Speech Communication 155 (2023) 102988

Contents lists available at ScienceDirect

Speech Communication

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

Subband fusion of complex spectrogram for fake speech detection

Cunhang Fan [a], Jun Xue [a], Shunbo Dong [a], Mingming Ding [a], Jiangyan Yi [b], Jinpeng Li [c], Zhao Lv [a,*]

[a] Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China
[b] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[c] Ningbo Institute of Life and Health Industry, University of Chinese Academy of Sciences, China

ARTICLE INFO

Keywords:
Automatic speaker verification
Complex spectrogram
Fake speech detection
Phase information
Subband

ABSTRACT

The phase information was shown useful in fake speech detection. However, the most common reason why phase-based features are not widely used is phase wrapping. This makes the original phase hard to model directly. Therefore, it remains a challenge how to utilize the phase information effectively. To address this issue, this paper proposes a novel subband fusion of the complex spectrogram method for fake speech detection. The complex spectrogram is used as the input feature, containing both amplitude and phase spectrogram. In addition, subbands of the complex spectrogram are modeled separately. The idea is motivated by the fact that each frequency band has a different effect on the fake speech detection task. Finally, to make full use of the subbands, the subband results are fused. Experimental results on the ASVspoof 2019 LA dataset show that our proposed system achieves an equal error rate (EER) of 0.68% and a minimum tandem detection cost function (min t-DCF) of 0.0224.

1. Introduction

Automatic speaker verification (ASV) system aims to confirm the claimed speaker identity from speech utterances. However, the spoofing attacks can fool the ASV systems, such as replay (prerecorded audio), impersonation (mimics or twins), text-to-speech (TTS), and voice conversion (VC) (Wu et al., 2015a). With the development of TTS and VC, they can generate high-quality speech and threaten ASV systems.

In order to address the problem of spoofing attacks, the ASVspoof challenges series have been held in 2015 (Wu et al., 2015b), 2017 (Kinnunen et al., 2017), 2019 (Todisco et al., 2019) and 2021 (Yamagishi et al., 2021). The datasets of the ASVspoof challenge consist of logical access (LA) and physical access (PA). The LA refers to attacks from TTS and VC, and the PA focuses on attacks from replay. This paper focuses on the LA attacks.

To improve the performance of fake speech detection systems, different front-end acoustic features are investigated, such as log power spectrogram (LPS), constant Q cepstral coefficients (CQCC), linear frequency cepstral coefficients (LFCC) and so on Kamble et al. (2020), Pal et al. (2018), Das et al. (2020), Paul et al. (2017) and Wang and Yamagishi (2021). However, the above features are all based on the amplitude spectrogram and lose the phase spectrogram information. Numerous studies (Kulmer and Mowlaee, 2015; Paliwal et al., 2011; Fan et al., 2020; Masuyama et al., 2019; Gurugubelli and Vuppala, 2020; Guo et al., 2022) have shown that the phase spectrogram is very important for speech quality and intelligibility. It is unreasonable to ignore the phase spectrogram. However, the phase spectrum does not contain stable patterns (Eldar et al., 2015; Spoorthi et al., 2019), due to rapidly varying phase changes and phase discontinuities, so we cannot use it directly. Group delay (GD) (Xiao et al., 2015) is proposed as the derivative of the phase spectrum along the frequency axis and acquires a good performance in fake speech detection. Furthermore, in order not to lose phase information, many works apply the raw waveform as the input feature (Jung et al., 2019, 2020; Tak et al., 2021; Ma et al., 2021; Hua et al., 2021; Jung et al., 2022). But compared to time–frequency (T–F) domain-based features, the raw time-domain waveforms contain a wealth of information that requires a powerful model to extract.

To address these issues, this paper proposes a novel subband fusion of the complex spectrogram method for fake speech detection, which utilizes complex spectrograms as the input features. The complex spectrogram contains both amplitude and phase spectrogram, which has more voiceprint information that can be used for fake speech detection. Moreover, several studies (Patel and Patil, 2017; Sahidullah et al., 2015; Sriskandaraja et al., 2016; Witkowski et al., 2017; Garg et al., 2019; Lin et al., 2018; Yang et al., 2019; Soni et al., 2016; Chettri et al., 2020; Ling et al., 2021; Liu et al., 2021; Zhang et al., 2021)

* Corresponding author.
E-mail addresses: cunhang.fan@ahu.edu.cn (C. Fan), kjlz@ahu.edu.cn (Z. Lv).

https://doi.org/10.1016/j.specom.2023.102988
Received 27 September 2022; Received in revised form 5 August 2023; Accepted 20 September 2023
Available online 29 September 2023

have shown that different frequency bands have different effects on the fake speech detection task. Motivated by Zhang et al. (2021), we model different subbands of the complex spectrogram, respectively. Finally, to make full use of the information of different subbands, we fuse their output results. Experimental results on the ASVspoof 2019 LA dataset show that our proposed system achieves an EER of 0.68%. The main contributions of this study can be summarized as follows:

- To the best of our knowledge, this is the first work that applies complex spectrogram to the fake speech detection task, which can exploit both amplitude and phase spectrograms.
- Moreover, we model different subbands of complex spectrogram respectively and fuse their results finally.

The rest of this paper is organized as follows: The related works are presented in Section 2. Section 3 explains our proposed method. Section 4 demonstrates experiments and results. This paper concludes in Section 5.

2. Related works

Artifacts are generally believed to exist in specific subbands (Patel and Patil, 2017; Sahidullah et al., 2015; Li and Horaud, 2019), so many works are focusing on subband effects. The authors in Sriskandaraja et al. (2016) propose three triangular filter bank design approaches, and identify 0–1 kHz, 2.5–5.5 kHz and 7–8 kHz as the most informative subbands. Several studies (Witkowski et al., 2017; Garg et al., 2019; Lin et al., 2018) focus on the high-frequency subband, and it is proven to be more robust against unseen spoofing attacks. Yang et al. propose three novel subbands transform methods (Yang et al., 2019), experimental results show that they perform better than traditional full-band transform methods. Because not all frequency bands are helpful for fake speech detection tasks, subband autoencoder (Soni et al., 2016) or subband modeling framework (Chettri et al., 2020) enable the system to learn subband-specific features. The latest research in subband Zhang et al. (2021) demonstrates high-frequency subbands usually lead to overfitting, while low-frequency subbands are more effective.

There are also some works focus on phase information (Saratxaga et al., 2016; Bharath and Kumar, 2022; Peng et al., 2021). The commonly used phase features are the phase spectrogram (Balamurali et al., 2019) and cosine phase based on short-time Fourier transform, group delay function (GDF) and instantaneous frequency distribution (IFD) (Alsteris and Paliwal, 2007). Saratxaga et al. (2009) propose the relative phase shift (RPS) representation of harmonic phase, which achieves good results when used for synthetic speech detection (De Leon et al., 2011). In Wu et al. (2012), in order to distinguish the converted speech from the bonafide signal, cosine-normalized phase and modified group delay function phase spectrum based features are proposed. Based on GDF, various phase-based features are proposed, such as modified group delay (MGD) (Xiao et al., 2015) and relative phase information (Wang et al., 2015). Recent work proposes a phase network (Kim and Ban, 2022) that processes phase information separately but phase features are still difficult to use effectively currently.

3. Our proposed subband fusion of complex spectrogram

In this section, we will illustrate the process of the complex spectrogram and our proposed subbands fusion algorithm. Fig. 1 indicates the schematic diagram of our proposed method for fake speech detection.

Many of the previous works for fake speech detection tasks lose the phase spectrogram information, which may lose much critical information. To take advantage of the speech information, we apply the complex spectrogram as the input feature that contains both amplitude and phase spectrogram. Moreover, different subbands are modeled respectively. Finally, their results are fused to improve the performance of the fake speech detection systems furthermore.

3.1. Complex spectrogram

The complex spectrogram is applied as the input feature for the fake speech detection task. It can be acquired as follows:

$$\mathbf{X}_r[t,f] + i * \mathbf{X}_i[t,f] = STFT(\mathrm{x}[k]) \tag{1}$$

where the $\mathbf{x}[k]$ denotes the raw speech waveform in the time-domain, k is the time index of speech signals. $STFT$ means the operation of short-time Fourier transformation (STFT), which converts the time-domain speech into the T–F domain. $\mathbf{X}_r \in \mathbb{R}^{F\times T}$ and $\mathbf{X}_i \in \mathbb{R}^{F\times T}$ are the corresponding real and imaginary part of STFT, respectively. t is the index of the time frame and f is the index of the frequency bin. $*$ denotes a multiplication operation.

Then the real $\mathbf{X}_r$ and imaginary $\mathbf{X}_i$ part of STFT are stacked together as the $\mathbf{X}_{complex}$:

$$\mathbf{X}_{complex} = stack(\mathbf{X}_r, \mathbf{X}_i) \in \mathbb{R}^{2\times F\times T} \tag{2}$$

where the $stack(*)$ means the stack operation. F and T are the number of frequency bin and time frame, respectively.

Different from the complex spectrogram, the log power spectrogram (LPS) loses the phase information:

$$\mathbf{X}_{LPS} = log\sqrt{(\mathbf{X}_r)^2 + (\mathbf{X}_i)^2} \in \mathbb{R}^{F\times T} \tag{3}$$

where $\mathbf{X}_{LPS}$ represents LPS feature.

3.2. Other phase features

To compare the performance of other phase features, we used phase angle (PA), group delay (GD) and modified group delay (MGD).

3.2.1. Phase angle

The phase angle is applied as an input feature to the fake speech detection task. It can be acquired as follows:

$$\mathbf{X}_{PA} = \tan^{-1}\left(X_i/X_r\right) \tag{4}$$

where $\mathbf{X}_{PA}$ represents PA feature.

3.2.2. Group delay

We also use the group delay function (Hegde et al., 2004) as a feature, which is represented as follows

$$\mathbf{X}_{GD} = \frac{X_r * Y_r + X_i * Y_i}{|X|^2} \tag{5}$$

The X_r and X_i represent the real and imaginary parts of X, respectively. Y_r and Y_i are the real and imaginary parts of the Fourier transform spectrum of $kx(k)$, respectively. $\mathbf{X}_{GD}$ is the group delay (GD) feature.

3.2.3. Modified group delay

The modified group delay function (Wu et al., 2013) is defined as follows:

$$\tau_\rho = \frac{X_r * Y_r + Y_i * X_i}{|S|^{2\rho}} \tag{6}$$

$$\mathbf{X}_{MGD} = \frac{\tau_\rho}{\left|\tau_\rho\right|}\left|\tau_\rho\right|^\gamma \tag{7}$$

$|S|^2$ is the smoothed version to X. ρ and γ are in $(0, 1]$ hyperparameters. $\mathbf{X}_{MGD}$ is the modified group delay (MGD) feature.

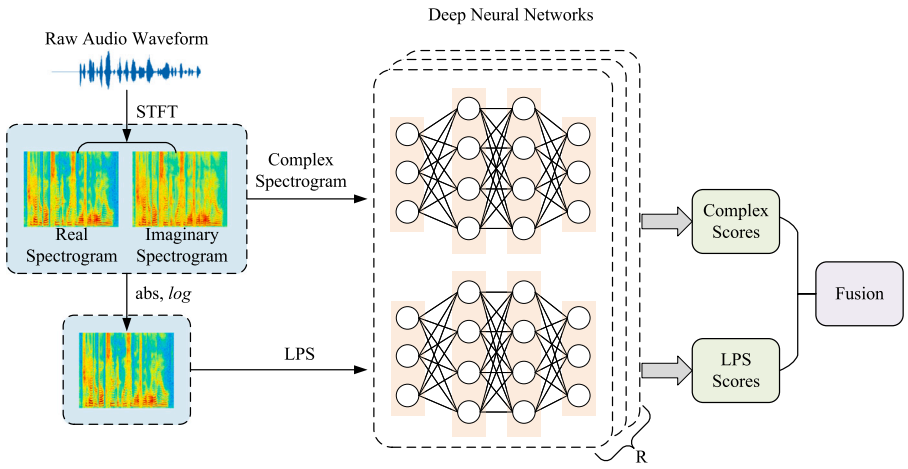

Fig. 1. The schematic diagram of our proposed method for fake speech detection.

3.3. Subband fusion of complex spectrogram

It has been demonstrated that different frequency bands have different effects on the fake speech detection task. In this paper, we divide the complex spectrogram into two subbands, namely low subband (0–4 kHz) $\mathbf{X}^{L}_{complex}$ and high subband (4–8 kHz) $\mathbf{X}^{H}_{complex}$. Same as the complex spectrogram, the LPS is also divided into low and high subbands: $\mathbf{X}^{L}_{LPS}$ and $\mathbf{X}^{H}_{LPS}$.

Motivated by Zhang et al. (2021), we use the squeeze-and-excitation ResNet (SENet) (Hu et al., 2018) as the classifier for different subbands. These different subbands are fed into each SENet to acquire their scores. To take full advantage of the information of each subband, we propose a two-stage fusion algorithm and further improve the performance of fake speech detection systems. Firstly, we fuse the $\mathbf{X}^{L}_{complex}$ and $\mathbf{X}^{H}_{complex}$ at score-level:

$$S_{complex} = \alpha * S^{L}_{complex} + (1-\alpha) * S^{H}_{complex} \tag{8}$$

where the $S^{L}_{complex}$ and $S^{H}_{complex}$ denote the scores of $\mathbf{X}^{L}_{complex}$ and $\mathbf{X}^{H}_{complex}$, respectively. The $S_{complex}$ is the fusion score of dual-band complex spectrogram. α is the weight of the first-stage fusion.

Furthermore, since the low subband Zhang et al. (2021) of LPS is more effective for the fake speech detection task. We apply low subband LPS and perform the second-stage fusion as follows:

$$S = \beta * S_{complex} + (1-\beta) * S^{L}_{LPS} \tag{9}$$

where S is the final fusion score, S^{L}_{LPS} is the low subband score of LPS. β is the weight of the second-stage fusion.

4. Experiments and results

4.1. Dataset

ASVspoof 2019 LA dataset: We conduct our experiments on the ASVspoof 2019 LA database, which consists of three parts: training, development and evaluation sets. The sampling rate of all generated speech waveforms is 16000 Hz and 16-bit quantization. As for the LA subset, the spoofing attacks are generated by various TTS and VC algorithms. The training and development subsets contain the same 6 attacks (A01–A06). There are 2 VC and 4 TTS algorithms. The evaluation set contains 13 attacks (7 TTS and 6 VC), consisting of 2 known algorithms and 11 unseen algorithms.

ASVspoof 2021 LA dataset: Unlike the ASVspoof 2019 LA dataset, the ASVspoof 2021 LA dataset has only the evaluation set, and we still need the data from ASVspoof 2019 LA for training. In addition, the ASVspoof 2021 LA dataset introduces communication interference based on different bandwidths and codecs with about 180,000 speech. These interferences have the potential to obscure the distinguishing information of the speech, further adding to the challenge.

In this work, to evaluate the results of different fake speech detection systems, the EER and the minimum tandem detection cost function (min t-DCF) (Kinnunen et al., 2020) are used as the evaluation metrics. The EER is the operating point where the false acceptance rate (FAR) meets the false rejection rate (FRR).

4.2. Experimental setup

As for the STFT operation, the length of the Blackman window is 1728 with 130 hop length. Therefore, the dimension of the spectrogram is 865. To form batches, the time frames are fixed to 600 by truncating or concatenating. The first 0–433 dimension is used as the low subband, and the last 433–865 dimension is used as the high subband. Therefore, the shape of the low subband and the high subband are set to 433×600 and 432×600, respectively. To reduce the impact from communication interference, we used data augmentation when training the model used for the evaluation of the ASVspoof 2021 LA dataset. Specifically, we introduced the RawBoost (Tak et al., 2022) data enhancement method, using a combination of linear and nonlinear noise and impulse signal independent noise.

Same as Zhang et al. (2021), the SENet34 (Hu et al., 2018) is applied as the classifier for different subbands. Moreover, to verify the effectiveness of features, we also used the other network to model, which are Light-CNN (LCNN) (Lavrentyeva et al., 2017) and Attention-CNN (ACNN) (Ling et al., 2021). All settings are the same as above except for the network. All experiments in this paper were run three times and the best results were taken.

We utilize the angular margin-based softmax (A-softmax) (Liu et al., 2017) as loss function. The batch size is 64, and the initial learning rate is 0.0001. For the first 1000 warm-up steps, the learning rate increases linearly and decreases proportionally to the inverse square root of the number of steps. In addition, we use the Adam as our optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$ and weight decay 10^{-4}. The number of epochs is 32. The weights of the first-stage and second-stage fusion α and β are set to 0.5.

4.3. Experimental results

4.3.1. Analysis of subband fusion results

Table 1 shows the EER results of our proposed second-stage fusion system. Table 2 shows the EER results fused between different subbands, calculated separately for each attack algorithm. Due to space limitations, the Table 2 shows the results based on the ACNN network. From Tables 1 and 2 we can derive the following:

Table 1
The EER (%) for our proposed second stage fusion systems on the ASVspoof 2019 LA dataset. "+" denotes the fusion operation.

Systems	EER(SENet)	EER(LCNN)	EER(ACNN)
PA(L+H)+LPS(L)	0.95	2.21	1.03
Complex(L+H)+LPS(L) Ours	**0.70**	1.32	**0.68**
GD(L+H)+LPS(L)	1.15	**1.29**	1.34
MGD(L+H)+LPS(L)	1.11	1.53	1.04
PA(L+H)+LPS(Full)	1.15	4.08	2.36
Complex(L+H)+LPS(Full)	1.46	3.92	3.17
GD(L+H)+LPS(Full)	2.46	4.52	4.11
MGD(L+H)+LPS(Full)	2.14	4.52	2.75
PA(L+H)+LPS(L+H)	0.96	1.82	1.40
Complex(L+H)+LPS(L+H)	1.15	2.70	2.36
GD(L+H)+LPS(L+H)	1.11	2.66	1.37
MGD(L+H)+LPS(L+H)	1.04	3.03	1.06

Table 2
Separately computed EER (%) results based ACNN for evaluation (A07–A19) subsets on the ASVspoof 2019 LA dataset. Where "F1, F2, F3, and F4" denote the PA(L+H)+LPS(L), Complex(L+H)+LPS(L), GD(L+H)+LPS(L), and MGD(L+H)+LPS(L), respectively.

Systems	Evaluation set												
	Seen attacks		Unseen attacks										
	A16	A19	A07	A08	A09	A10	A11	A12	A13	A14	A15	A17	A18
F1	0.34	1.19	0.10	1.99	0.08	0.32	0.42	0.20	0.13	0.09	0.26	2.28	0.73
F2	0.16	0.42	0.01	0.81	0.00	0.40	0.36	0.16	0.18	0.05	0.20	1.78	0.97
F3	0.57	1.42	0.24	2.50	0.05	0.58	0.51	0.32	0.13	0.13	0.46	2.35	1.91
F4	0.36	1.07	0.04	2.23	0.02	0.36	0.34	0.12	0.17	0.08	0.24	1.98	1.28

(1) Table 1 shows the fusion results based on different subbands and different phase features of LPS. Overall, LPS-based low subband and phase signatures give the best results. This is because the low subband modeling performance of LPS is inherently better than the result of full subband and high and low subband fusion. In addition, we also utilize several existing networks for modeling, among which the ACNN-based network has the best results. This is because the ACNN network contains frequency attention and channel attention modules. When modeling phase features and amplitude spectrum features, it can pay more attention to the essential information of different features, and the final fusion stage can be more complementary. Among them, Complex (L + H) + LPS (L) performed the best, and its EER result reached 0.68%.

(2) Table 2 shows the fusion results of phase features and the LPS low subband for specific individual attacks. For visible attacks, especially the A19 attack, the performance of PA, GD, and MGD features is significantly inferior to that of complex spectral features. For invisible attacks, the A08 and A17 algorithms are the most difficult to detect. However, the EERs of the complex spectral features in the A08 and A17 attacks are 0.81% and 1.78% respectively, which are also significantly better than other phase features. This may be because the complex spectrum contains both amplitude information and phase information, and has more discriminative information.

4.3.2. Analysis of different subband results

Fig. 2 shows the single system results (EER and t-DCF) based on different networks. "L" and "H" denote the low and high subbands, respectively. "Full" means applying the full frequency bands. Table 3 shows the first-stage fusion results for disjoint subbands between their individual features. Table 4 shows the EER results for each subband, calculated separately for each attack algorithm. From Fig. 2, Tables 3 and 4 we can draw the following points:

(1) Regardless of whether it is LPS or PA, etc., the high subband has worse performance than the low subband. Specifically, the complex spectrum low subband based on SENet has an EER of 5.36%, but the high subband is 20.00%. As for LPS, the EER of the low subband is 1.14%, but the high subband is 14.10%. The reason is that although the high-frequency features are more discriminative because of the worse generation for TTS and VC on high frequency, this may lead to overfitting.

Table 3
The results of EER (%) for our proposed first stage fusion systems on the ASVspoof 2019 LA dataset. "+" denotes the fusion operation.

Systems	EER(SENet)	EER(LCNN)	EER(ACNN)
LPS(L+H)	1.16	2.42	4.63
PA(L+H)	3.17	5.91	3.64
PA(L)+LPS(H)	3.41	5.28	6.97
PA(H)+ LPS(L)	1.00	2.70	1.37
Complex(L+H)	3.67	5.00	5.03
Complex(L)+LPS(H)	2.95	4.66	5.51
Complex(H)+LPS(L)	1.56	2.95	3.33
GD(L+H)	5.16	6.37	6.89
GD(L)+LPS(H)	3.38	4.43	7.05
GD(H)+LPS(L)	1.64	4.18	2.61
MGD(L+H)	5.69	6.36	4.47
MGD(L)+LPS(H)	3.15	4.93	7.69
MGD(H)+LPS(L)	1.65	3.73	1.46

(2) Compared with the full band system, fusing the low and high subbands for complex spectrograms can improve the performance of fake speech detection systems. For example, the min t-DCF and EER of "Complex(Full)" are 0.2301 and 9.34%, but for "Complex(L + H)", they are 0.1212 and 3.67%, respectively. These results indicate that different frequency bands have different effects on the fake speech detection task, and dealing with different subbands respectively is beneficial to the fake speech detection task.

(3) For the LPS feature, the EER results are 1.14% and 1.16% for LPS(L) and LPS(L + H), respectively, but LPS(L) + PA(H)'s EER result is 1.00%. The reason is that the performance of LPS(L) is already excellent, and LPS(H) may not be able to obtain further gain effect by compensating high-frequency information only. However, PA(H) can compensate for both phase information and high-frequency information. This indicates that both the high frequency and phase of the features contain effective information to distinguish between real and fake speech.

(4) As can be seen from Table 1, the amplitude-based LPS features are still the most discriminative, but the phase itself is irregular, so the phase-based features still perform poorly when modeled alone. It is well known that A17 (waveform filtering) is the most notorious attack algorithm. However, among the four based-phase features, the complex spectrum and MGD features of the fusion system perform better for the A17 attack. But, the MGD feature performs significantly worse for the visible attacks A16 and A19. Taken together, we conclude that the complex spectrum feature can be well modeled and has good generalization ability.

4.3.3. Comparison with other systems

To evaluate the performance of our proposed method, we compared the proposed system with other state-of-the-art systems on the evaluation dataset of the ASVspoof 2019 LA database. Besides the top-performance systems (T05, T45 and T60) of ASVspoof 2019 challenge, the other recent published novel systems are also compared, such as subband models (Zhang et al., 2021), raw waveform based models (Tak et al., 2021; Ma et al., 2021) and frequency attention model (Ling et al., 2021).

Table 5 shows the performance comparison with other systems. We can observe that our proposed complex subbands fusion system achieves an EER of 0.70%, which outperforms the second-ranked system (EER 1.12%) among all known systems. In terms of T45, T60, Ling et al. and FFT-L-SENet, features of the above systems are based on the amplitude spectrogram, which loses the phase information. Although the RW-Resnet and RawNet2 systems are based on the raw waveform without the phase information lost, their performance is worse than

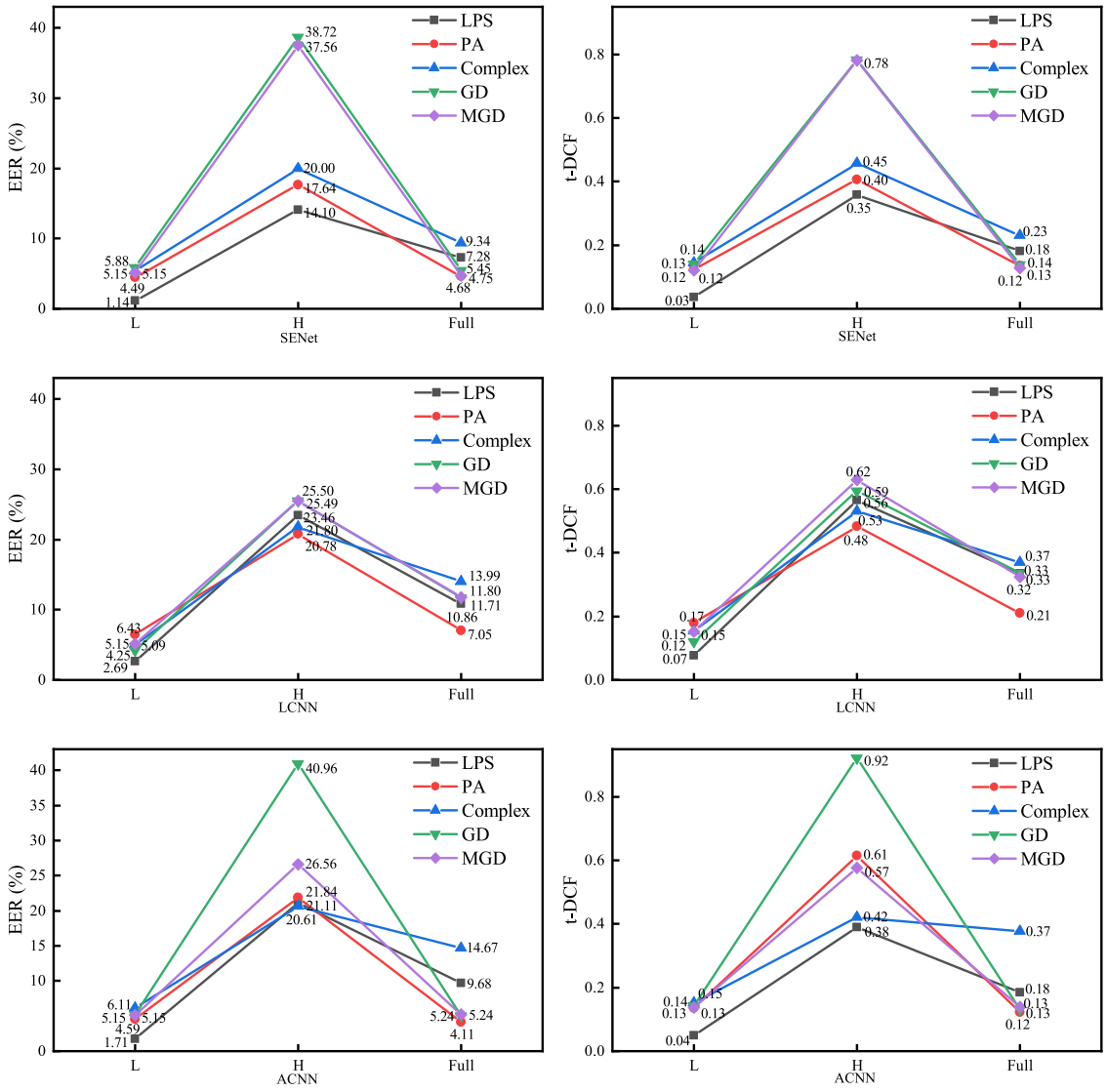

Fig. 2. The results of min t-DCF and EER for our proposed different single systems on ASVspoof 2019 LA dataset. "L" and "H" denote the low and high subbands, respectively. "Full" means applying the full frequency bands.

Table 4
Separately computed EER (%) results based ACNN for evaluation (A07-A19) subsets on the ASVspoof 2019 LA dataset. For visualization purposes, the poorer performing cells have a darker gray background.

Systems	Evaluation set												
	Seen attacks		Unseen attacks										
	A16	A19	A07	A08	A09	A10	A11	A12	A13	A14	A15	A17	A18
LPS_L	0.6	1.9	0.2	4.3	0.1	0.5	0.6	0.2	0.1	0.1	0.4	3.1	1.32
LPS_H	0.0	0.1	0.0	0.4	0.1	10.2	87.66	25.3	47.9	3.4	11.4	43.0	7.4
LPS_Full	0.0	0.0	1.1	4.0	0.2	1.7	1.5	1.4	0.6	1.1	1.1	5.8	23.2
Complex_L	1.7	9.9	1.1	4.0	0.2	1.7	1.5	1.4	0.6	1.1	1.1	5.8	23.2
Complex_H	0.3	0.3	0.0	0.8	0.3	15.6	23.5	39.3	61.7	6.4	19.3	38.8	17.3
Complex_Full	0.0	0.0	0.0	0.0	0.4	9.8	26.9	25.1	19.8	0.8	15.6	28.3	14.7
PA_L	2.7	9.5	1.2	4.3	1.1	1.8	2.0	1.8	1.4	1.3	2.3	12.3	5.0
PA_H	10.2	16.5	4.3	8.5	17.6	28.4	32.4	38.8	28.6	16.5	13.4	36.4	14.1
PA_Full	1.2	3.8	0.6	1.3	1.5	3.0	2.5	3.2	1.6	1.4	3.5	13.5	5.7
GD_L	1.3	4.9	0.5	2.7	0.2	1.7	2.2	1.6	0.2	0.4	1.0	4.2	19.1
GD_H	53.6	43.2	39.8	43.1	46.5	48.3	22.7	50.1	35.2	39.3	50.2	41.5	42.5
GD_Full	1.5	4.7	0.5	4.4	0.3	1.6	2.0	1.3	0.3	0.7	1.4	5.7	21.5
MGD_L	1.4	4.9	0.9	4.3	0.3	2.2	2.7	1.7	0.8	0.9	1.8	7.2	20.4
MGD_H	30.2	37.5	11.0	12.0	3.6	20.5	6.2	36.1	64.6	9.8	20.5	43.9	25.4
MGD_Full	1.5	5.5	0.9	4.2	0.4	2.4	3.0	1.9	0.6	0.8	1.5	6.8	20.7

our proposed method. The results verify that our proposed complex subbands fusion system is quite effective for fake speech detection. Because the front-end feature is based on the complex spectrogram, which can fully use all the speech information.

Table 5
Comparison with other systems on the evaluation set of the ASVspoof 2019 LA database.

Single systems	t-DCF	EER%
FFT-L-SENet (Zhang et al., 2021)	0.0368	1.14
Ling et al. (2021)	0.0510	1.87
GMM-LFCC (Tak et al., 2020)	0.0904	3.50
RawNet2 (Tak et al., 2021)	0.1294	4.66
Fusion systems	**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 et al., 2020)	0.0740	2.92
FFT dual-band fusion (Zhang et al., 2021)	0.0498	1.56
RW-Resnet (Ma et al., 2021)	0.0820	2.98
Complex(L+H)+LPS(L) (ours)	**0.0224**	**0.68**

Table 6
The EER (%) for our proposed second stage fusion systems on the ASVspoof 2021 LA dataset. "+" denotes the fusion operation.

Systems	EER(SENet)	EER(LCNN)	EER(ACNN)
PA(L+H)+LPS(L)	15.71	17.90	18.71
Complex(L+H)+LPS(L) Ours	**5.99**	9.49	6.88
GD(L+H)+LPS(L)	6.78	11.31	9.65
MGD(L+H)+LPS(L)	7.20	10.62	10.17
PA(L+H)+LPS(Full)	26.00	30.70	29.07
Complex(L+H)+LPS(Full)	7.51	11.43	11.10
GD(L+H)+LPS(Full)	11.47	16.18	14.76
MGD(L+H)+LPS(Full)	11.76	16.17	15.73
PA(L+H)+LPS(L+H)	18.09	19.91	20.64
Complex(L+H)+LPS(L+H)	6.54	10.57	7.59
GD(L+H)+LPS(L+H)	11.47	16.18	14.76
MGD(L+H)+LPS(L+H)	8.41	11.83	12.18

Table 7
Comparison of experimental results based on SENet in ASVspoof 2021 LA database with other systems. DA denotes data augmentation.

Systems	EER	t-DCF
CQCC-GMM (Yamagishi et al., 2021)	15.62	0.4974
LFCC-GMM (Yamagishi et al., 2021)	19.30	0.5758
LFCC-LCNN (Yamagishi et al., 2021)	9.26	0.3445
RawNet2 (Yamagishi et al., 2021)	9.50	0.4257
LPS(L)	14.43	0.3688
Complex(L)	6.76	0.3166
Complex(H)	41.79	0.8859
Complex(L+H)	6.58	0.3154
Complex(L+H)+LPS(L)	**5.99**	**0.3081**
LPS(L)(DA)	5.16	0.3148
Complex(L)(DA)	7.04	0.3342
Complex(H)(DA)	19.51	0.4737
Complex(L+H)(DA)	7.00	0.3304
Complex(L+H)+LPS(L)(DA)	**3.98**	**0.2794**

We also have an advantage over the state-of-the-art system (T05) (Nautsch et al., 2021), which is obtained by fusing seven single systems, including two ResNet systems, four MobileNet systems, and one DenseNet system. Therefore, our proposed "Complex(L + H) + LPS(L)" system has only three systems fused, which shows that the number of fused systems can be further reduced by making full use of the voice information, and also has good performance.

4.3.4. Analysis of the results in the ASVspoof 2021 LA dataset

To further validate the effectiveness of our approach, we evaluated it on the ASVspoof 2021 LA dataset. Table 6 shows the results of the system with different feature fusions. Table 7 shows the results of our proposed system and the benchmark system. The first four rows are the results of the benchmark system. From Table 7, we can see the following two points: (1) the results of ASVspoof 2021 LA are worse than those of ASVspoof 2019 LA due to the introduction of communication interference in the dataset, which may mask the distinguishing information between real and false speech; (2) our proposed method is still valid on the ASVspoof 2021 LA dataset set, where the EER of the "LPS(L)" system is 14.43% and t-DCF is 0.3688, while the EER and t-DCF of the "Complex(L + H) + LPS(L)" system are 5.99% and 0.3081, respectively, which show a significant improvement in the system performance. This further proves that our method can be generalized to other datasets and is very effective. In addition, data augmentation can effectively improve the generalization of the system, and the EER of the "Complex(L + H) + LPS(L)" system can reach 3.98%.

5. Conclusions

In order to make full use of speech information, this paper proposes a novel complex spectrogram subbands fusion method for fake speech detection. We apply the complex spectrogram as the input feature, containing both amplitude and phase spectrogram information. In addition, different subbands are modeled respectively. Finally, the two-stage fusion algorithm is applied to improve the performance of fake speech detection further. The experimental results on the ASVspoof 2019 LA dataset show that our proposed method achieves a min t-DCF of 0.0224 and an EER of 0.68%, second only to the T05 system. In the future, we will explore frequency attention with complex spectrogram to automatically fuse each frequency band.

CRediT authorship contribution statement

Cunhang Fan developed the model, planned as well as performed all the experiments. Cunhang Fan also wrote the main part of the manuscript. Jun Xue, Shunbo Dong, Mingming Ding, Yijiang Yan, Jinpeng Li and Jianhua Tao took part in the development of the model, planned the experiments, analyzed the results. Shunbo Dong, Yijiang Yan and Zhao Lv took part in planned the experiments and analyzed the results. Jun Xue and Shunbo Dong also participated in the coordination of the study and reviewed the manuscript. Shunbo Dong and Mingming Ding reviewed the manuscript. All authors read and approved the final manuscript.

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 National Key Research and Development Program of China (No. 2021ZD0201502), the National Natural Science Foundation of China (NSFC) (No. 61972437, No. 62201002), Excellent Youth Foundation of Anhui Scientific Committee, China (No. 2208085J05), 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) and the Open Projects Program of National Laboratory of Pattern Recognition, China (NO. 202200014).

References

Alsteris, L.D., Paliwal, K.K., 2007. Short-time phase spectrum in speech processing: A review and some experimental results. Digit. Signal Process. 17 (3), 578–616.

Balamurali, B., Lin, K.E., Lui, S., Chen, J.-M., Herremans, D., 2019. Toward robust audio spoofing detection: A detailed comparison of traditional and learned features. IEEE Access 7, 84229–84241.

Bharath, K., Kumar, M.R., 2022. Replay spoof detection for speaker verification system using magnitude-phase-instantaneous frequency and energy features. Multimedia Tools Appl. 81 (27), 39343–39366.

Chettri, B., Kinnunen, T., Benetos, E., 2020. Subband modeling for spoofing detection in automatic speaker verification. In: Proc. Odyssey 2020 the Speaker and Language Recognition Workshop. pp. 341–348.

Chettri, B., Stoller, D., Morfi, V., Ramírez, M.A.M., Benetos, E., Sturm, B.L., 2019. Ensemble models for spoofing detection in automatic speaker verification. In: Proc. Interspeech 2019. pp. 1018–1022.

Das, R.K., Yang, J., Li, H., 2020. Assessing the scope of generalized countermeasures for anti-spoofing. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6589–6593.

De Leon, P.L., Hernaez, I., Saratxaga, I., Pucher, M., Yamagishi, J., 2011. Detection of synthetic speech for the problem of imposture. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4844–4847.

Eldar, Y.C., Sidorenko, P., Mixon, D.G., Barel, S., Cohen, O., 2015. Sparse phase retrieval from short-time Fourier measurements. IEEE Signal Process. Lett. 22 (5), 638–642.

Fan, C., Tao, J., Liu, B., Yi, J., Wen, Z., Liu, X., 2020. End-to-end post-filter for speech separation with deep attention fusion features. IEEE/ACM Trans. Audio Speech Lang. Process. 28, 1303–1314.

Garg, S., Bhilare, S., Kanhangad, V., 2019. Subband analysis for performance improvement of replay attack detection in speaker verification systems. In: 2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA). IEEE, pp. 1–7.

Guo, L., Wang, L., Dang, J., Chng, E.S., Nakagawa, S., 2022. Learning affective representations based on magnitude and dynamic relative phase information for speech emotion recognition. Speech Commun. 136, 118–127.

Gurugubelli, K., Vuppala, A.K., 2020. Analytic phase features for dysarthric speech detection and intelligibility assessment. Speech Commun. 121, 1–15.

Hegde, R.M., Murthy, H.A., Gadde, V., 2004. The modified group delay feature: a new spectral representation of speech. In: Proceedings of 8th International Conference on Spoken Language Processing (INTERSPEECH'04), Vol. 2. pp. 913–916.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Hua, G., Teoh, A.B.J., Zhang, H., 2021. Towards end-to-end synthetic speech detection. IEEE Signal Process. Lett. 28, 1265–1269.

Jung, J.-w., Heo, H.-S., Tak, H., Shim, H.-j., Chung, J.S., Lee, B.-J., Yu, H.-J., Evans, N., 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 (ICASSP). IEEE, pp. 6367–6371.

Jung, J.-w., Kim, S.-b., Shim, H.-j., Kim, J.-h., Yu, H.-J., 2020. Improved RawNet with feature map scaling for text-independent speaker verification using raw waveforms. In: Proc. Interspeech 2020. pp. 1496–1500.

Jung, J.-w., Shim, H.-j., Heo, H.-S., Yu, H.-J., 2019. Replay attack detection with complementary high-resolution information using end-to-end DNN for the ASVspoof 2019 challenge. In: Proc. Interspeech 2019. pp. 1083–1087.

Kamble, M.R., Tak, H., Patil, H.A., 2020. Amplitude and frequency modulation-based features for detection of replay spoof speech. Speech Commun. 125, 114–127.

Kim, J., Ban, S.M., 2022. Phase-aware spoof speech detection based on Res2Net with phase network. arXiv preprint arXiv:2203.10793.

Kinnunen, T., Delgado, H., Evans, N., Lee, K.A., Vestman, V., Nautsch, A., Todisco, M., Wang, X., Sahidullah, M., Yamagishi, J., Reynolds, D.A., 2020. Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals. IEEE/ACM TASLP 28, 2195–2210.

Kinnunen, T., Sahidullah, M., Delgado, H., Todisco, M., Evans, N., Yamagishi, J., Lee, K.A., 2017. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In: Proc. Interspeech 2017. pp. 2–6.

Kulmer, J., Mowlaee, P., 2015. Phase estimation in single channel speech enhancement using phase decomposition. IEEE Signal Process. Lett. 22 (5), 598–602.

Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Kudashev, O., Shchemelinin, V., 2017. Audio replay attack detection with deep learning frameworks. In: Interspeech. pp. 82–86.

Lavrentyeva, G., Novoselov, S., Tseren, A., Volkova, M., Gorlanov, A., Kozlov, A., 2019. STC antispoofing systems for the asvspoof2019 challenge. In: Proc. Interspeech 2019. pp. 1033–1037.

Li, X., Horaud, R., 2019. Narrow-band deep filtering for multichannel speech enhancement. arXiv preprint arXiv:1911.10791.

Lin, L., Wang, R., Diqun, Y., 2018. A replay speech detection algorithm based on sub-band analysis. In: International Conference on Intelligent Information Processing. Springer, pp. 337–345.

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

Liu, M., Wang, L., Lee, K.A., Chen, X., Dang, J., 2021. Replay-attack detection using features with adaptive spectro-temporal resolution. In: ICASSP 2021. IEEE, pp. 6374–6378.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L., 2017. SphereFace: Deep hypersphere embedding for face recognition. In: CVPR 2017. pp. 6738–6746.

Ma, Y., Ren, Z., Xu, S., 2021. RW-resnet: A novel speech anti-spoofing model using raw waveform. In: Proc. Interspeech 2021. pp. 4144–4148.

Masuyama, Y., Yatabe, K., Oikawa, Y., 2019. Griffin–lim like phase recovery via alternating direction method of multipliers. IEEE Signal Process. Lett. 26 (1), 184–188.

Nautsch, A., Wang, X., Evans, N., Kinnunen, T.H., Vestman, V., Todisco, M., Delgado, H., Sahidullah, M., Yamagishi, J., Lee, K.A., 2021. ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech. IEEE Trans. Biometr. Behav. Identity Sci. 3 (2), 252–265.

Pal, M., Paul, D., Saha, G., 2018. Synthetic speech detection using fundamental frequency variation and spectral features. Comput. Speech Lang. 48, 31–50.

Paliwal, K., Wójcicki, K., Shannon, B., 2011. The importance of phase in speech enhancement. Speech Commun. 53 (4), 465–494.

Patel, T.B., Patil, H.A., 2017. Significance of source–filter interaction for classification of natural vs. spoofed speech. IEEE J. Sel. Top. Sign. Proces. 11 (4), 644–659.

Paul, D., Sahidullah, M., Saha, G., 2017. Generalization of spoofing countermeasures: A case study with asvspoof 2015 and BTAS 2016 corpora. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 2047–2051.

Peng, J., Qu, X., Gu, R., Wang, J., Xiao, J., Burget, L., Cernockỳ, J., 2021. Effective phase encoding for end-to-end speaker verification. In: Interspeech. pp. 2366–2370.

Sahidullah, M., Kinnunen, T., Hanilçi, C., 2015. A comparison of features for synthetic speech detection. In: Proc. Interspeech 2015. pp. 2087–2091.

Saratxaga, I., Hernaez, I., Erro, D., Navas, E., Sanchez, J., 2009. Simple representation of signal phase for harmonic speech models. Electron. Lett. 45 (7), 381–383.

Saratxaga, I., Sanchez, J., Wu, Z., Hernaez, I., Navas, E., 2016. Synthetic speech detection using phase information. Speech Commun. 81 (C), 30–41.

Soni, M.H., Patel, T.B., Patil, H.A., 2016. Novel subband autoencoder features for detection of spoofed speech. In: Interspeech 2016. pp. 1820–1824.

Spoorthi, G.E., Gorthi, S., Gorthi, R.K.S.S., 2019. PhaseNet: A deep convolutional neural network for two-dimensional phase unwrapping. IEEE Signal Process. Lett. 26 (1), 54–58.

Sriskandaraja, K., Sethu, V., Le, P.N., Ambikairajah, E., Investigation of sub-band discriminative information between spoofed and genuine speech. In: Interspeech 2016. 1710–1714.

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 (ICASSP). IEEE, pp. 6382–6386.

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 2020. pp. 1106–1110.

Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N., Larcher, A., 2021. End-to-end anti-spoofing with RawNet2. In: ICASSP 2021. pp. 6369–6373.

Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T.H., Lee, K.A., 2019. ASVspoof 2019: Future horizons in spoofed and fake audio detection. In: Proc. Interspeech 2019. pp. 1008–1012.

Wang, X., Yamagishi, J., 2021. Investigating self-supervised front ends for speech spoofing countermeasures. arXiv preprint arXiv:2111.07725.

Wang, L., Yoshida, Y., Kawakami, Y., Nakagawa, S., 2015. Relative phase information for detecting human speech and spoofed speech. In: Proc. Interspeech 2015. pp. 2092–2096.

Witkowski, M., Kacprzak, S., Żelasko, P., Kowalczyk, K., Gałka, J., 2017. Audio replay attack detection using high-frequency features. In: Proc. Interspeech 2017. pp. 27–31.

Wu, Z., Chng, E.S., Li, H., 2012. Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In: Thirteenth Annual Conference of the International Speech Communication Association.

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., Li, H., 2015a. Spoofing and countermeasures for speaker verification: A survey. Speech Commun. 66, 130–153.

Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilçi, C., Sahidullah, M., Sizov, A., 2015b. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In: Proc. Interspeech 2015. pp. 2037–2041.

Wu, Z., Xiao, X., Chng, E.S., Li, H., 2013. Synthetic speech detection using temporal modulation feature. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 7234–7238.

Xiao, X., Tian, X., Du, S., Xu, H., Chng, E.S., Li, H., 2015. Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 challenge. In: Proc. Interspeech 2015. pp. 2052–2056.

Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., Liu, X., Lee, K.A., Kinnunen, T., Evans, N., Delgado, H., 2021. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. In: Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. pp. 47–54.

Yang, J., Das, R.K., Li, H., 2019. Significance of subband features for synthetic speech detection. IEEE Trans. Inf. Forensics Secur. 15, 2160–2170.

Zhang, Y., Wang, W., Zhang, P., 2021. The effect of silence and dual-band fusion in anti-spoofing system. In: Proc. Interspeech 2021. pp. 4279–4283.