2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) | 979-8-3503-2372-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/WASPAA58266.2023.10248132

ROBUST AUDIO ANTI-SPOOFING SYSTEM BASED ON LOW-FREQUENCY SUB-BAND INFORMATION

Menglu Li, Xiao-Ping Zhang

Toronto Metropolitan University
Department of Electrical, Computer and Biomedical Engineering
Toronto, ON, Canada
menglu.li@torontomu.ca, xzhang@ee.torontomu.ca

ABSTRACT

The current audio anti-spoofing systems usually have a computationally complex architecture without providing the fundamental discriminative factors for the detection judgments. The state-of-the-arts also highly depend on voice information to develop detector systems, which may become vulnerable when the spoofing algorithms have further improved the quality of fake speech. Therefore, we conduct a series of experiments on different frequency sub-bands to investigate the underlying discriminative features. We find the lowest frequency sub-band in the range from 0 to 1600Hz contains the most critical features that distinguish between Deepfake and real speech. We also focus on forensic evidence and identify that the basis of detectors' judgment exists in non-speech parts in audio samples. Based on the findings, our single detection system, with only 57K parameters and utilizing a one-tenth segment of the entire spectrogram as input, demonstrates its robustness by outperforming all official baselines of the ASVspoof2021 DF track. Our lightweight system can be easily applied in practical use cases, such as automated Deepfake screening or protecting voice-able devices.

Index Terms— Deepfake Detection, Subband Frequency, Speech Anti-Spoofing, ASVspoof2021, Robustness

1. INTRODUCTION

With the rapid development of audio spoofing technology, automatic speaker verification (ASV) systems [1] and voice-able devices [2] have faced severe secure challenges. Therefore, the need for Deepfake audio detection algorithms is emerging. To motivate the development of anti-spoofing audio technologies, a series of ASVspoof Challenges have been successfully held. Two major categories of spoofing audio generation techniques, which are Text-to-Speech (TTS) and Voice Conversion (VC), are both included in the Logical Access (LA) Track of ASVspoof Challenge of 2019 [3] and 2021 [4]. In ASVspoof2021 Challange, a new track, Speech Deepfake (DF), has been added. The DF Track focus on detecting the spoofing audio that aims to fool a human listener in more generalized scenarios. And its evaluation set contains speech audio generated by more than hundreds of spoofing algorithms with different data conditions and compression methods compared to the training dataset [4], which are created for the purpose of evaluating the robustness of the anti-spoofing algorithms.

We observe some limitations in the current stage of anti-spoofing detectors. Most works, even in the post-challenge time, only focus on improving the detection performance for the LA track dataset. The state-of-the-arts that produce a promising result on the LA dataset result in a high degree of overfitting for the DF track data [4]. Furthermore, the anti-spoofing systems with the top performance usually have a relatively complex architecture that requires a large number of parameters, the help of ensemble systems and data augmentation techniques [5], rather than stating the basis discriminative factors for the detection judgments. To address these issues, we conduct a series of systematic experiments to investigate the existence of underlying discriminative information in different frequency sub-bands. The power spectrogram of audio samples is divided into eight segments and passed into the identical back-end classifier which utilizes the Graph Attention Network (GAT) architecture. We have evaluated the detection performance of each frequency sub-band and found the lowest frequency sub-band in the range from 0 to 1600Hz contains the most critical characteristics that distinguish between Deepfake and real speech. With a one-tenth segment of the entire spectrogram, our proposed anti-spoofing detector is able to outperform all official baselines provided in the ASVspoof2021 Challenge DF Track. Furthermore, the reasons for the promising performance of low-frequency sub-band feature is analyzed and explained. The contribution of this work includes:

- We propose a robust low-frequency anti-spoofing system with only 57K parameters, which outperforms all the baselines for the ASVspoof2021 Challenge DF Track.
- We identify the discriminative information presented in the low-frequency sub-band of spectrograms, and demonstrate its power in detecting spoofing audio, especially for TTS-generated speech.
- We investigate the effectiveness of average-pooling operation on retaining the essential information of temporal features in audio samples, such as silence segments.

2. BACKGROUND

In the recent literature, the hand-crafted acoustic features, such as Mel-frequency Cepstral Coefficients (MFCC) [6, 7], Constant Q Cepstral Coefficients (CQCC) [8, 9], have been widely used in the Deepfake detectors to distinguish the spoofing audio based on the speech quality and similarity [10]. However, this type of detector becomes vulnerable when the spoofing algorithms have further improved the quality of fake speech. There is a need to find an explainable foundation for the countermeasure's judgment besides the audio quality. Some studies [11, 12] have shown the impacts of silence segments on fake audio detection to prevent overfitting, but there is a lack of detailed reasons to support the effect of non-speech segments. In this work, we further investigate to explore the under-

979-8-3503-2372-6/23/$31.00 ©2023 IEEE

lying difference between real and spoofing speech. We assume that the non-speech parts in the speech audio, including the breathing or silence segments, should contain more essential discriminative information than the voice parts. It is because the spoofing algorithms focus more on the actual speech content in order to fool the listeners rather than paying attention to fabricate other factors in real-world scenarios, such as electronic noise caused by the recording devices, which may not even be noticeable to the human. Therefore, it motivates us to investigate the existence of discriminative information in different frequency sub-bands, especially the low-frequency bands, and determine the effective identification features with an explanation for robust detectors. It also makes our work different from existing models that use sub-band analysis, which either incorporates more than half of the spectrogram range as a single sub-band or combines all sub-band features to encompass the full frequency region [13, 14].

3. EXPERIMENTAL SETUP

3.1. Dataset and Metrics

We chose to use the ASVspoof2021 DF Track dataset [15] to reflect the robustness of our detection algorithm. To follow the procedure of the ASVspoof2021 Challenge [4], the training and the development stages of the DF track use the same set of data as the ASVspoof2019 Logical Access (LA) Track, which both contain the spoofed audio generated by four TTS and two VC algorithms. In the training set, the ratio of real speech and spoofed speech is 1:9. The evaluation set of DF Track contains Deepfake audios generated from hundreds of unseen algorithms which are different from those in the training and development set. The DF Track evaluation data are collected from three data sources, ASVspoof2019 LA evaluation set [16], VCC2018 [17], and VCC2020 [18]. Therefore, it can better emphasize the ability of algorithms to fake speech detection across unknown and multiple conditions. Equal Error Rate (EER) is adopted as the indicator to measure the performance of detectors.

3.2. Front-End feature

We utilize the Fast Fourier Transform (FFT) power spectrogram as the main acoustic feature to describe the input audio signals, rather than using a Mel-scale spectrogram or the constant-Q transform coefficients. It is because the FFT power spectrogram keeps the relationship between each frequency bin as linear, which retains the audio information across the entire frequency spectrum instead of exaggerating the information at the lower frequencies or being lack of resolution at the high frequencies. Therefore, we can obtain reasonable evidence to determine the impact of different sub-bands on spoofing audio detection.

Before transforming to the spectrogram, we first fix all audio clips to the same length of 4 seconds by either truncating the longer audio clips or concatenating the shorter audio clips repeatedly. Then, the spectrograms are extracted using a Hann window function with a window size of 1000. Each input audio clip results in a spectrogram feature matrix with a shape of 501×259, of which 259 is the number of total time frames and 501 is the number of frequency bins. In order to investigate the importance of each sub-band, we divide the entire feature matrix into eight parts vertically with the same size, and each sub-feature matrix contains 50 frequency bins with 259 frames, which cover the acoustic information in the range of 1600Hz. The amplitude of each feature element is converted in a Decibel Scale to reflect the sound intensity.

The resulting 2-dimensional (2D) feature matrix can be considered as an image input with one channel, which is represented as $\mathbf{M} \in \mathbb{R}^{(C \times F \times T)}$. F and T stand for the number of frequency bins and time frames respectively, while C refers to the number of feature channels, which is 1 initially. Then, this feature matrix is passed into four ResNet blocks [19] to obtain a deep feature embedding. Our ResNet blocks are different from [20] in two ways: (1) The batch normalization layer with the ReLU activation function is applied before both convolutional layers, which can reduce the overfitting during training since the weights are normalized before the convolution operation. (2) We use a stride of 2 on convolutional layers instead of max-pooling to avoid information loss in the temporal dimension.

3.3. Model architecture

The GAT is applied as the classifier for the experiment. Since we focus on the effectiveness of the spectral domain, an average-pooling operation is applied across the temporal dimension of the feature embedding matrix, $\mathbf{M} \in \mathbb{R}^{(C \times F \times T)}$, which results in a spectral feature matrix $\mathbf{S} \in \mathbb{R}^{(C \times F)}$. We choose to apply average pooling rather than the max-pooling operation because max pooling may more focus on the extreme features, which will ignore the impacts brought by the silence segments inside the audio clips. Applying average pooling can consider all features present in the temporal domain, which will extract a more smooth representation.

In the spectral feature matrix $\mathbf{S} \in \mathbb{R}^{(C \times F)}$, each frequency bin of F is regarded as a node in the graph with a feature dimensionality of C. All nodes are fully connected inside the graph, and a GAT layer with an attention mechanism is applied to each node while assigning a learnable weight to each frequency bin. We then adopt the graph pooling technique inspired by [20] to select a subset of nodes with the largest values of weights to prevent overfitting. The final fully-connection layer is applied to the selected nodes to obtain the binary classification result.

Layer	Structure
Conv layer	conv2D((2,3), 16, stride = 1, padding = 1)
	BatchNorm, ReLU
	AveragePool((1,2), stride = 2)
[Res Block1] $\times 2$	BatchNorm, ReLU
	conv2D((2,3), 16, stride = 1, padding = 1)
	BatchNorm, ReLU
	conv2D((2,3), 16, stride = 1, padding = 1)
[Res Block2] $\times 4$	BatchNorm, ReLU
	conv2D((2,3), 32, stride = 2, padding = 1)
	BatchNorm, ReLU
	conv2D((2,3), 32, stride = 1, padding = 1)
GAT Block	AdaptiveAveragePool2D(26,1)
	GAT layer (#node = 26)
	Graph pooling (#node = 16)
Output	Fully-Connected layer

Table 1: The detailed architecture of the anti-spoofing detector used in the experiment

3.4. Training Configurations and Strategies

The detailed structure of the detection algorithm is shown in Table 1, which utilized a spectral sub-band of the spectrogram as input.

In the first two ResNet blocks, a convolutional filter of a size of 16 is applied. Then, the filter size increases to 32 for the other four ResNet blocks. The filter size is chosen as a fairly small value to prevent the overfit towards the training data. An asymmetric kernel size of the convolutional filter is chosen with a height of 2 and a width of 3 to aggregate the different lengths of features in the spectral and temporal dimensions. After the ResNet blocks and average-pooling operation, the number of frequency bins becomes 26 with a feature dimension of 32. The graph-pooling layer will then select 16 frequency bins with the largest value of weights to produce the final prediction.

The Weighted Cross Entropy is used as the loss function. The weight of the real speech and the spoofing speech is set to 0.9 and 0.1 because of the data imbalance in the training dataset. An Adam optimizer [21] with a weight decay of $1e$-4 is used. The learning rate increases linearly for the first 10 epochs as a warm-up to $1e$-4, and then decreases to zero by a cosine function. The model was trained with 300 epochs with a mini-batch size of 32. The model with the minimum validation loss for the development dataset was selected as the best model for evaluation. The detection model in our experiment only utilizes 57K parameters in total.

4. EXPERIMENTAL RESULTS AND ANALYSIS

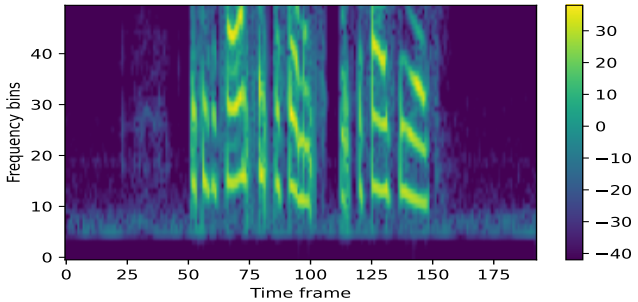
4.1. Frequency sub-band performance

The experiment results are shown in Table 2, which indicates the effectiveness of each frequency subband on spoofing audio detection. Table 2 demonstrates that the lowest sub-bands which only contain the first 50 frequency bins (0-1600Hz) produce the best EER score of 22.13% on the ASVspoof2021 DF evaluation set. Other higher frequency sub-bands do not contain such discriminate features to distinguish real and spoofing speech compared to the lowest band, while the full-band features lead to overfitting.

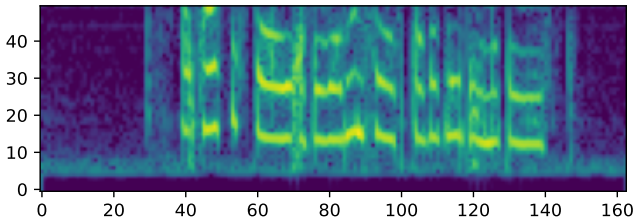
We also report the detection results of each sub-band on different data sources within the DF evaluation set. The spoofing data in VCC2018 and VCC2020 are all generated by various VC algorithms. The speech in ASV2019 contain both TTS and VC spoofing methods, while the majority of data are created via TTS systems. We can observe that the lowest sub-band has a significantly advanced ability to detect spoofing speech generated by TTS than VC as it gives the minimum EER of 13.34% for the ASV2019 subset. The full band of frequency features performs better in detecting VC-generated speech for both VCC2018 and VCC2020 sets, which indicates that the information in the actual voice segment is essential to detect the VC-based spoofing speech.

4.2. Sub-band feature analysis

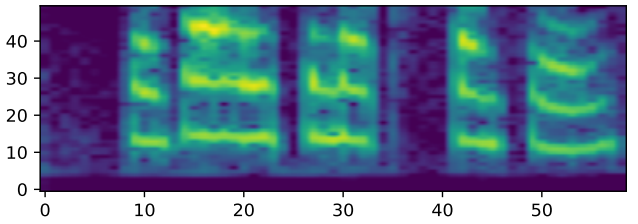
We further investigate the characteristics of the sub-band features. Figure 1 shows the power spectrogram of the lowest-frequency sub-band (0-50 frequency bins) of both real and spoofing audio in the training dataset. The x-axis stands for the time frames, and the y-axis represents the frequency bins, while the colour indicates the sound intensity in the Decibel scale. We can observe that there is a continuous noise band between 0 to 10 frequency bins in the real speech throughout the entire time frame. This noise band occurs in all real speech samples in the ASVspoof2019 training data, which



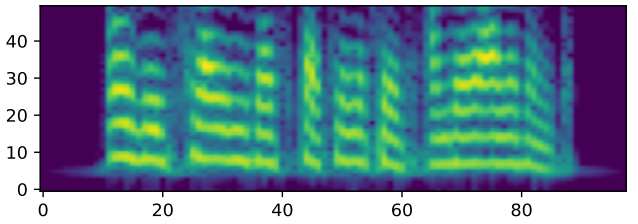
(a) Bonafide Speech (LA_T_1011221.flac)



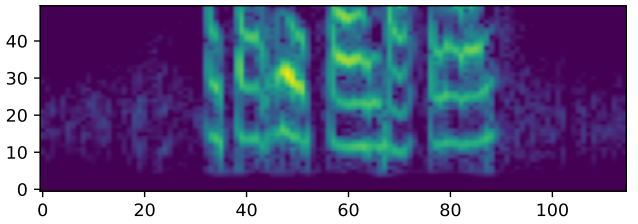
(b) Bonafide Speech (LA_T_1078988.flac)



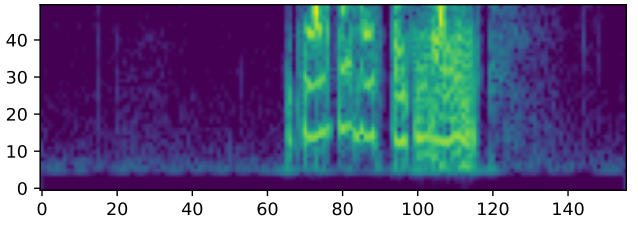
(c) A01 TTS-based Spoof (LA_T_1076943.flac)



(d) A03 TTS-based Spoof (LA_T_1014095.flac)



(e) A05 VC-based Spoof (LA_T_1035039.flac)



(f) A06 VC-based Spoof (LA_T_1064797.flac)

Figure 1: The spectrogram of low-frequency sub-band (first 50 bins) for both bonafide and spoofing speech

Freq-band	ASV2019	VCC2018	VCC2020	EER
0 - 50 bins	**13.34**	39.5	32.03	**22.13**
51-100 bins	17.43	41.9	44.09	25.94
101-150 bins	18.77	42.04	39.87	25.70
151-200 bins	19.12	43.81	33.08	24.62
201-250 bins	17.47	42.93	31.09	23.25
251-300 bins	19.48	42.84	31.61	24.44
301-350 bins	18.17	41.49	30.52	24.16
351-400 bins	19.59	41.22	30.95	24.01
401-450 bins	18.71	45.61	31.33	24.91
451-501 bins	35.63	43.12	40.41	36.41
0 - 501 bins	21.13	**32.52**	**30.13**	25.53

Table 2: The performance results of different frequency sub-band inputs in terms of EER (%) on each subset in the ASVspoof2021 DF track evaluation dataset based on the original source (ASV2019, VCC2018, VCC2020). The pooled EER scores for the entire evaluation set are shown in the last column.

should be caused by electronic recording devices. However, for the TTS-generated speech samples, there is no such continuous noise band at the low-frequency range, or the noises shown in the spectrogram are either unstable or abnormal. This situation occurs because TTS algorithms create audio directly from the text data, which more focus on improving the quality of voice rather than generating some noises in real life. Therefore, the low-band frequency feature contains the most discriminative information to detect TTS-based spoofing data, which is consistent with our experiment result in the previous section.

For the VC-based spoofing audio, since they are generated directly from the real recording speech, the electronic noise may be retained depending on the particular architecture of the VC algorithm. For example, the spoofing speech in A05 and A06 subsets, which are generated by different VC algorithms, result in different observations for the noise bands on the spectrogram. A06 still keeps the continuous noise band after altering the speakers' characteristics, which provides the explanation for the relatively poor performance of the low-frequency band on detecting VC-based spoofing audios.

4.3. Performance comparison

Table 3 illustrates the performance of our proposed system using the lowest frequency band compared to other single systems reported in the literature. The models, B01 to B04, are the official baselines of the ASVspoof2021 Challenge, and the corresponding evaluation results are reported in [4]. The other two chosen models, RawGAT [20] and Res2Net [22], are state-of-the-art for the ASVspoof LA track with open-source code available. We train the online source codes with the provided hyperparameters to obtain the detection results on the DF track evaluation set. As Table 3 indicates, our low-frequency system outperforms all baselines by utilizing a one-tenth segment of the entire spectrogram. This result emphasizes the discriminative information that exists in the lower frequency has the powerful and robust ability to distinguish spoofing speech. With only 57K parameters, our system with the least complex architecture has a higher possibility of being applied in real-world use cases.

Model	EER (%)	Parameters
Low-frequency system (Ours)	**22.13**	**57K**
B04: RawNet2 [4]	22.38	25000K
RawGAT [20]	22.47	440K
B03: LFCC-LCNN [4]	23.48	-
Res2Net [22]	24.47	923K
B02: LFCC-GMM [4]	25.25	-
B01: CQCC-GMM [4]	25.56	-

Table 3: Performance on the ASVspoof2021 DF evaluation set for our proposed model and different state-of-the-art single systems and baseline systems

4.4. Ablation study

We measure the effectiveness of our design choices in ablation experiments as Table 4 shows. The first experiment is to demonstrate the impact of the average-pooling operation in generating the spectral feature matrix. After replacing the average-pooling layer with max pooling, the detection performance degrades by 6.6% (23.70% cf. 22.13%), which indicates the benefit of the average-pooling operation in retaining informative temporal features. The second experiment investigates the effectiveness of the cosine-based scheduling technique in updating the learning rate. An unchanged learning rate of $1e$-4 throughout the training process leads to a performance degradation of 3.2% (22.88% cf. 22.13%) caused by overfitting.

Model	EER (%)
Low-frequency system	**22.13**
w/o averaging-pooling	23.70
w/o learning rate scheduling	22.88

"w/o averaging-pooling" refers to using max-pooling layers instead of average-pooling; "w/o learning rate scheduling" refers to keeping the learning rate as $1e$-4 without decreasing by a cosine function

Table 4: Ablation experiments of our model design

5. CONCLUSION

In this work, the effect of different frequency sub-bands on detecting spoofing audio is explored. We propose a robust GAT-based Deepfake audio detector, utilizing a low-frequency band as input, to detect fake speech across various conditions. With only 57K parameters, our single system has obtained a compelling performance that outperforms all official baselines of the ASVspoof2021 DF track. Because of its good robustness and usage efficiency, our model, as the most lightweight system reported in the literature, has gained a higher possibility of being applied in real-world use cases, such as automated Deepfake screening. We also identify the discriminative information presented in the low-frequency sub-band of spectrograms, which provides an explanation of the effectiveness of our low-frequency detector. We propose that the non-speech parts contain more identification features to make the detection judgment, especially for detecting TTS-generated speech. This finding provides a direction for the potential design of future anti-spoofing systems.

6. REFERENCES

[1] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[2] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, X.-Y. Li, Y. Wang, and Y. Deng, "Voicemask: Anonymize and sanitize voice input on mobile devices," *arXiv preprint arXiv:1711.11460*, 2017.

[3] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *Interspeech 2019*, 2019.

[4] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, *et al.*, "Asvspoof 2021: Accelerating progress in spoofed and deepfake speech detection," *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021.

[5] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, *et al.*, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *arXiv preprint arXiv:2210.02437*, 2022.

[6] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," *Interspeech 2019*, 2019.

[7] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "A comparative study on physical and perceptual features for deepfake audio detection," *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 2022.

[8] R. K. Das, J. Yang, and H. Li, "Long range acoustic and deep features perspective on asvspoof 2019," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 1018–1025.

[9] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," *Interspeech 2015*, 2015.

[10] J. Xue, C. Fan, Z. Lv, J. Tao, J. Yi, C. Zheng, Z. Wen, M. Yuan, and S. Shao, "Audio deepfake detection based on a combination of f0 information and real plus imaginary spectrogram features," in *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 2022, pp. 19–26.

[11] Y. Zhang, W. Wang, and P. Zhang, "The effect of silence and dual-band fusion in anti-spoofing system," *Interspeech 2021*, 2021.

[12] N. M. Müller, F. Dieckmann, P. Czempin, R. Canals, K. Böttinger, and J. Williams, "Speech is silver, silence is golden: What do asvspoof-trained models really learn?" *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021.

[13] J. Yang, R. K. Das, and H. Li, "Significance of subband features for synthetic speech detection," *IEEE Transactions on Information Forensics and Security*, vol. 15, p. 2160–2170, 2020.

[14] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, "Spoofing attack detection using the non-linear fusion of subband classifiers," *Interspeech 2020*, 2020.

[15] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang, and J. Yamagishi, "Asvspoof 2021 challenge - speech deepfake database," May 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4835108

[16] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[17] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," *The Speaker and Language Recognition Workshop (Odyssey 2018)*, 2018.

[18] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020:intra-lingual semi-parallel and cross-lingual voice conversion," *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 2016, pp. 630–645.

[20] H. Tak, J.-w. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] X. Li, N. Li, C. Weng, X. Liu, D. Su, D. Yu, and H. Meng, "Replay and synthetic speech detection with res2net architecture," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6354–6358.