Multi-grained Backend Fusion for Manipulation Region Location of Partially Fake Audio

Jun Li[1], Lin Li[2,3], Mengjie Luo[2,3], Xiaoqin Wang[2,3,*], Shushan Qiao[2,3] and Yumei Zhou[2,3]

[1]Nanjing Institute of Intelligent Technology, Nanjing, China
[2]Institute of Microelectronics of the Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China

Abstract

Fake audio detection is an important research area to prevent the misuse of speech synthesis and voice conversion technologies. While progress has been made in detecting partially fake audio at the utterance level, accurately locating the manipulation region at the segment level remains challenging. Aiming to promote the development of manipulation region location of partially fake audio, ADD 2023 is organized and Track 2 seeks to locate the fake clips. This paper introduces our system submitted to ADD 2023 Track 2, combining AASIST-based and Wav2Vec2-based subsystems through multi-grained backend fusion. With the proposed method, the bias of AASIST towards fake class, and Wav2Vec2 towards genuine class are mitigated. Our system achieves a $Score$ of 59.12%, a 40.7% increase compared to the best baseline system in this paper.

Keywords

Fake Audio Detection, Audio Deepfake Detection, Partially Fake, Wav2Vec2, AASIST, Backend Fusion

1. Introduction

The advancement of speech synthesis and voice conversion(VC) technologies has significantly enhanced the quality and naturalness of synthesized speech [1, 2, 3, 4]. However, an issue of potential technology abuse such as telecom fraud may be brought up. Consequently, there is a growing concern about fake audio detection(FAD), where the synthesized audio for inappropriate uses is defined as fake audio or spoofing attacks.

The Asvspoof challenges have gathered attention from researchers who aim to protect automatic speaker verification(ASV) systems from spoofing attacks, [5, 6, 7, 8]. The Asvspoof 2015, 2017 and 2019 focused on the logical access(LA) task, physical access(PA) task or both. The LA task involved detecting spoofing audio generated by statistical or neural text-to-speech(TTS) and VC methods, While the PA task aimed to distinguish replay audio implemented in various simulated acoustic environments. In Asvspoof 2021, a new deepfake track was introduced to detect compressed manipulated audio, aiming to enhance system robustness. Furthermore, the spoofing-aware speaker verification(SASV) challenge in 2022 attempted to jointly optimize FAD and ASV systems instead of utilizing a FAD system as a gate to start the ASV system[9]. Among these challenges, ResNet-based frameworks[10] have been widely adopted in ASVspoof challenges [11, 12], and the AASIST[13] was served as a baseline model and employed by several top-ranked participants in SASV 2022 who would like to achieve a low equal error rate of FAD[14, 15, 16, 17].

Previous challenges have primarily focused on detecting fully fake audio at the utterance level, without addressing realistic scenarios involving partially fake audio. Partially fake audio refers to fake audio with small fake clips hidden in genuine speech audio[18, 19]. To address this gap, ADD challenges are launched to encourage researchers to explore new frameworks for detecting partially fake audio[20, 21]. In ADD 2022(Audio Deep Synthesis Detection Challenge), Track 2 targeted at detecting partially fake audio at the utterance level. In ADD 2023(Audio Deepfake Detection Challenge), the goal of Track 2 is localizing manipulated clips within a speech sentence. In ADD 2022 Track 2, the best partially FAD system at the utterance level is based on pretrained self-supervised Wav2Vec2[22, 23], but it fails to spot fake clips[24]. On the other hand, methods focusing on the frame-wisely boundary detection of manipulated clips have shown capability in locating fake clips [25, 26].

This paper presents our system for the manipulation region location of partially fake audio in ADD 2023 Track 2. The backend fused system combines AASIST for detecting fake audio at the utterance level, and Wav2Vec2 at the frame level. The main contribution of this paper is the proposal of multi-grained backend fusion, which aims to mitigate the biases of AASIST towards fake audio and Wav2Vec2 towards genuine audio. Our submitted system achieves a $Score$ of 59.12%, a relative increase of 40.7% compared to the best baseline system.

IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023), August 19, 2023, Macao, S.A.R
*Corresponding author: Xiaoqin Wang.
lj@niit.ac.cn (J. Li); lilin2020@ime.ac.cn (L. Li); luomengjie@ime.ac.cn (M. Luo); wangxiaoqin@ime.ac.cn (X. Wang); qiaoshushan@ime.ac.cn (S. Qiao); ymzhou@ime.ac.cn (Y. Zhou)

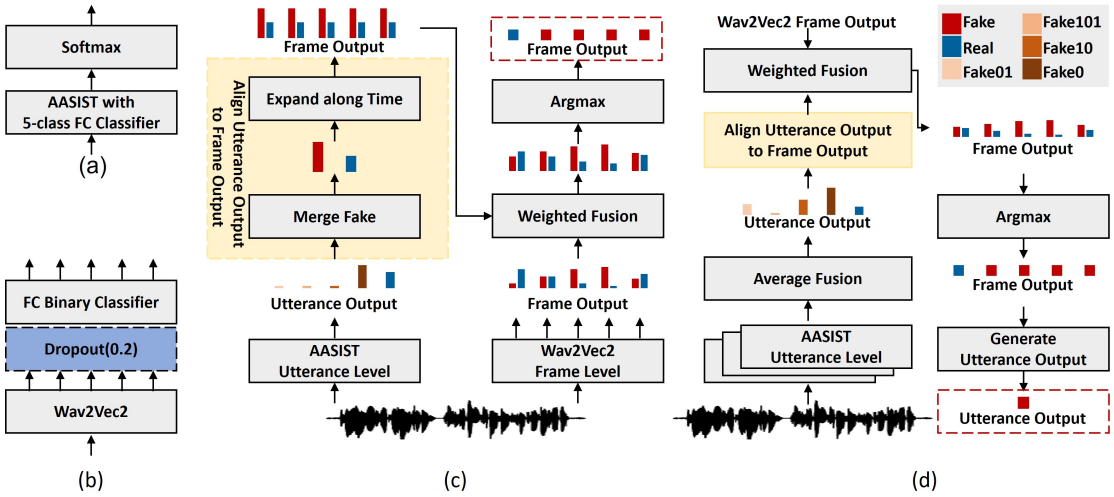

Figure 1: The overview of our proposed system. (a) AASIST-based subsystem at the utterance level. (b) Wav2Vec2-based subsystem at the frame level. (c) Manipulation region location system. (d) FAD system at the utterance level.

The rest of this paper is as follows. The proposed method is described in Section 2. Section 3 details the experiment settings. Experimental results and analysis are discussed in Section 4. Finally, Section 5 concludes the paper.

2. Method

2.1. Task Definition

FAD at the utterance level is a binary classification task to detect if a sentence is genuine or fake. In contrast, manipulation region location identifies fake segments at a finer granularity. In Track 2 of ADD 2023, the duration of each segment is 10ms. Therefore, given an utterance X with N segments, represented as $X = [x_1, x_2, ..., x_N]$, the output at the utterance level should be $y_{sen} \in \{0, 1\}$, while the output at the segment level is a vector $\mathbf{y}_{seg} = [y_1, y_2, ..., y_N] \in \{0, 1\}^N$, where *0* denotes *fake* and *1* denotes *genuine*. Besides, since the duration of segments is similar to that of a frame commonly used in speech processing, models generating frame outputs can be used to detect segments.

2.2. Proposed System

Figure 1 shows an overview of our proposed system. It comprises subsystems, namely AASIST for utterance-level FAD and Wav2Vec2 for frame-level analysis. The final results at different levels are obtained by the fusion of multi-grained outputs from subsystems.

2.2.1. Subsystems

AASIST-based subsystem at the utterance level AASIST is an end-to-end architecture based on graph attention network, proposed to detect different spoofing attacks[13]. The raw waveform is adopted as input, with a minimum of 64,600 samples, about 4s at a sampling rate of 16kHz. While the original AASIST aims to classify genuine and spoofed utterances with a binary classifier[1], the classifier is replaced by a 5-class FC(fully connected) layer to detect 4 types of fake audio along with a genuine class. In the training and development set of ADD 2023 Track 2, we refer to the 4 fake forms as Fake01, Fake101, Fake10 and Fake0, where *0* denotes the presence of manipulated fake clips. Finally, the logits of the last FC layer are fed into a softmax function.

Wav2Vec2-based subsystem at the frame level To address the limitation of AASIST in fake clips location, Wav2Vec2-based subsystem is employed to determine the authenticity of each frame. A self-supervised pretrained model called XLS-R-300M with 300M parameters is utilized to capture contextualized acoustic representations [2][23]. Similar to AASIST, Wav2Vec2 also takes the raw waveform as input. It generates frame representations at a hop length of 20ms, with each frame length 25ms, given an input sampling rate of 16kHz. The last hidden output of Wav2Vec2 is passed through a dropout layer, followed by a binary linear layer for frame classification.

[1]https://github.com/clovaai/aasist
[2]https://huggingface.co/facebook/Wav2Vec2-xls-r-300m

Table 1
The statistics of datasets in ADD 2023 Track 2.

Name	Genuine	Fake					Total
		Fake01	Fake101	Fake10	Fake0	Sum	
Train	26554	8487	14319	2547	1185	26538	53092
Develop	8913	2890	4751	839	430	8910	17823
Test	-	-	-	-	-	-	50000

2.2.2. Multi-grained Backend Fusion

Manipulation region location system As depicted in Figure 1(c), the manipulation region location system consists of an AASIST-based subsystem and a Wav2Vec2-based subsystem. By fusing multi-grained results from these subsystems, the system aims to mitigate biases observed in experiments detailed in Section 4. The alignment of utterance results to frame level involves two main steps. Firstly, probabilities of all types of fake audio are summed, converting the 5-class to binary classification. Then the binary classification outputs at the utterance level are expanded along the time domain to match the number of frames in Wav2Vec2 outputs. The expanded utterance outputs with frame outputs are combined by weighted fusion, and the argmax function is applied to determine the authenticity of each frame.

FAD system at the utterance level To enhance sentence accuracy, average fusion of AASIST models trained on different datasets is utilized. The averaged utterance result is then fused with the frame output of a Wav2Vec2-based subsystem. Following the definition in ADD 2023, if any frame is identified as fake, the label of fake is assigned to the entire utterance. Only when all frames are classified as genuine, the utterance is labeled as genuine.

3. Experiment Settings

3.1. Data Preparation

Various datasets are used for training. The sampling rate of all data is 16kHz. The details of the datasets provided by ADD 2023 Track 2 are presented in Table 1. This includes a train set used for model fitting, a development set used for an early stop during training, and a test set whose labels are unknown, and used to evaluate the FAD system. The distributions of genuine and fake utterances in both train set and development set are balanced, However, the percentages of each fake type vary, with *Fake101* being the majority, and *Fake0* being the minority. To enhance the generalization capability of our system, new training data is constructed as outlined in Table 2. The *RS* represents the individual genuine sentences obtained by splitting continuous real segments from each genuine or partially fake sentence in the train set. The *FS* denotes fake sentences generated by splitting continuous fake clips from each fake sentence in the train set. Three traditional vocoders, namely GL(griffin-Lim) [3] [27], Straight [4] [28] and World [5] [29] are employed to synthesize fake audio from the real segments of *RS*. Additionally, utterances in *MidAug* are created by randomly inserting newly constructed fake clips into the audio of *RS*. In *MidAug*, the duration range of fake clips is [0.2s, 3s], and any utterances shorter than 0.2s are discarded.

Table 2
Constructed Train Data.

Name	Label	Total
RS	Genuine	66226
FS	Fake0	26538
GL	Fake0	66226
Straight	Fake0	6352
World	Fake0	2390
MidAug	Fake101	51907

During training, Online data augmentation is employed. The MUSAN dataset[30] is utilized to add background noise with noise and music, while the RIR database[31] is used to simulate reverberation. Dynamic padding is applied. Additionally, the duration of audio is fixed to 5s during the training of AASIST, whereas the full length when testing. For Wav2Vec2, 4s is mainly employed as the maximum duration both for training and testing.

3.2. Training

The system is mainly built on top of [32] and each model is trained on an Nvidia 3090 GPU card. Cross entropy is adopted as loss function and Adam[33] as optimizer. The train batch size is 16. Baseline subsystems are trained with the train set from ADD 2023 Track 2 with max epoch 50. The initial learning rate is 1e-3, and it decreases by 5% after every epoch. To quickly converge to new data, we finetune the baseline models with lowest loss on development set for another 20 epochs. The finetune learning rate starts from 1e-4.

[3] https://librosa.org/doc/main/generated/librosa.griffinlim.html
[4] https://github.com/HidekiKawahara/legacy_STRAIGHT
[5] http://www.isc.meiji.ac.jp/ mmorise/world/english/download.html

Table 3
Experimental Results. *Str.* is the abbreviation of *Straight*, and *Wor.* is *World*.

Model	Name	Train set/Fused System	A_{sen}(%)	P_{seg}(%)	R_{seg}(%)	$F1_{seg}$(%)	Score(%)
AASIST	B1	train	67.54	20.87	60.90	31.09	42.02
	A1	+FS	58.87	21.71	48.07	29.91	38.60
	A2	+GL	55.23	20.64	40.03	27.23	35.63
	A3	+Straight	67.64	21.99	85.44	34.98	44.78
	A4	+World	51.58	**22.02**	32.02	26.10	33.74
	A5	+FS+GL+Str.	67.92	21.53	89.96	34.74	44.69
	A6	+FS+GL+Str.+Wor.	69.73	21.93	84.30	34.80	45.28
	A7	+RS	59.15	21.91	33.33	26.44	36.25
	A8	+MidAug	68.80	21.32	69.28	32.61	43.47
	A9	+FS+GL+Str.+RS+MidAug	72.48	21.22	92.10	34.49	45.88
	A10	+FS+GL+Str.+Wor.+RS+MidAug	**72.61**	21.54	**94.13**	**35.05**	**46.32**
Fusion	FA1	(A10, A9)	72.68	21.30	**93.75**	34.71	46.10
	FA2	(A10, A9, A6)	72.86	21.42	92.58	34.79	46.21
	FA3	(A10, A9, A6, A8)	**73.36**	21.51	91.85	34.85	**46.40**
	FA4	(A10, A9, A6, A8, A5)	72.68	21.52	91.77	34.87	46.21
	FA5	(A10, A9, A6, A8, A5, A3)	70.42	21.85	89.09	**35.09**	45.69
	FA6	(A10, A9, A6, A8, A5, A3, B1)	70.48	**21.87**	88.27	35.06	45.69
Wav2Vec2	B2	train	45.33	62.02	5.83	10.66	21.06
	W1	+MidAug	54.03	**75.95**	16.83	27.56	35.50
	W2	+FS+GL+Str.+Wor.+RS+MidAug	**58.10**	67.99	**22.33**	**33.62**	**40.96**
Fusion	FW1	(W1, W2)	54.30	**81.43**	16.76	27.79	35.75
Multi-grained	B3	0.91*B1+0.09*B2	49.48	28.71	29.28	28.99	35.14
Backend	FS1	0.90*A10+0.10*FW1	58.51	40.82	54.37	46.63	50.19
Fusion	FS2	0.90*A10+0.10*W1	74.50	45.23	60.77	51.86	**58.65**
	FS3	0.85*A10+0.15*W2	71.37	47.97	58.04	**52.53**	58.18
	FS4	0.90*FA1+0.10*FW1	58.89	42.52	51.18	46.45	50.18
	FS5	0.91*FA1+0.09*W1	**74.52**	42.19	**63.70**	50.76	57.89
	FS6	0.85*FA1+0.15*W2	71.36	48.51	55.70	51.86	57.71
	FS7	0.90*FA3+0.10*FW1	59.47	42.81	52.39	47.12	50.82
	FS8	0.90*FA3+0.10*W1	73.50	47.87	55.80	51.54	58.13
	FS9	0.85*FA3+0.15*W2	70.78	**49.82**	53.98	51.81	57.50

3.3. Evaluation Metrics

The sentence accuracy(A_{sen}) and segment F1 score($F1_{seg}$) are simultaneously adopted as evaluation metrics for ADD 2023 Track 2. Taking *fake* as *positive* and *genuine* as *negative*, TP, TN, FP, FN are the numbers of true positive, true negative, false positive, false negative samples.

At the utterance level, TP, TN, FP, FN samples denote utterances, A_{sen} is defined as

$$A_{sen} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{1}$$

The metrics at the segment level aim to measure the ability of models to correctly identify fake clips from fake audio[21], including P_{seg} for segment precision, R_{seg} for segment recall, and $F1_{seg}$ for F1 score, they are defined as follows:

$$P_{seg} = \frac{TP}{TP + FP} \tag{2}$$

$$R_{seg} = \frac{TP}{TP + FN} \tag{3}$$

$$F1_{seg} = \frac{2PR}{P + R} \tag{4}$$

where TP, FP and FN samples denote segments.

The final $Score$ of ADD 2023 Track 2 is defined as

$$Score = 0.3 \times A_{sen} + 0.7 \times F1_{seg}. \tag{5}$$

4. Results and Analysis

Table 3 presents the experimental results of the proposed systems. Baselines B1, B2 are AASIST and Wav2Vec2 trained with the train set of ADD 2023 Track 2 respectively. A1-A10 are AASIST trained with the constructed data in Table 2 in addition to the train set. FA1-FA6 represent the results of fusing probabilities of AASIST model at the utterance level, where the subsystems are chosen based on the sorted A_{sen}. W1 and W2 are Wav2Vec2

trained with additional constructed data. The number of Wav2Vec2 experiments is limited due to the unacceptable training and evaluation time. FW1 is the result of average fusion of Wav2Vec2 at the frame level. B3 and FS1-FS9 are the results of fused systems shown in Figure 1 (b) and (c), where B3 is a baseline, and AASISTs are chosen based on A_{sen} and R_{seg}, Wav2Vec2 based on P_{seg}. The weighted fusion factors of each subsystem are provided.

Baselines. Comparing the results obtained from B1 and B2, it can be observed that AASIST performs better in terms of A_{sen} and R_{seg}, while Wav2Vec2 achieves a higher score in P_{seg}. The reason may be AASIST at the utterance level tends to use global information and the process of transforming utterance to frame outputs of AASIST makes fake segments majority, leading to misidentification of genuine segments. Conversely, as the genuine segments are the majority in the train set, Wav2Vec2 at the frame level has a bias to the genuine class. To address the biases in AASIST and Wav2Vec2, B3 utilizes multi-grained fusion. Although most metrics in B1, and P_{seg} in B2 decrease, the R_{seg} improves relatively 402.2% compared to B2, and P_{seg} by 37.6% to B1, revealing the deviations of AASIST towards fake, and Wav2Vec2 towards genuine are lessened to some extent.

AASIST. When only one kind of constructed data is added to the train set, A3 with Straight exhibits a notable improvement in R_{seg}, a relative 47.7% increase compared to B1. The highest R_{seg} 94.13% is achieved by A10, indicating that the generalization can be improved by using all train data. Finally, through the fusion of top-ranked A_{sen} AASIST subsystems, the A_{sen} rises to 73.36% in FA3, P_{seg} to 21.87% in FA6, $F1_{seg}$ to 35.09% in FA5, and 46.40% in FA3.

Wav2Vec2. When all available data is utilized in W2, there is an improvement in all metrics compared to B2, with a growth of 28.2% in A_{sen}, 9.6% in P_{seg}, 283.0% in R_{seg}, 215.4% in $F1_{seg}$, 94.5% in $Score$. The highest P_{seg} 81.43% is achieved by combining W1 and W2 in FW1.

Multi-grained Backend Fusion. Having discussed in *Baselines*, though the performance of AASIST and Wav2Vec2 is improved by adding more constructed data or fusing subsystems separately, there remain biases for AASIST towards fake and Wav2Vec2 towards genuine. The selection of top-performing subsystems aims to mitigate the biases by multi-grained backend fusion. However, it can be observed that the adoption of FW1 such as FS7 with a $Score$ of 50.18% performs inferior to the fused systems with a single Wav2Vec2. This could be attributed to the decrease in R_{seg}, as the confidence of real segments generated by fused Wav2Vec2 increases. Additionally, as shown in Table 3, the best $F1_{seg}$ of 52.35% is achieved by combining A10 and W2, both are single subsystems trained with all data, indicating the importance of model generalization. Conversely, the best A_{sen} is acquired in FS5, with a relatively high R_{seg} of 63.70% in FS1-FS9, suggesting the significance to recognize fake audio when evaluating. The best $Score$ of 58.65% of a system is obtained in FS2, with a balanced A_{sen} and $F1_{seg}$, Finally, the submitted results for ADD 2023 Track2 utilize the results of FS5 at the utterance level, and FS3 at the segment level, achieving a $Score$ of 59.12%.

5. Conclusions

In this paper, a system based on multi-grained backend fusion is proposed to locate the manipulation region. The performance is improved with the proposed system by mitigating the biases brought by AASIST at the utterance level and Wav2Vec2 at the frame level. Our method achieves an A_{sen} of 74.52%, a $F1_{seg}$ of 52.53%, and the final $Score$ is 59.12%. Compared to the best baseline system B1 with a $Score$ of 42.02%, the proposed system achieves a relative improvement of 40.7%.

References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al., Tacotron: Towards end-to-end speech synthesis, arXiv preprint arXiv:1703.10135 (2017).

[2] J.-M. Valin, J. Skoglund, Lpcnet: Improving neural speech synthesis through linear prediction, in: ICASSP, IEEE, 2019, pp. 5891–5895.

[3] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, H.-M. Wang, Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks, arXiv preprint arXiv:1704.00849 (2017).

[4] T. Kaneko, H. Kameoka, K. Tanaka, N. Hojo, Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion, in: ICASSP, IEEE, 2019, pp. 6820–6824.

[5] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, A. Sizov, Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge, in: Sixteenth annual conference of the international speech communication association, 2015.

[6] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, K. A. Lee, The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection (2017).

[7] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, K. A. Lee, Asvspoof 2019: Future horizons in spoofed and fake audio detection, arXiv preprint arXiv:1904.05441 (2019).

[8] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen,

N. Evans, H. Delgado, ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection, in: Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 47–54.

[9] J. weon Jung, H. Tak, H. jin Shim, H.-S. Heo, B.-J. Lee, S.-W. Chung, H.-J. Yu, N. Evans, T. Kinnunen, SASV 2022: The First Spoofing-Aware Speaker Verification Challenge, in: Proc. Interspeech 2022, 2022, pp. 2893–2897.

[10] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition (2015). arXiv:1512.03385.

[11] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, K. A. Lee, Asvspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech, IEEE Transactions on Biometrics, Behavior, and Identity Science 3 (2021) 252–265.

[12] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, K. A. Lee, Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild (2022). arXiv:2210.02437.

[13] J. weon Jung, H.-S. Heo, H. Tak, H. jin Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, N. Evans, Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks (2021). arXiv:2110.01200.

[14] X. Wang, X. Qin, Y. Wang, Y. Xu, M. Li, The DKU-OPPO System for the 2022 Spoofing-Aware Speaker Verification Challenge, in: Proc. Interspeech 2022, 2022, pp. 4396–4400.

[15] J.-H. Choi, J.-Y. Yang, Y.-R. Jeoung, J.-H. Chang, HYU Submission for the SASV Challenge 2022: Reforming Speaker Embeddings with Spoofing-Aware Conditioning, in: Proc. Interspeech 2022, 2022, pp. 2873–2877.

[16] P. Zhang, P. Hu, X. Zhang, Norm-constrained Score-level Ensemble for Spoofing Aware Speaker Verification, in: Proc. Interspeech 2022, 2022, pp. 4371–4375.

[17] L. Zhang, Y. Li, H. Zhao, Q. Wang, L. Xie, Backend Ensemble for Speaker Verification and Spoofing Countermeasure, in: Proc. Interspeech 2022, 2022, pp. 4381–4385.

[18] J. Yi, Y. Bai, J. Tao, Z. Tian, C. Wang, T. Wang, R. Fu, Half-truth: A partially fake audio detection dataset, arXiv preprint arXiv:2104.03617 (2021).

[19] H. Ma, J. Yi, C. Wang, X. Yan, J. Tao, T. Wang, S. Wang, L. Xu, R. Fu, Fad: A chinese dataset for fake audio detection (2022). arXiv:2207.12308.

[20] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, S. Liang, S. Wang, S. Zhang, X. Yan, L. Xu, Z. Wen, H. Li, Add 2022: the first audio deep synthesis detection challenge, in: ICASSP, 2022, pp. 9216–9220.

[21] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Zhang, X. Zhang, Y. Zhao, Y. Ren, L. Xu, J. Zhou, H. Gu, Z. Wen, S. Liang, Z. Lian, H. Li, Add 2023: the second audio deepfake detection challenge, accepted by IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023) (2023).

[22] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations (2020). arXiv:2006.11477.

[23] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, M. Auli, Xls-r: Self-supervised cross-lingual speech representation learning at scale (2021). arXiv:2111.09296.

[24] Z. Lv, S. Zhang, K. Tang, P. Hu, Fake audio detection based on unsupervised pretraining models, in: ICASSP, IEEE, 2022, pp. 9231–9235.

[25] H. Wu, H.-C. Kuo, N. Zheng, K.-H. Hung, H.-Y. Lee, Y. Tsao, H.-M. Wang, H. Meng, Partially fake audio detection by self-attention-based fake span discovery, in: ICASSP, IEEE, 2022, pp. 9236–9240.

[26] Z. Cai, W. Wang, M. Li, Waveform boundary detection for partially spoofed audio, in: ICASSP, IEEE, 2023, pp. 1–5.

[27] N. Perraudin, P. Balazs, P. L. Søndergaard, A fast griffin-lim algorithm, in: 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2013, pp. 1–4.

[28] H. Kawahara, Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds, Acoustical science and technology 27 (2006) 349–353.

[29] M. Morise, F. Yokomori, K. Ozawa, World: a vocoder-based high-quality speech synthesis system for real-time applications, IEICE TRANSACTIONS on Information and Systems 99 (2016) 1877–1884.

[30] D. Snyder, G. Chen, D. Povey, Musan: A music, speech, and noise corpus (2015). arXiv:1510.08484.

[31] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, S. Khudanpur, A study on data augmentation of reverberant speech for robust speech recognition, in: ICASSP, IEEE, 2017, pp. 5220–5224.

[32] J. S. Chung, J. Huh, S. Mun, M. Lee, H.-S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, I. Han, In Defence of Metric Learning for Speaker Recognition, in: Proc. Interspeech 2020, 2020, pp. 2977–2981.

[33] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization (2014). arXiv:1412.6980.