feature maps.

Therefore, the Res2Net can increase the interaction between channel groups through residual connections in the block. Through the residual connections in the block, each channel group obtains a different amount of information, thereby it can generate multiple-scale features. Such a multi-scale representation increases the receptive field. Finally, all channel groups are aggregated and the original channel size is maintained by the 1 × 1 convolution.

3.2.2. Spatial reconstructed block

The Res2Net has a strong feature representation ability, which relies on the information transfer of multiple internal channel groups. However, as the feature information continues to be superimposed, more artifacts will appear in its spatial structure. This greatly affects the performance of the Res2Net model. Inspired by Woo, Park, Lee, and Kweon (2018), we design a SR block for residual connection between channel groups. It aims to reconstruct the feature space and remove its artifacts when the information is passed to the next channel group.

Firstly, to reduce subsequent parameter computations, we compress the channel dimension:

$$\mathcal{F}_{\mathrm{Mean}} \in \mathbb{R}^{1\times F\times T} = \mathrm{Mean}\left(\mathcal{F}_{\mathrm{in}} \in \mathbb{R}^{C\times F\times T}\right) \tag{5}$$

where $\mathrm{Mean}(\cdot)$ means the mean operation, $\mathcal{F}_{\mathrm{in}}$ is the input feature, and C is the number of channels.

Then, to further expand the receptive field, the depth-wise dilation convolution is applied:

$$\mathcal{F}_{\text{DW-DC}} \in \mathbb{R}^{1\times F\times T} = \mathrm{DW{-}DC}\left(\mathcal{F}_{\mathrm{Mean}} \in \mathbb{R}^{1\times F\times T}\right) \tag{6}$$

where DW−DC represents the operation of depth-wise dilation convolution.

Finally, $\mathcal{F}_{\text{DW-DC}} \in \mathbb{R}^{1\times F\times T}$ reconstructs the feature space through the sigmoid layer. Then it multiplies with the input feature $\mathcal{F}_{\mathrm{in}}$:

$$\mathcal{F}_{\mathrm{sr}} \in \mathbb{R}^{C\times F\times T} = \mathcal{F}_{\mathrm{in}} \otimes \mathrm{sigmoid}\left(\mathcal{F}_{\text{DW-DC}} \in \mathbb{R}^{1\times F\times T}\right) \tag{7}$$

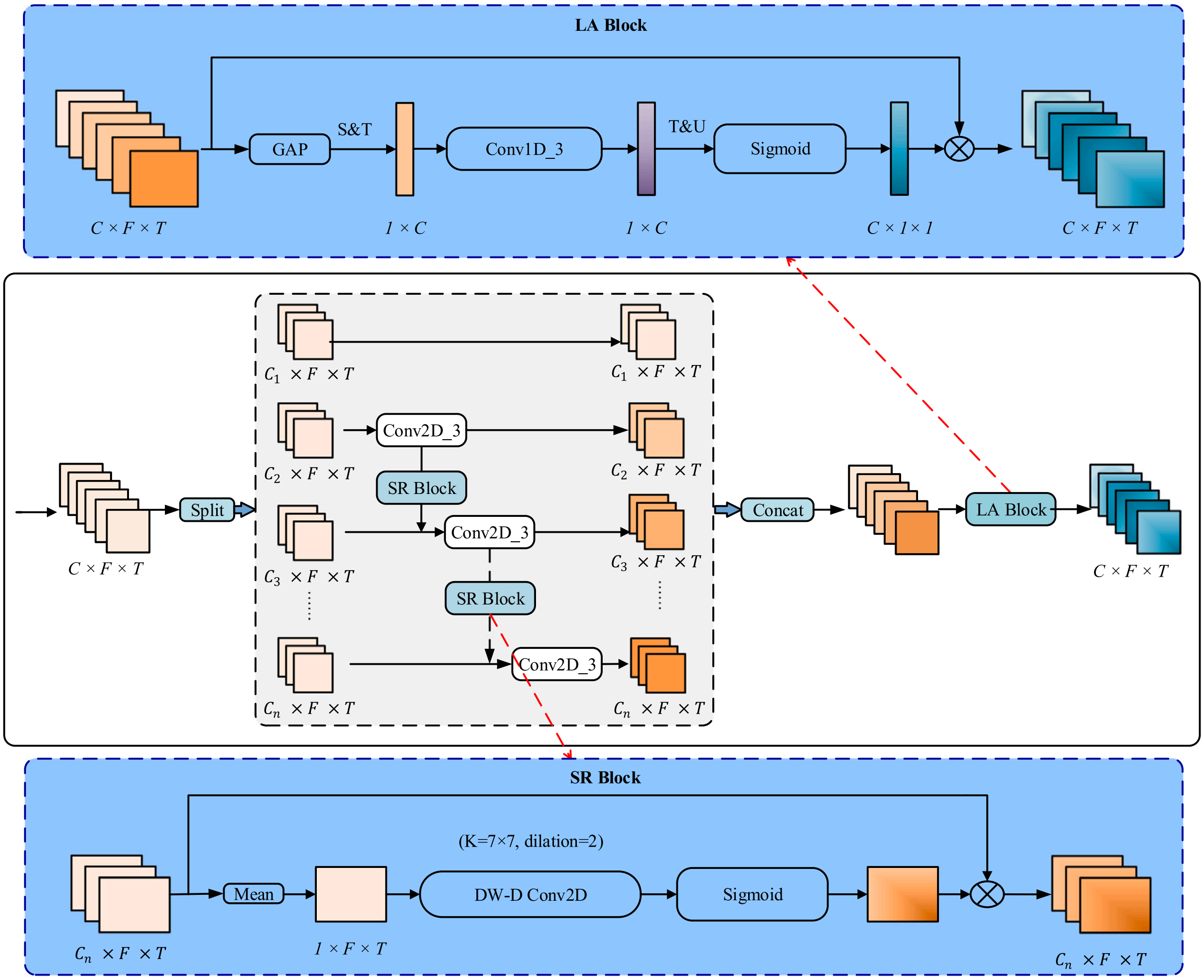


Fig. 3. The schematic diagram of the proposed SR-LA Res2Net architecture. The spatial reconstructed (SR) block is used to remove the spatial structure artifacts of the channel group, and the local attention (LA) block aims to highlight important information and remove redundant information.

where $\mathcal{F}_{\text{sr}}$ represents the reconstructed feature space. ⊗ denotes the element-wise multiplication.

3.2.3. Local attention block

Although Res2Net can obtain multi-scale global information, due to the uneven interaction information between the channel groups, this uneven global interaction may lead to information redundancy and the useful information may not be focused. To address this problem, motivated by Wang, Wu, Zhu, Li, Zuo and Hu (2020), the LA block is applied to focus on the local information.

Firstly, the global average pooling (GAP) is used to squeeze the dimensions of the input features:

$$\mathcal{F}_{\text{GAP}} \in \mathbb{R}^{C\times1\times1} = \text{GAP}\left(\mathcal{F}_g \in \mathbb{R}^{C\times F\times T}\right) \tag{8}$$

Where GAP(⋅) represents a GAP operation, $\mathcal{F}_g$ is the output of Res2Net feature aggregation.

In order to squeeze the one-dimensional channel of the convolution, squeeze and transpose operations are then applied:

$$\mathcal{F}_{\text{S\&T}} \in \mathbb{R}^{1\times C} = \text{S\&T}\left(\mathcal{F}_{\text{GAP}} \in \mathbb{R}^{C\times1\times1}\right) \tag{9}$$

where S&T(⋅) means the operation of squeeze and transpose.

In Hu, Shen, Albanie, Sun, and Wu (2019), the authors proposed the Squeeze-and-Excitation (SE) block, which is different from the LA block in that they use two fully connected layers to learn global channel attention. The first FC layer is used for dimensionality reduction, and the second FC layer is used to restore the dimensionality. Although the parameters of this method are reduced by the dimensionality reduction, the complexity of the model is still very high. In addition, the dimensionality reduction can affect the performance of the model. To address this issue, motivated by Wang, Wu et al. (2020), we apply the one-dimensional convolution to acquire the local attention information. The details are as follows:

$$\mathcal{F}_{\text{Conv}} = \text{Conv}\left(\mathcal{F}_{\text{S\&T}} \in \mathbb{R}^{1\times C}\right) \tag{10}$$

where the Conv(⋅) is the operation of one-dimensional convolution.

Then, the feature size is gradually restored by the transpose and unsqueeze operations, which is defined as follows:

$$\mathcal{F}_{\text{T\&U}} \in \mathbb{R}^{C\times1\times1} = \text{T\&U}\left(\mathcal{F}_{\text{Conv}} \in \mathbb{R}^{1\times C}\right) \tag{11}$$

where T&U(⋅) means the operation of transpose and unsqueeze.

Finally, the $\mathcal{F}_{\text{T\&U}} \in \mathbb{R}^{C\times1\times1}$ is passed by the sigmoid layer to acquire the vector of attention weight. The finally local attention vector can be obtained by multiplying the attention weight and the input feature $\mathcal{F}_g$ of local attention block.

$$\mathcal{F}_{\text{la}} = \mathcal{F}_g \otimes \text{sigmoid}\left(\mathcal{F}_{\text{T\&U}} \in \mathbb{R}^{C\times1\times1}\right) \tag{12}$$

where the $\mathcal{F}_{\text{la}}$ denotes the output of local attention block.

4. Experiments and results

4.1. Dataset

We conduct our experiments on the ASVspoof 2019 LA database and the ASVspoof 2021 LA database.

Table 1
The detailed information of ASVspoof 2019 LA Dataset. Where the "utt." means the number of utterance.

Partition	Bonafide	Spoof	Spoof
	utt.	utt.	attacks type.
Train.	2580	22 800	A01–A06
Dev.	2548	22 296	A01–A06
Eval.	7355	63 882	A07–A19

4.1.1. *ASVspoof 2019 LA database*

ASVspoof 2019 LA[4] mainly has 19 spoofing attack algorithms (A01–A19), including three types of spoofing attacks: TTS, voice conversion (VC), and audio playback. Table 1 details the components of the ASVspoof 2019 LA dataset. It can be seen that the LA subset has three parts: training, development, and evaluation. Among them, the training set is used to train the model, the development set is used to select the best performing model in training, and finally, the model performance is evaluated through the evaluation set. The training set and development set mainly include four TTS and two VC algorithms, namely A01–A06. To better evaluate the performance of the system, unseen spoofing attacks were added to the evaluation set, including two known spoofing attacks (A16 and A19) and 11 unseen spoofing attacks (A07–A15, A17, and A18).

4.1.2. *ASVspoof 2021 LA database*

The difference between ASVspoof 2019 and 2021[5] LA database is the evaluation set. Therefore, the ASVspoof 2019 LA training and dev sets are used to train the model. The evaluation set contains about 180,000 utterances transmitted through real telephone systems with different bandwidths and different codecs. The transmission interference of this data set greatly affects the performance of the FSD system and makes it more challenging.

To quantitatively evaluate the performance of different FSD systems, EER and the minimum normalized tandem detection cost function (min t-DCF) are applied. EER is the working point where the false rejection rate (FRR) and false acceptance rate (FAR) are equal.

4.2. *Experimental setup*

First, we perform STFT operations on the original audio waveform. We use Blackman as the window function of the STFT and set the window length and hop length to 1728 and 130, respectively, to obtain a spectrogram of size 865. Then, we fix the number of frames to 600 by truncating and concatenating. So the feature dimension is 865 × 600. We take the 0–400 Hz LPS feature as the F0 subband, so the corresponding frequency dimension is 45. Therefore, the first 0–45 dimensions are taken as the F0 subband features, and 45 × 600 is obtained by cutting the above features.

Network architecture: In this paper, Res2Net[6] is used as the backbone network, the proposed SR block is embedded in the internal channel residual connections of Res2Net, and the proposed LA block is embedded after the channel aggregation of Res2Net. As shown in Figs. 3 and Table 2, the details including convolution kernel, channels, and repetition counts are provided. In addition, we use Adam as the optimizer, and the parameters of the optimizer are set to: $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$ and weight decay is 10^{-4}. The number of the epoch is 32. Fig. 4 shows the EER results for different numbers of channel groups (n) based on the SR-LA Res2Net architecture. From Fig. 4, we can see that the best performance is achieved for n=8, so we set the number of channel groups to 8 in the experiment.

Table 2
The proposed SR-LA Res2Net model architecture and configuration. Dimensions refer to (channels, frequency, and time). Batch normalization (BN) and Rectified Linear Unit (ReLU), SR and LA are the spatial reconstruction block and the local attention block, respectively.

Layer	Input: 27000 samples	Output shape
Front-end	F0 subband	(45,600)(F,T)
Post-processing	Add channel Conv2D_1 BN & ReLU	(1,45,600) (16,45,600)
Res2-block	2 × {Conv2D_1 Conv2D_3 & SR Conv2D_1 LA}	(32,45,600)
Res2-block	2 × {Conv2D_1 Conv2D_3 & SR Conv2D_1 LA}	(64,23,300)
Res2-block	2 × {Conv2D_1 Conv2D_3 & SR Conv2D_1 LA}	(128,12,150)
Res2-block	2 × {Conv2D_1 Conv2D_3 & SR Conv2D_1 LA}	(256,6,75)
Output	Avgpool2D(1,1) AngleLinear	(256,1,1) 2

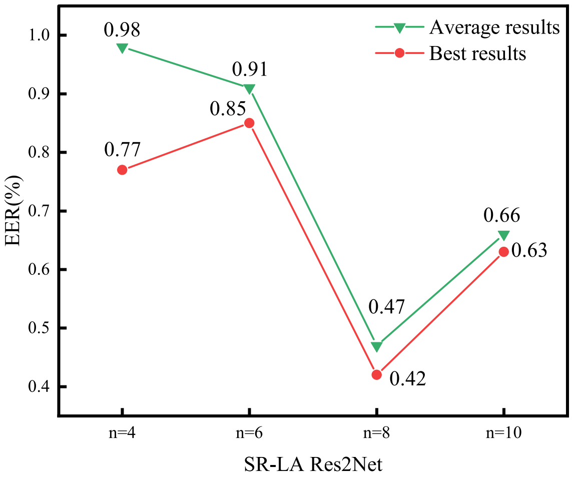


Fig. 4. The EER results of SR-LA Res2Net for different number of channel groups (n). To avoid randomness in the experiments, the results are averaged for the three runs. Where the green line meas the average results and the red line denotes the best results of the three runs.

In addition, since the ASVspoof 2021 LA dataset was interfered with by transmissions such as telephone communications, we used the Rawboost[7] (Tak, Kamble, Patino, Todisco, & Evans, 2022) data enhancement method when training the evaluation used for the ASVspoof 2021 LA dataset. Specifically, we added impulse signal-independent additive noise and stationary signal-independent additive noise to the original waveform.

[4] https://datashare.ed.ac.uk/handle/10283/3336
[5] https://zenodo.org/records/4837263
[6] https://github.com/Res2Net
[7] https://github.com/TakHemlata/RawBoost-antispoofing

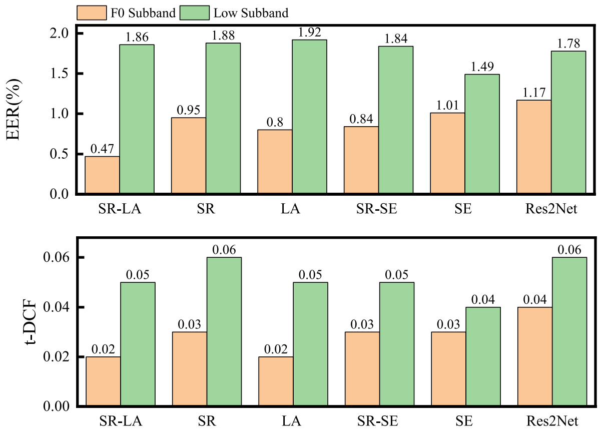


Fig. 5. The EER and t-DCF results for our proposed different systems in the ASVspoof 2019 LA evaluation set (A07-A19). Where SR, LA, and SE are the different components embedded in the Res2Net.

Table 3
Results of our proposed ablation experiments for different components. The results are the average of three runs, with the best of the three results in parentheses.

Systems	t-DCF	EER(%)
SR-LA Res2Net (F0)	**0.0159**(**0.0143**)	**0.47**(**0.42**)
SR-SE Res2Net (F0)	0.0270(0.0227)	0.84(0.74)
SR Res2Net (F0)	0.0306(0.0302)	0.96(0.95)
LA Res2Net (F0)	0.0246(0.0229)	0.80(0.77)
SE Res2Net (F0)	0.0310(0.0292)	1.01(0.95)
Res2Net (F0)	0.0353(0.0335)	1.17(1.14)
LA ResNet (F0)	0.0388(0.0364)	1.26(1.14)
SE ResNet (F0)	0.0424(0.0392)	1.36(1.23)
ResNet (F0)	0.0493(0.0406)	1.64(1.34)

4.3. Experimental results on ASVspoof 2019 LA dataset

4.3.1. Effectiveness of the F0 subband

This section evaluates the effectiveness of F0 subband feature on different network structures. Fig. 5 shows the minimal t-DCF and EER results for the different systems we proposed. To avoid randomness in the experiments, the results are averaged for the three runs, with the best of the three runs in parentheses. "F0" represents based on the F0 subband feature, and "L" represents based on the low frequency (0–4000 Hz) (Zhang, Wang et al., 2021) subband feature. The first six lines are the results of the F0 subband feature. The last six lines are experimental results based on low frequency subband feature. Fig. 7 shows the EER histograms based on the F0 subband and low subband features, with EERs calculated separately for different attack types.

From Figs. 5 and 7, we can see the following.

(1) In the LPS features, the performance of the F0 subband is better than that of the lower subband features in all cases. For example, in our proposed state-of-the-art classifier-SR-LA Res2Net, the average EER of its F0 subband is 0.47%, while the EER of the low subband is 1.86%. Even though the low subband (0–4000 Hz) has ten times more band information than F0 (0–400 Hz), it performs much worse in the FSD task. This is because the main discriminative information is concentrated in the F0 sub-band, and the other frequency band information may make it overfitting. The experimental results show that the F0 subband is an important identification feature.

(2) The F0 subband feature has general applicability for different types of attacks. For low subband features, the attack types of A08 (neural waveform), A17 (waveform filtering), and A18 (vocoder) are difficult to detect. For example, Al-Radhi, Csapó, and Németh (2018) proposed a source-filter based vocoder, in order to refine the output of the contF0 estimation, the authors used post-processing to reduce the unwanted vocalized components of unvoiced speech, resulting in a smoother contF0 trace. It can be seen in Fig. 7 that the F0 subband has good performance for different attack types, even for the notorious attack type like A17. Overall, the F0 subband features outperform the lower subbands on different classifiers.

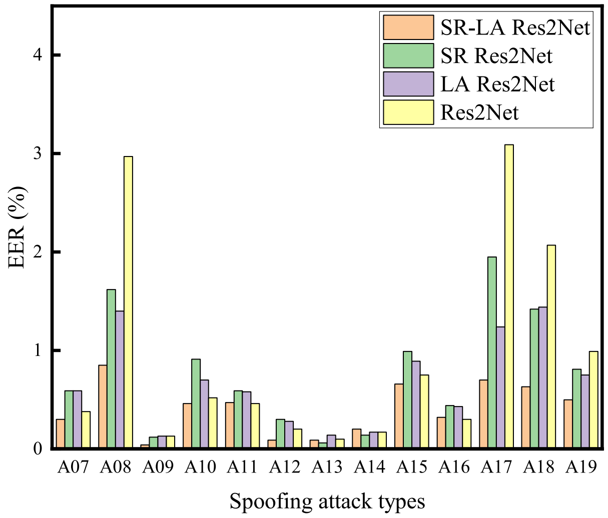


Fig. 6. EER(%) results for the evaluation subset (A07-A19). The EER(%) of each spoof method is calculated separately.

(3) SR-LA Res2Net classifier can fully exploit the discriminatory ability of F0 subbands. For example, from ResNet to SR Res2Net and then to SR-LA Res2Net, attacks such as A08 and A17 are greatly optimized under the F0 subband, but it is more difficult to develop for the low subband, which may be due to the interference of having a lot of redundant information in the low subband.

4.3.2. Effectiveness of the SR-LA Res2Net architecture

To verify the effectiveness of our proposed FSD system based on the F0 subband and the SR-LA Res2Net, we performed a series of ablation experiments. Table 3 shows the min t-DCF and EER results of our proposed different systems. In addition, to validate the effectiveness of our proposed SR-LA Res2Net, we used some recently published advanced network to model the F0 subband. Fig. 6 shows the EER for the different networks.

Firstly, the multi-scale feature representation of F0 subband is an effective way to improve the performance of the pseudo-speech detection system. the EER result of "ResNet (F0)" is 1.64%, while the EER result of "Res2Net (F0)" is 1.17%. This is due to the fact that Res2Net is designed with residual connections within the channel group so that the model can learn information at different scales to discriminate. However, when the Res2Net extracts multi-scale features, the larger the number of channel groups, the more information is superimposed, and the spatial structure of the features will also have more artifacts, which affects the ability of the model to capture fundamental discriminative information. Therefore, we design the spatial reconstructed block to be integrated into the residual connections of Res2Net to reconstruct the feature space before transferring the information. Experimental results show that SR-Res2Net (F0) has a good performance improvement over Res2Net (F0). This suggests that the spatial reconstruction mechanism helps remove spatial structure artifacts and further improves the performance of the model.

Next, we integrate a local attention block at the bottom of Res2Net. This is because the deepest channel group of the Res2Net superimposes all the information, and the rest also superimposes a lot of information, which leads to the generation of a lot of redundant information and further covers the important discriminative information. We propose

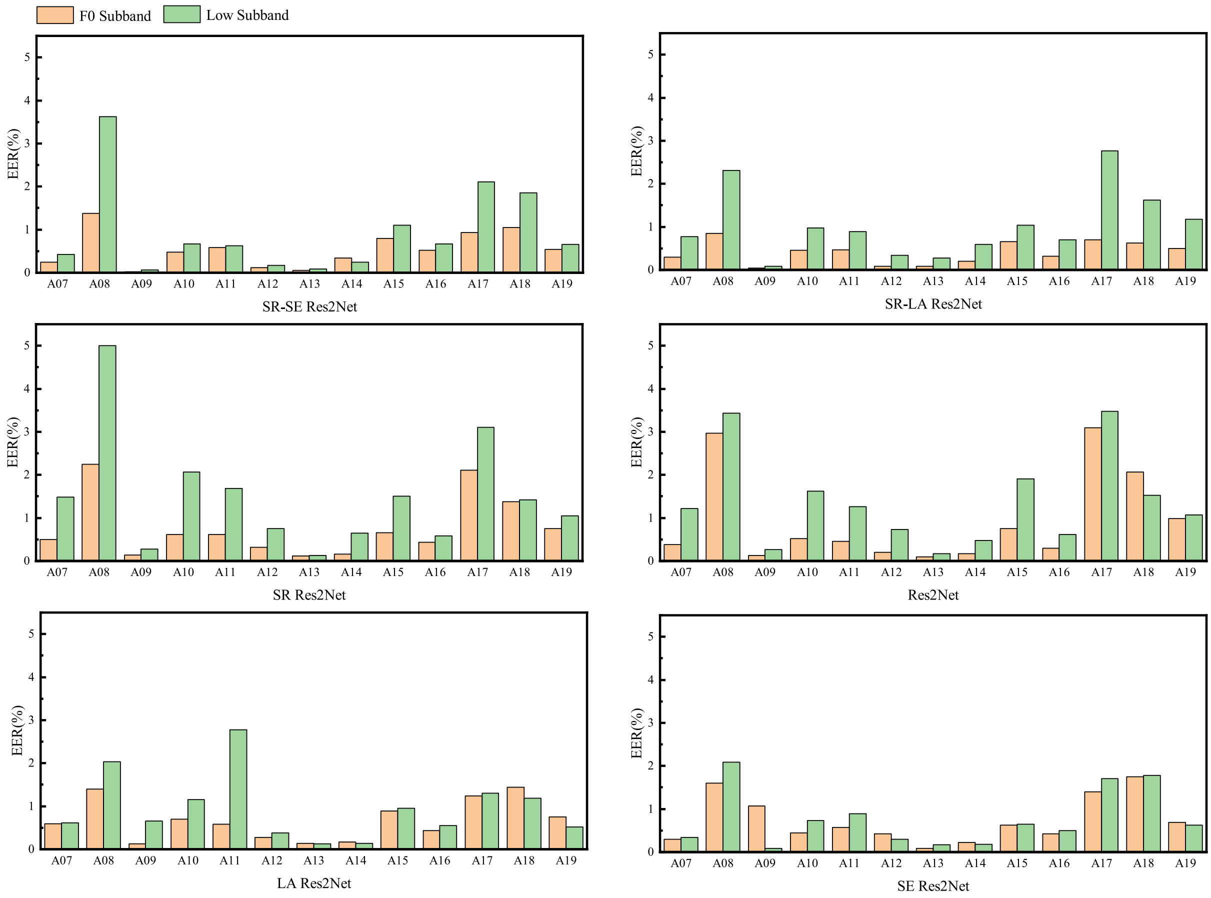


Fig. 7. EER results of our proposed different systems on the ASVspoof 2019 LA evaluation set (A07-A19). The EER of each spoof method is calculated separately.

Table 4
Comparison of the results of F0 subband features on other advanced classifiers.

Systems	Input	t-DCF	EER(%)	#Parm
ACNN (Ling et al., 2021)	FFT	0.0510	1.87	1.04M
	F0 subband	**0.0454**	**1.46**	
MCG-Res2Net (Li, Wu et al., 2021)	CQT	0.0520	1.78	1.09M
	F0 subband	**0.0299**	**1.03**	
LCNN (Lavrentyeva, Tseren, Volkova, Gorlanov, Kozlov, & Novoselov, 2019)	FFT	0.1028	4.53	0.78M
	F0 subband	**0.0417**	**1.30**	
ResNet18-L-FM (Chen et al., 2020)	LFBs	0.0520	1.81	0.68M
	F0 subband	**0.0465**	**1.48**	
SR-LA Res2Net (**Ours**)	F0 subband	**0.0159**	**0.47**	0.95M

to restore the weights of important information through local attention blocks after this information is aggregated. Specifically, we also compare the performance of local attention (LA) and global attention (SE) (Hu et al., 2019), and the experimental results show that local attention is better than global attention. We think there are two reasons: ① When generating the attention map, much long-distance information in the global information cannot accurately capture its specific connection, and the short-distance information can better judge the weight of the central information; ② The global attention block uses two fully connected layers, which are used for dimensionality reduction and expansion, respectively. The dimensionality reduction operation may result in the loss of some information, which can affect the discriminability of the model. According to the experimental results of integrating local attention blocks and global attention blocks respectively, LA Res2Net achieves an EER result of 0.80%, and SR-LA Res2Net achieves an EER result of 0.47%. This shows that local attention can capture important information in more detail and reduce the influence of redundant information left over from the network.

Finally, for the setting of the number of channel groups in the SR-LA Res2Net, we believe that n=8 is most appropriate for right in the fake audio detection task, thus achieving state-of-the-art performance. This is because the information exchange at this time is sufficient and the feature representation is reasonable.

Moreover, to verify the effectiveness of our proposed SR-LA Res2Net, we simulate the F0 subband with some recently published advanced networks. Table 4 shows the EER and t-DCF of different networks, which demonstrates the differences between different networks when using either the original features or the F0 subband. Fig. 8 shows the specific EER results of the attack. From Figs. 6 and 8, we see that the F0 subband feature can achieve excellent performance when paired with other networks, and SR-LA Res2Net can fully access the discriminative information of the F0 subband and perform extremely well in all aspects.

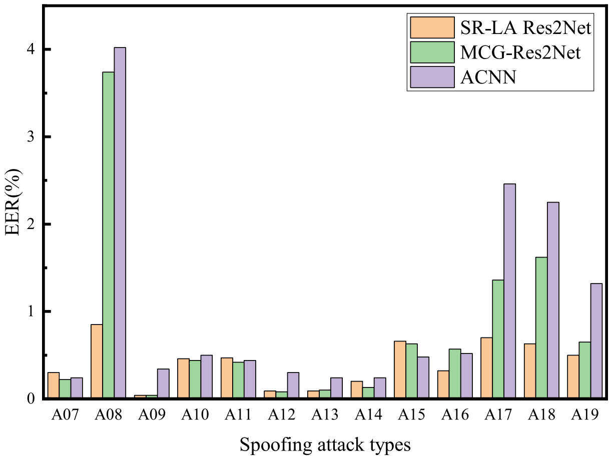
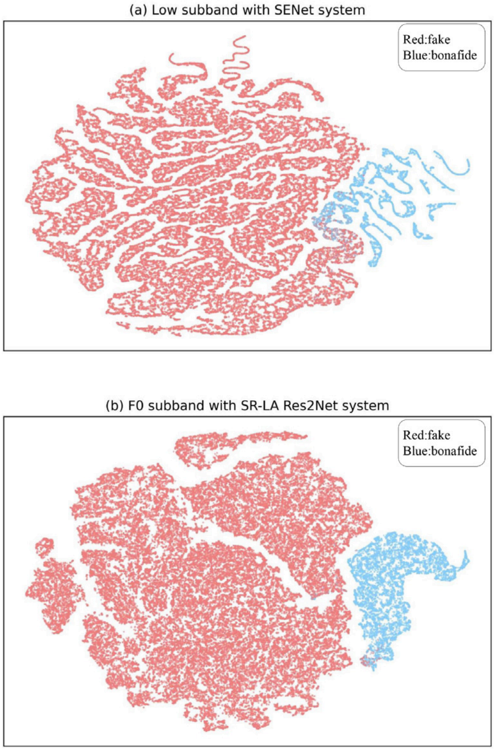


Fig. 8. EER results of the F0 subband feature on other advanced classifiers.


(a) Low subband with SENet system

(b) F0 subband with SR-LA Res2Net system

Fig. 9. Subplot (a) shows the t-SNE visualization for the low-frequency subband and SENet system, and subplot (b) shows the t-SNE visualization for the F0 subband and SR-LA Res2Net system. The blue dots represent real speech and the red dots represent false speech.

4.3.3. *Effective generalization ability of SR-LA Res2Net architecture*

Fig. 6 counts the EER of each attack algorithm for different systems. From Fig. 6, we can draw the following points. First, when different classifiers capture the details of the feature and then generalize to unknown attacks, their biases will be large. For example, the GMM-based baseline algorithm is very effective against attack algorithms such as A08, but the performance of attack algorithms such as A10, A13, A14, and A17 becomes extremely poor. Even so, it is not our original intention to only effectively target a certain attack algorithm. For example, our proposed SR-LA Res2Net (n=8, F0) can be generalized to each unseen attack more evenly, which is the focus of our work.

Second, it is well known that the A17 algorithm is notorious, and the method was judged to have the highest spoofing ability in the 2018 Speech Transformation Challenge (Kinnunen et al., 2018). However, our proposed SR-LA Res2Net system can obtain 0.70% EER on the A17 attack, which is the best performance among all systems. We believe that the multi-scale feature representation enables the FSD system to be generalized to spoofing attacks like A17. the EER results of the ResNet (F0) and Res2Net (F0) systems on A17 are 4.43% and 3.09%, respectively, which indicates that multi-scale features can greatly extend the feature receptive field and enhance its generalizability. However, the channel group information of its Res2Net architecture is constantly superimposed, which requires a spatial reconstruction block to reconstruct each channel group information, and the EER result of its SR Res2Net (F0) at A17 is 1.95%, and the experimental results prove that the spatial reconstruction block can reduce the influence of redundant information. In addition, other systems have poor performance for unseen attacks like A08, A17, and A18, but the SR-LA Res2Net (F0) system achieves high performance in the face of all unseen attacks. This further validates the need to integrate spatial reconstruction block and local attention block in Res2Net, which can greatly improve the generalization ability of the model.

Third, compared to our proposed SR-LA Res2Net (F0) system, other systems are difficult to pass in some individual deception algorithms. For example, the B1 system achieves an EER result of 26.15% in the A13 algorithm, and our SR-LA Res2Net (F0) system has an EER of 0.09 for A13, and other systems also performed well. For A18, the SR-LA Res2Net (F0) system leads the way.

In summary, the strong generalization of the SR-LA Res2Net (F0) system comes from the spatial reconstructed block and the local attention block. By reconstructing the feature space and focusing on local information, it reduces the multi-scale sequelae brought by feature representation, which greatly improves the performance of the FSD system.

Table 5
EER and t-DCF of single systems and primary systems based on the top performance of ASVspoof 2019 LA dataset.

(a) Single systems

System	t-DCF	EER%
CQCC+GMM (B1)	0.2316	9.57
LFCC+GMM (B2)	0.2116	8.09
LFCC-Siamese CNN (Lei, Yang, Liu, & Ye, 2020)	0.0930	3.79
RW-ResNet (Ma, Ren, & Xu, 2021)	0.0820	2.98
ACNN (Ling et al., 2021)	0.0510	1.87
MCG-Res2Net50 (Li, Wu et al., 2021)	0.0520	1.78
FFT-L-SENet (Zhang, Wang et al., 2021)	0.0368	1.14
AASIST (Jung et al., 2022)	0.0347	1.13
SAMO (Ding, Zhang, & Duan, 2023)	0.0356	1.08
PA-Res2Net (Kim & Ban, 2023)	0.0300	1.07
ECANet_SD (Xue et al., 2023)	0.0295	0.88
Ours (single system)	**0.0159**	**0.47**

(b) Fusion systems

System	t-DCF	EER%
T05 (Todisco et al., 2019)	0.0069	0.22
T45 (Lavrentyeva et al., 2019)	0.0510	1.84
T60 (Chettri et al., 2019)	0.0755	2.64
GMM fusion (Tak, Patino, NAutsch, Evans, & Todisco, 2020)	0.0740	2.92
T24 (Todisco et al., 2019)	0.0953	3.45
T50 (Yang et al., 2019)	0.1671	3.56
Ours (single system)	**0.0159**	**0.47**

4.3.4. *Comparison with other systems*

Table 5 shows the results of the eight best-performing single systems, the six main systems, and our best system on the ASVspoof 2019 LA evaluation set. Where B1 and B2 are the baseline systems. The results of single systems are shown in Table 5a. These systems include some top-performance systems from the ASVspoof 2019 challenge and systems from recently published papers. Table 5b shows the results of

Table 6
EER and t-DCF of single systems and primary systems based on the top performance of ASVspoof 2021 LA dataset.
(a) Single systems

System	t-DCF	EER%
B03 (Yamagishi et al., 2021)	0.3445	9.26
B04 (Yamagishi et al., 2021)	0.4257	9.50
B01 (Yamagishi et al., 2021)	0.4974	15.62
B02 (Yamagishi et al., 2021)	0.5758	19.30
Ours (single system)	**0.2642**	**3.61**

(b) Fusion systems

System	t-DCF	EER%
T23 (Tomilov et al., 2021)	0.2177	1.32
T20 (Chen, Khoury, Phatak, & Sivaraman, 2021)	0.2608	3.21
T04 (Cáceres, Font, Grau, & Molina, 2021)	0.2747	5.58
T06 (Kang, Alam, & Fathan, 2021)	0.2853	5.66
Ours (single system)	**0.2642**	**3.61**

the primary systems, where T05, T45, T60, T24, and T50 represent the anonymous identifiers of the teams in the ASVspoof 2019 challenge. These primary systems may contain multiple front-end features and neural network architectures. The GMM fusion system consists of the nonlinear fusion of its six subbands. From the performance comparison of different systems in Table 5, it can be seen that our proposed system achieves state-of-the-art performance in a single system, and also outperforms the second-ranked system on the ASVspoof 2019 LA challenge among the primary systems.

To the best of our knowledge, among all fusion systems, only the T05 system outperforms us. Here we want to emphasize that the fusion system is obtained by fusing multiple single systems, which means that multiple models need to be trained and finally fused, so that the overall model parameters are huge. Moreover, T05 is a fusion of 7 single systems, including 2 Resnet models, 4 MobileNet models, and 1 DenseNet model, and the final results are obtained by combining the equal weights of these 7 single systems. It can be seen that the T05 system architecture is extremely complex. Therefore, our proposed single system has advantages in terms of performance and network architecture.

4.3.5. *t-SNE visualization analysis*

To visualize the effectiveness of the proposed approach, we also visualize the baseline system and my proposed system using t-SNE (van der Maaten & Hinton, 2008), respectively. Both models are trained on the LA dataset in Asvspoof 2019 and take the penultimate layer of the network. As shown in Fig. 9, we can see that the real and fake speech of the Low subband and SENet systems are not distinguished, and there are many blue dots embedded inside the red dots. While the true and false speech of the F0 subband and SR-LA Res2Net systems are separated, there is only a little blending at the boundary. The above visualization results further validate our experimental results.

4.4. *Experimental results on ASVspoof 2021 LA dataset*

Table 6 shows the results of the ASVSpoof 2021 LA Challenge for single and fusion systems. Among them, the T23 (Tomilov et al., 2021) system is a fusion of twelve subsystems, including ten MSTFT-LCNN systems, one MSTFT-ResNet system, and one RawNet system, finally fused in the scoring stage by a fine weight assignment; the T20 (Chen et al., 2021) system is a fusion of three subsystems based on ResNet system with equal weight fusion of scoring; the T04 (Cáceres et al., 2021) system is a scoring fusion of three subsystems, namely LFCC-LCNN, RawNet2, and lightweight TDNN Focal, and all three subsystems use a data enhancement strategy; the T06 (Kang et al., 2021) system is a fusion of eight subsystems, namely an LFCC-LCNN system (baseline), one RawNet2 system (baseline), one LFCC-GMM, four LFCC-SENet systems, and one PSCC-TDNN system, all of which use data enhancement strategies except for the baseline system. B01–B04 are the four baseline systems for the ASVSpoof 2021 LA Challenge. The following points can be observed from the table:

(1) Most of the systems perform data enhancement in the face of transmission interference in the ASVspoof 2021 LA data set, and the performance is poor for the baseline systems that do not perform data enhancement.

(2) Several of the most advanced systems submitted at the ASVspoof 2021 LA challenge are fusion systems, while we can obtain good performance for our single system.

In conclusion, our proposed single system is competitive in the face of both single and fusion systems, which further validates the effectiveness of our proposed method.

4.5. *Comparison with conventional F0 extraction methods*

We extracted traditional 3D additional pitch features as auxiliary features through the Kaldi tool, which was used in combination with the 80-dimensional MFCC, 60-dimensional LFCC, and low subband. As shown in Table 7, 3D denotes the additional three-dimensional pitch features. The results in Table 7 show that the additional pitch features can indeed improve the Kaldi performance of the FSD system, which also verifies that F0 has effective discriminative information. However, the system performance is still limited compared to the F0 subband, which also shows the superiority of the F0 subband features.

5. Discussion

In this section, we discuss the advantages, shortcomings, and future directions of the proposed method in the current research of FSD.

First, our proposed system has a high-performance advantage, because the F0 subband features have a strong discriminative ability to distinguishing between real and fake speech, and the SR-LA Res2Net can further utilize the multi-scale discriminative information of F0 subband to improve the performance of FSD. In addition, from the results of the ASVspoof 2021 LA database, it can be found that our proposed single system can still achieve competitive performance even in a fusion system.

Further, the proposed system also has obvious advantages in real scenarios and practical applications. The first one is to deal with a large amount of data, like about 180,000 entries on the ASVsopoof 2021 LA dataset, which can still be handled efficiently by the proposed system. The second is the challenge of practical deployment, most of the current systems are deployed in automatic speaker verification (ASV), automatic Speech Recognition (ASR), etc., which may require lower computational requirements, and the proposed system is also less than 1MB in model parameters, which is also very convenient to use for ensembling other existing systems. On the other hand, since the system is mostly deployed for ASV and ASR systems, the communication interference encountered in real scenarios is the most common. Thus, the ASVspoof 2021 LA dataset models a large number of communication disturbances, specifically real and spoofing speech transmitted using a variety of telephony systems, including Voice over IP (VoIP) and Public Switched Telephone Network (PSTN).

Finally, we summarize the future directions of FSD systems. (1) Improve the generalization ability of the system in the face of unknown attacks; (2) Enhance the robustness of the system in complex scenarios, such as bandwidth noise, communication interference, and other real-world conditions; (3) Explore the study of lightweight of the system, which can have efficient flexibility in specific applications; (4) Explain the decision-making process of the system in depth. The current research on FSD is not able to do the above points better, which is a future research trend, including the ASVspoof dataset also needs to make more progress for more application scenarios.

Table 7
EER and t-DCF results for conventional pitch features and F0 subbands under different models. F0 (3D) indicates that additional pitch features (probability of voicing, pitch value and delta pitch value) are used together.

	Front-end	Res2Net	LA Res2Net	SR-LA Res2Net
EER(%)	MFCC	9.06	8.26	7.77
	MFCC + F0 (3D)	7.92	7.72	7.39
	LFCC	4.92	2.28	2.05
	LFCC + F0 (3D)	3.76	2.42	1.99
	Low subband	1.85	2.82	2.06
	Low subband + F0 (3D)	1.82	1.53	1.47
	F0 subband (**Ours**)	**1.17**	**0.80**	**0.47**
t-DCF	MFCC	0.2220	0.1825	0.2146
	MFCC + F0 (3D)	0.1836	0.1779	0.2312
	LFCC	0.1358	0.0652	0.0552
	LFCC + F0 (3D)	0.0815	0.0666	0.0538
	Low subband	0.0510	0.0561	0.0577
	Low subband + F0 (3D)	0.0597	0.0455	0.0389
	F0 subband (**Ours**)	**0.0353**	**0.0246**	**0.0159**

6. Conclusions

In this paper, we propose an F0 subband with SR-LA Res2Net for FSD. The F0 distribution of bonafide speech is often difficult to replicate so it is very different from the fake one. Therefore, we think the F0 contains the discriminative information. In addition, to effectively model the F0 subband, we propose a novel SR-LA Res2Net for FSD. Specifically, the SR block is designed to eliminate spatial artifacts when information is transmitted between channel groups. The LA block is used to focus on local information. Experimental results on the ASVspoof 2019 LA dataset show that our proposed approach is very effective against unseen spoofing attacks and achieves a minimum t-DCF of 0.0159 and an EER of 0.47%, which achieves state-of-the-art performance among all single systems. One of the limitations of this work is that we only use the F0 subband for FSD. The other speech information is abandoned, which may also contain some important discriminative information. In the future, to make full use of the speech information, we will explore combining the F0 subband with other speech features to further improve the performance of FSD.

CRediT authorship contribution statement

Cunhang Fan: Writing – review & editing, Validation, Supervision. **Jun Xue:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology. **Jianhua Tao:** Writing – review & editing, Supervision. **Jiangyan Yi:** Methodology, Supervision, Validation, Writing – review & editing. **Chenglong Wang:** Methodology, Validation, Visualization, Writing – review & editing. **Chengshi Zheng:** Writing – review & editing, Methodology, Supervision. **Zhao Lv:** Writing – review & editing, Supervision, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work is supported by the STI 2030—Major Projects (No. 2021ZD0201500), the National Natural Science Foundation of China (NSFC) (No. 62201002, No. 62322120), Excellent Youth Foundation of Anhui Scientific Committee, China (No. 208085J05), Special Fund for Key Program of Science and Technology of Anhui Province, China (No. 202203a07020008), the Open Research Projects of Zhejiang Lab, China (NO.2021KH0AB06).

References

Al-Radhi, M. S., Csapó, T. G., & Németh, G. (2018). A continuous vocoder using sinusoidal model for statistical parametric speech synthesis. In *Speech and computer: 20th international conference, SPECOM 2018, leipzig, Germany, September 18–22, 2018, proceedings 20* (pp. 11–20). Springer.

Ali, M., Sabir, A., & Hassan, M. (2021). Fake audio detection using hierarchical representations learning and spectrogram features. In *2021 international conference on robotics and automation in industry* (pp. 1–6). IEEE.

Cáceres, J., Font, R., Grau, T., & Molina, J. (2021). The biometric vox system for the asvspoof 2021 challenge. In *Proc. 2021 edition of the automatic speaker verification and spoofing countermeasures challenge* (pp. 68–74).

Chen, T., Khoury, E., Phatak, K., & Sivaraman, G. (2021). Pindrop labs' submission to the asvspoof 2021 challenge. In *Proc. 2021 edition of the automatic speaker verification and spoofing countermeasures challenge* (pp. 89–93).

Chen, T., Kumar, A., Nagarsheth, P., Sivaraman, G., & Khoury, E. (2020). Generalization of audio deepfake detection. In *Proc. odyssey 2020 the speaker and language recognition workshop* (pp. 132–137).

Chettri, B., Kinnunen, T., & Benetos, E. (2020). Subband modeling for spoofing detection in automatic speaker verification. In *Proceedings of odyssey 2020: the speaker and language recognition workshop* (pp. 341–348). ISCA.

Chettri, B., Stoller, D., Morfi, V., Ramírez, M., Benetos, E., & Sturm, B. (2019). Ensemble models for spoofing detection in automatic speaker verification. In *Proc. interspeech* (pp. 1018–1022).

Das, R. K., Yang, J., & Li, H. (2019). Long range acoustic features for spoofed speech detection. In *Interspeech* (pp. 1058–1062).

Ding, S., Zhang, Y., & Duan, Z. (2023). SAMO: Speaker attractor multi-center one-class learning for voice anti-spoofing. In *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing* (pp. 1–5). IEEE.

Doan, T.-P., Nguyen-Vu, L., Jung, S., & Hong, K. (2023). BTS-e: Audio deepfake detection using breathing-talking-silence encoder. In *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing* (pp. 1–5). IEEE.

Fan, C., Ding, M., Yi, J., Li, J., & Lv, Z. (2023). Two-stage deep spectrum fusion for noise-robust end-to-end speech recognition. *Applied Acoustics*, *212*, Article 109547.

Fan, C., Xue, J., Dong, S., Ding, M., Yi, J., Li, J., et al. (2023). Subband fusion of complex spectrogram for fake speech detection. *Speech Communication*, *155*, Article 102988.

Fan, C., Zhang, H., Li, A., Xiang, W., Zheng, C., Lv, Z., et al. (2023). CompNet: Complementary network for single-channel speech enhancement. *Neural Networks*, *168*, 508–517.

Gao, S., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., & Torr, P. H. (2019). Res2Net: a new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 652–662.

Hajipour, M., Akhaee, M. A., & Toosi, R. (2021). Listening to sounds of silence for audio replay attack detection. In *2021 7th international conference on signal processing and intelligent systems* (pp. 1–6). IEEE.

He, J., Xu, J., Zhang, L., & Zhu, J. (2023). An interpretive constrained linear model for ResNet and mgnet. *Neural Networks*, *162*, 384–392.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition* (pp. 770–778). IEEE.

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2019). Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(8), 2011–2023.

Huang, B., Cui, S., Huang, J., & Kang, X. (2023). Discriminative frequency information learning for end-to-end speech anti-spoofing. *IEEE Signal Processing Letters*, *30*, 185–189.

Huang, S.-F., Lin, C.-J., Liu, D.-R., Chen, Y.-C., & Lee, H.-y. (2022). Meta-tts: Meta-learning for few-shot speaker adaptive text-to-speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 1558–1571.

Jung, J.-w., Heo, H.-S., Tak, H., Shim, H.-j., Chung, J. S., Lee, B.-J., et al. (2022). Aasist: audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing* (pp. 6367–6371). IEEE.

Kang, W. H., Alam, J., & Fathan, A. (2021). Crim's system description for the asvspoof2021 challenge. In *Proc. 2021 edition of the automatic speaker verification and spoofing countermeasures challenge* (pp. 100–106).

Kim, J., & Ban, S. M. (2023). Phase-aware spoof speech detection based on res2net with phase network. In *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing* (pp. 1–5). IEEE.

Kinnunen, T., Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., et al. (2018). A spoofing benchmark for the 2018 voice conversion challenge: Leveraging from spoofing countermeasures for speech artifact assessment. In *The speaker and language recognition workshop* (pp. 187–194). ISCA.

Kinnunen, T., Sahidullah, M., Delgado, H., Todisco, M., Evans, N., Yamagishi, J., et al. (2017). The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection. In *Proc. interspeech* (pp. 2–6).

Kinnunen, T., Sahidullah, M., Falcone, M., Costantini, L., Hautamäki, R. G., Thomsen, D., et al. (2017). Reddots replayed: a new replay spoofing attack corpus for text-dependent speaker verification research. In *2017 IEEE international conference on acoustics, speech and signal processing* (pp. 5395–5399). IEEE.

Kinnunen, T., Wu, Z.-Z., Lee, K. A., Sedlak, F., Chng, E. S., & Li, H. (2012). Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In *2012 IEEE international conference on acoustics, speech and signal processing* (pp. 4401–4404). IEEE.

Łańcucki, A. (2021). Fastpitch: Parallel text-to-speech with pitch prediction. In *IEEE international conference on acoustics, speech and signal processing* (pp. 6588–6592). IEEE.

Lavrentyeva, G., Tseren, A., Volkova, M., Gorlanov, A., Kozlov, A., & Novoselov, S. (2019). STC antispoofing systems for the AsVspoof2019 challenge. In *Proc. interspeech* (pp. 1033–1037).

Lei, Z., Yang, Y., Liu, C., & Ye, J. (2020). Siamese convolutional neural network using Gaussian probability feature for spoofing speech detection. In *Proc. interspeech* (pp. 1116–1120).

Li, X., Li, N., Weng, C., Liu, X., Su, D., Yu, D., et al. (2021). Replay and synthetic speech detection with Res2Net architecture. In *IEEE international conference on acoustics, speech and signal processing* (pp. 6354–6358). IEEE.

Li, J., Wang, H., He, P., Abdullahi, S. M., & Li, B. (2022). Long-term variable q transform: A novel time-frequency transform algorithm for synthetic speech detection. *Digital Signal Processing*, *120*, Article 103256.

Li, X., Wu, X., Lu, H., Liu, X., & Meng, H. (2021). Channel-wise gated res2net: towards robust detection of synthetic speech attacks. In *Proc. Interspeech 2021*.

Ling, H., Huang, L., Huang, J., Zhang, B., & Li, P. (2021). Attention-based convolutional neural network for ASV spoofing detection. In *Proc. interspeech* (pp. 4289–4293).

Liu, R., Zhang, J., & Gao, G. (2024). Multi-space channel representation learning for mono-to-binaural conversion based audio deepfake detection. *Information Fusion*, *105*, Article 102257.

Lv, Z., Zhang, S., Tang, K., & Hu, P. (2022). Fake audio detection based on unsupervised pretraining models. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing* (pp. 9231–9235). IEEE.

Ma, Y., Ren, Z., & Xu, S. (2021). RW-resnet: a novel speech anti-spoofing model using raw waveform. In *Proc. interspeech* (pp. 4144–4148).

Paul, A., Das, R. K., Sinha, R., & Prasanna, S. M. (2016). Countermeasure to handle replay attacks in practical speaker verification systems. In *2016 international conference on signal processing and communications* (pp. 1–5). IEEE.

Paul, D., Pal, M., & Saha, G. (2017). Spectral features for synthetic speech detection. *IEEE Journal of Selected Topics in Signal Processing*, *11*(4), 605–617.

Qian, K., Jin, Z., Hasegawa-Johnson, M., & Mysore, G. J. (2020). F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6284–6288). IEEE.

Shang, W., & Stevenson, M. (2008). A preliminary study of factors affecting the performance of a playback attack detector. In *2008 Canadian conference on electrical and computer engineering* (pp. 459–464). IEEE.

Shchemelinin, Vadim, & Simonchik, K. (2013). Examining vulnerability of voice verification systems to spoofing attacks by means of a TTS system. In *Proceedings of the 15th international conference on speech and computer-volume 8113* (pp. 132–137).

Sun, T., Ding, S., & Guo, L. (2022). Low-degree term first in ResNet, its variants and the whole neural network family. *Neural Networks*, *148*, 155–165.

Tak, H., Jung, J.-W., Patino, J., Kamble, M., Todisco, M., & Evans, N. (2021). End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. In *ASVSPOoF 2021, automatic speaker verification and spoofing countermeasures challenge* (pp. 1–8). ISCA.

Tak, H., weon Jung, J., Patino, J., Todisco, M., & Evans, N. (2021). Graph attention networks for anti-spoofing. In *Proc. interspeech 2021* (pp. 2356–2360).

Tak, H., Kamble, M., Patino, J., Todisco, M., & Evans, N. (2022). Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing* (pp. 6382–6386). IEEE.

Tak, H., Patino, J., NAutsch, A., Evans, N., & Todisco, M. (2020). Spoofing attack detection using the non-linear fusion of sub-band classifiers. In *Proc. interspeech* (pp. 1106–1110).

Tian, X., Lee, S. W., Wu, Z., Chng, E. S., & Li, H. (2017). An exemplar-based approach to frequency warping for voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(10), 1863–1876.

Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., et al. (2019). ASVspoof 2019: future horizons in spoofed and fake audio detection. In *Proc. interspeech* (pp. 1008–1012).

Tomilov, A., Svishchev, A., Volkova, M., Chirkovskiy, A., Kondratev, A., & Lavrentyeva, G. (2021). STC antispoofing systems for the asvspoof2021 challenge. In *Proc. 2021 edition of the automatic speaker verification and spoofing countermeasures challenge* (pp. 61–67).

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11531–11539).

Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., et al. (2020). Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech and Language*, *64*, Article 101114.

Wei, L., Long, Y., Wei, H., & Li, Y. (2022). New acoustic features for synthetic and replay spoofing attack detection. *Symmetry*, *14*(2), 274.

Williams, J., & Rownicka, J. (2019). Speech replay detection with x-vector attack embeddings and spectral features. In *Proc. Interspeech 2019* (pp. 1053–1057).

Witkowski, M., Kacprzak, S., Zelasko, P., Kowalczyk, K., & Galka, J. (2017). Audio replay attack detection using high-frequency features. In *Interspeech* (pp. 27–31).

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision* (pp. 3–19).

Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilçi, C., Sahidullah, M., et al. (2015). ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Proc. interspeech* (pp. 2037–2041).

Xue, J., Fan, C., Yi, J., Wang, C., Wen, Z., Zhang, D., et al. (2023). Learning from yourself: A self-distillation method for fake speech detection. In *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing* (pp. 1–5). IEEE.

Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., et al. (2021). Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *ASVspoof 2021 workshop-automatic speaker verification and spoofing coutermeasures challenge*.

Yang, J., & Das, R. K. (2020). Long-term high frequency features for synthetic speech detection. *Digital Signal Processing*, *97*, Article 102622.

Yang, J., Das, R. K., & Li, H. (2019a). Significance of subband features for synthetic speech detection. *IEEE Transactions on Information Forensics and Security*, *15*, 2160–2170.

Yang, J., Das, R. K., & Zhou, N. (2019b). Extraction of octave spectra information for spoofing attack detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(12), 2373–2384.

Yang, Y., Wang, H., Dinkel, H., Chen, Z., Wang, S., Qian, Y., et al. (2019). The sjtu robust anti-spoofing systems for the asvspoof 2019 challenge. In *Proc. interspeech* (pp. 1038–1042).

Yi, J., Fu, R., Tao, J., Nie, S., Ma, H., Wang, C., et al. (2022). Add 2022: the first audio deep synthesis detection challenge. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing* (pp. 9216–9220). IEEE.

Zhang, Z., Gu, Y., Yi, X., & Zhao, X. (2022). FMFCC-a: a challenging mandarin dataset for synthetic speech detection. In *Digital forensics and watermarking: 20th international workshop, IWDW 2021, Beijing, China, November 20–22, 2021, revised selected papers* (pp. 117–131). Springer.

Zhang, Y., Jiang, F., & Duan, Z. (2021). One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters*, 937–941.

Zhang, Y., Wang, W., & Zhang, P. (2021). The effect of silence and dual-band fusion in anti-spoofing system. In *Proc. interspeech* (pp. 4279–4283).

Zhang, Z., Yi, X., & Zhao, X. (2021). Fake speech detection using residual network with transformer encoder. In *Proceedings of the 2021 ACM workshop on information hiding and multimedia security* (pp. 13–22).