ADD 2023: the Second Audio Deepfake Detection Challenge

Jiangyan Yi[1,*], Jianhua Tao[2,*], Ruibo Fu[1], Xinrui Yan[1,†], Chenglong Wang[1,†], Tao Wang[1,†], Chu Yuan Zhang[1,†], Xiaohui Zhang[1], Yan Zhao[1], Yong Ren[1], Le Xu[1], Junzuo Zhou[1], Hao Gu[1], Zhengqi Wen[1], Shan Liang[1], Zheng Lian[1], Shuai Nie[1] and Haizhou Li[3,4]

[1] *State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China*
[2] *Department of Automation, Tsinghua University, Beijing, China*
[3] *Department of Electrical and Computer Engineering, National University of Singapore, Singapore*
[4] *The Chinese University of Hong Kong, Hong Kong*

Abstract
Audio deepfake detection is an emerging topic in the artificial intelligence community. The second Audio Deepfake Detection Challenge (ADD 2023) aims to spur researchers around the world to build new innovative technologies that can further accelerate and foster research on detecting and analyzing deepfake speech utterances. Different from previous challenges (e.g. ADD 2022), ADD 2023 focuses on surpassing the constraints of binary real/fake classification, and actually localizing the manipulated intervals in a partially fake speech as well as pinpointing the source responsible for generating any fake audio. Furthermore, ADD 2023 includes more rounds of evaluation for the fake audio game sub-challenge. The ADD 2023 challenge includes three subchallenges: audio fake game (FG), manipulation region location (RL) and deepfake algorithm recognition (AR). This paper describes the datasets, evaluation metrics, and protocols. Some findings are also reported in audio deepfake detection tasks.

Keywords
Audio deepfake, fake detection, audio fake game, manipulation region location, deepfake algorithm recognition

1. Introduction

Over the last decades, the development of artificial intelligence has brought forth great improvements in speech synthesis [1, 2, 3] and voice conversion [4, 5, 6] technologies. The models are able to generate realistic and human-like speech. The technology nevertheless poses a serious threat to the society if someone misuses it [7]. Therefore, audio deepfake detection is an emerging topic of interest. An increasing number of efforts have been made to detect the deepfake audio recently [8, 9, 10, 11, 12, 13, 14].

A series of challenges, including Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2021) [11], the First Audio Deepfake Detection Challenge (ADD 2022) [12] have played a critical role in fostering research on this area. The ASVspoof 2021 introduced a new task involving audio deepfake (DF) detection, accelerating progress in deepfake audio detection. To address more challenges in the real world, the ADD 2022 [1] included three Tracks: low-quality fake audio detection (LF), partially fake audio detection (PF) and audio fake game (FG). However, some limitations still existed in ADD 2022. The techniques used in the challenge focused more on performing binary classification between real and fake audio. In addition, there were limited rounds of evaluation for the FG Track.

Moreover, there is also an interest in surpassing the constraints of binary real/fake classification, and actually localizing the manipulated intervals in a partially fake speech as well as pinpointing the source responsible for generating any fake audio. Therefore, we launched a second Audio Deepfake Detection Challenge (ADD 2023 [2]) to spur researchers around the world to build new innovative technologies that can further accelerate and foster research on detecting and analysing deepfake utterances.

In the following sections, we describe the datasets and evaluation metrics designed for different subchallenges. Finally, we briefly report on the performance of the results submitted by the ADD 2023 participants to further explore the current state and future direction of real-world audio deepfake detection.

2. Subchallenges

The ADD 2023 challenge includes three subchallenges: audio fake game (FG) [15, 16], manipulation region loca-

IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023), August 19, 2023, Macao, S.A.R
[*] Corresponding author.
[†] These authors contributed equally.

jiangyan.yi@nlpr.ia.ac.cn (J. Yi); jhtao@tsinghua.edu.cn (J. Tao); ruibo.fu@nlpr.ia.ac.cn (R. Fu); yanxinrui2021@ia.ac.cn (X. Yan); chenglong.wang@nlpr.ia.ac.cn (C. Wang); tao.wang@nlpr.ia.ac.cn (T. Wang); zhangchuyuan2021@ia.ac.cn (C. Y. Zhang); haizhou.li@u.nus.edu (H. Li)

[1] http://addchallenge.cn/add2022
[2] http://addchallenge.cn/add2023

tion (RL) and deepfake algorithm recognition (AR). The RL and AR subchallenges are new to ADD.

Track 1. audio fake game (FG): Different from ADD 2022, ADD 2023 has two rounds of evaluations for the generation task and two rounds of evaluations for the detection task.

Track 1.1 generation task (FG-G): aiming to generate fake audio that can fool the fake detection model of Track 1.2.

Track 1.2 detection task (FG-D): attempting to detect fake utterances, especially the fake samples generated from Track 1.1.

Track 2. manipulation region location (RL): focusing on locating the manipulated regions in a partially fake audio in which the original utterances are manipulated with real or generated audio [13].

Track 3. deepfake algorithm recognition (AR): aiming to recognize the algorithms of the deepfake utterances, and the evaluation dataset includes samples from an unknown deepfake algorithm [17, 18].

2.1. Training and dev sets

The training and dev. sets of ADD 2023 contain four subsets, as summarized in Table 1.

Table 1
Numbers of utterances in training and dev. sets of each task

Subchallenges	Training		Dev.	
	#Real	#Fake	#Real	#Fake
Track 1.2	3,012	24,072	2,307	26,017
Track 2	26,554	26,539	8,914	8,910
Track 3	3,200	19,200	1,200	7,200

Table 2
Overview of the numbers of utterances with different labels for the training set, dev. set and test set of Track 3. The unknown category is labeled 7.

Label	Training	Dev.	Test
#0	3,200	1,200	9,512
#1	3,200	1,200	10,474
#2	3,200	1,200	7,169
#3	3,200	1,200	10,461
#4	3,200	1,200	10,391
#5	3,200	1,200	10,507
#6	3,200	1,200	10,507
#7	–	–	10,469

Track 1.1: We use the AISHELL-3 [19] corpus, which is a large-scale Chinese speech corpus containing over 88,000 utterances, composing 85 hours of speech.

Track 1.2: We use the same training and dev. sets as Track 3.2 of ADD 2022, including the real and fake utterances based on AISHELL-3.

Track 2: The dataset consists of real utterances and partially fake utterances. Fake utterances generated by manipulated the original genuine utterances with real or synthesized audios.

Track 3: The training and dev. sets include 7 classes (1 real and 6 counterfeit) as shown in Table 2. The 7 categories are labeled 0, 1, 2, 3, 4, 5, 6. Fake audio taken from speech synthesized by different speech generation algorithms and tools.

2.2. Test sets

The test sets of ADD 2023 are more challenging compared to the previous one. The number of utterances in the four subsets are shown in Table 3.

Table 3
Numbers of utterances in the test sets of each task

Test	Track 1.2		Track 2	Track 3
	R1	R2		
#Real	80,000	87,500	20,000	10,507
#Fake	31,976	30,977	30,000	68,983

Track 1.1: It consists of test sets for two rounds, with two speakers, one male and one female, randomly selected from the AISHELL3 dataset in each round. There are 499 text contents in the test set file, and the text content of each line corresponds to an audio file generated for each target speaker ID.

Track 1.2: The real audio of the test set for two rounds consists of sources including AISHELL-1 [20], Thchs30 [21], etc. The fake audio consists of audio generated by using TTS and voice conversion techniques, and a portion of audio generated from the two rounds of track 1.1 submissions.

Track 2: The test set includes unseen partially fake and real utterances. Additional noise addition and format conversions were done on this base.

Track 3: The test set includes 8 classes (the 7 classes included in the training and dev. sets and unknown counterfeit class, as shown in Table 2). The unknown category data was synthesized by an unknown speech generation tool.

3. Evaluation metrics

Track 1.1 aims to generate fake audio that can fool the detection models. Therefore, the deception success rate (DSR) [12] is chosen as the metric. The goal of Track 1.2 is audio deepfake detection. So the weighted equal error rate (WEER) [12] is used as the metric. To better evaluate the manipulation region location performance of Track 2, the final score is the weighted sum of sentence accuracy and segment F$_1$-score [13]. For Track 3, participants should recognize the known and unknown algorithms of the deepfake utterances. Therefore, we utilize the macro-average F$_1$-score [22] in open set recognition.

3.1. Track 1.1 FG-G

DSR reflects the degree to which the audio deepfake detection model is deceived by the generated utterances, and is defined as follows:

$$DSR = \frac{W}{A \times N} \tag{1}$$

where W is the count of wrong detection samples by all the detection models on the condition of reaching each own equal error rate (EER) [8] performance, A is the total number of evaluation samples, and N is the number of detection models. For the first round, the DSR against the Track 1.2 submissions forms the totality of generation performance metric, where as in the second round, weighted consideration is also given to the DSR against the model we release, effectively:

$$WDSR = \gamma DSR_{R1} + \delta WDSR_{R2} \tag{2}$$

$$WDSR_{R2} = \alpha DSR_{R2baseline} + \beta DSR_{R2} \tag{3}$$

where γ=0.4, δ=0.6, α =0.4 and β=0.6, and they represent the weights for DSR in our consideration. $\mathrm{DSR}_{\mathrm{R1}}$ and $\mathrm{WDSR}_{\mathrm{R2}}$ represent the generation performance metrics for the first and second rounds, respectively. $\mathrm{DSR}_{\mathrm{R2}}$ and $\mathrm{DSR}_{\mathrm{R2baseline}}$ refers to the DSR achieved by using the synthesized speech submitted by the participants to attack the model submitted in track 1.2 and the detection baseline model [3] provided by organizers, respectively.

3.2. Track 1.2 FG-D

The WEER is defined as:

$$WEER = \alpha EER_{R1} + \beta EER_{R2} \tag{4}$$

where α =0.4 and β =0.6, which represent the weights of $\mathrm{EER}_{\mathrm{R1}}$ obtained in the first round and $\mathrm{EER}_{\mathrm{R2}}$ obtained in the second round, respectively. The EER is defined and calculated in the same way as in ADD 2022.

3.3. Track 2 RL

For Track 2, sentence accuracy measures the ability of the model to correctly distinguish between genuine and fake audio, and is defined as follows:

$$A_{sentence} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Table 4
Description of detection baseline systems

ID	Model	Features	Task
S01	GMM	LFCC	Track 1.2
S02	LCNN	LFCC	Track 1.2
S03	LCNN	Wav2vec2	Track 1.2
S04	LCNN	LFCC	Track 2
S05	ResNet (Softmax with threshold)	LFCC	Track 3
S06	ResNet (Openmax)	LFCC	Track 3

where TP, TN, FP, and FN denote the numbers of true positive, true negative, false positive, and false negative samples. Additionally, we use Segment Precision $\mathrm{P}_{\mathrm{segment}}$, Segment Recall $\mathrm{R}_{\mathrm{segment}}$, and Segment F$_1$-score $\mathrm{F}_{1_{\mathrm{segment}}}$ to measure the ability of the model to correctly identify fake areas from fake audios, defined respectively as:

$$P_{segment} = \frac{TP}{TP + FP} \tag{6}$$

$$R_{segment} = \frac{TP}{TP + FN} \tag{7}$$

$$F_{1segment} = \frac{2 \times P \times R}{P + R} \tag{8}$$

The final score is the weighted sum of Sentence Accuracy and Segment F$_1$-score, as shown below.

$$Score = \alpha A_{sentence} + \beta F_{1segment} \tag{9}$$

where α =0.3 and β =0.7, which represent the weights of $A_{sentence}$ and $F_{1segment}$.

3.4. Track 3 AR

For the algorithm recognition tasks in Track 3, we use the macro-average F$_1$-score, defined as:

$$P = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i} \tag{10}$$

$$R = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FN_i} \tag{11}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{12}$$

where C denotes the number of known classes, TP_i, FP_i and FN_i denote the true positive, false positive, and false negative samples of class i [22]. Note that while the formulae iterate only over known classes, FP_i and FN_i take unknown class samples into consideration.

[3] https://github.com/asvspoof-challenge/2021/tree/main/DF/Baseline-RawNet2

Table 5
ADD 2023 Track 1.1 Rankings. The results of round 1 (R1) and round 2 (R2) were measured by DSR (%) and the final evaluation was performed with WDSR (%).

#	ID	DSR_{R1}	DSR_{R2}	$\text{DSR}_{\text{R2baseline}}$	WDSR	#	ID	DSR_{R1}	DSR_{R2}	$\text{DSR}_{\text{R2baseline}}$	WDSR
1	A01	37.91	49.60	49.80	44.97	9	A09	0.00	35.55	24.85	18.76
2	A02	37.80	27.81	77.05	43.63	10	A10	30.71	17.28	0.10	18.53
3	A03	43.20	51.58	23.45	41.48	11	A11	23.58	16.00	6.71	16.80
4	A04	33.16	36.25	51.30	38.63	12	A12	41.72	0.00	0.00	16.69
5	A05	36.63	38.52	36.77	37.35	13	A13	40.12	0.00	0.00	16.05
6	A06	38.14	36.66	9.32	30.69	14	A14	0.00	0.00	0.00	0.00
7	A07	39.68	22.79	25.45	30.18		Avg.				27.11
8	A08	34.83	29.03	5.51	25.71						

Table 6
ADD 2023 Track 1.2 Rankings. The results of round 1 (R1) and round 2 (R2) were measured by EER (%) and the final evaluation was performed with WEER (%).

#	ID	EER_{R1}	EER_{R2}	WEER	#	ID	EER_{R1}	EER_{R2}	WEER	#	ID	EER_{R1}	EER_{R2}	WEER
1	B01	11.56	13.05	12.45	19	B18	40.31	33.55	36.25	37	B34	38.06	100.0	75.22
2	B02	21.11	15.82	17.93	20	B19	54.71	28.46	38.96	38	B35	38.06	100.0	75.47
3	B03	23.44	21.26	22.13	21	B20	32.90	45.04	40.18	39	B36	41.34	100.00	76.53
4	B04	23.51	21.75	22.45	22	B21	44.29	41.35	42.52	40	B37	42.03	100.00	76.81
5	B05	24.06	22.59	23.17	23	B22	41.91	43.76	43.02	41	B38	42.38	100.00	76.95
6	B06	32.80	21.38	25.94	24	B23	49.52	40.23	43.94	42	B39	43.68	100.0	77.47
7	B07	32.73	24.94	28.05	25	B24	42.83	45.24	44.27	43	B40	44.50	100.0	77.80
8	B08	27.81	28.46	28.20	26	B25	42.73	51.90	48.23	44	B41	44.93	100.0	77.97
9	B09	32.91	25.50	28.46	27	B26	44.61	51.48	48.73	45	B42	47.04	100.0	78.81
10	B10	32.87	25.81	28.63	28	**S01**	36.90	63.80	53.04	46	B43	47.04	100.0	78.81
11	B11	35.09	24.81	28.92	29	B27	100.00	37.76	62.65	47	B44	47.70	100.0	78.96
12	**S03**	21.71	35.61	30.05	30	B28	100.00	38.80	63.28	48	B45	53.61	100.00	81.44
13	B12	32.37	29.93	30.90	31	**S02**	61.25	70.37	66.72	49	B46	54.77	100.00	81.90
14	B13	33.34	30.50	31.63	32	B29	25.94	100.00	70.37	50	B47	54.91	100.0	81.96
15	B14	34.77	29.78	31.77	33	B30	100.00	53.62	72.17	51	B48	58.19	100.0	83.27
16	B15	38.75	30.08	33.54	34	B31	35.33	100.00	74.13	52	B49	65.00	100.0	86.00
17	B16	32.37	34.43	33.60	35	B32	36.98	100.00	74.79		Avg.			49.94
18	B17	38.08	34.20	35.75	36	B33	100.00	58.04	74.82					

4. Challenge results

ADD 2023 has challenge data requests from 145 teams from 12 countries. Participants submit task results and receive scores through the CodaLab website. In this section, we report on the detection baselines provided by the organizers and the results and analysis submitted by the participants.

4.1. Detection baselines

ADD 2023 provides six baseline systems, which are described in summary as shown in Table 5. For the detection task of track 1.2, we present three different detection systems. The first system is a GMM-based system that operates on linear frequency cepstral coefficients (LFCCs) [23] (baseline S01). The feature extraction of LFCC is the same as that of ASVspoof 2021, where the window length and shift are set to 30 ms and 15 ms. The second LFCC-LCNN system (baseline S02) operates on LFCC features with a light convolutional neural network (LCNN) [24]. The frame length and shift are set to 20 ms and 10 ms. The back end is based on the LCNN reported in [24]. The third system operates on wav2vec2 features with an LCNN (baseline S03). The wav2vec2 [25] pretrained model variant "wav2vec XLSR" is used as a pretrained feature extractor, which is trained on 56k hours of audio samples in 53 languages using additional linear transformations and a larger context network.

For the detection task of track 2, the front-end LFCC feature extraction settings of the baseline system S04 are the same as those of S02. For back-end model architecture, we remove all pooling layers from the conventional LCNN to ensure the output size aligns with the segment label. For the recognition task of track 3, we introduce two different recognition systems. Both baselines are LFCC-ResNet based systems. The LFCC were extracted similar to baseline system S01. The model struc-

Table 7
ADD 2023 Track 2 Rankings. The results were performed by score (%).

#	ID	Score	#	ID	Score	#	ID	Score
1	C01	67.13	7	C07	53.99	13	**S04**	42.25
2	C02	62.49	8	C08	50.86	14	C13	42.11
3	C03	62.02	9	C09	48.55	15	C14	38.74
4	C04	59.62	10	C10	45.39	16	C15	27.57
5	C05	59.12	11	C11	44.65	17	C16	18.80
6	C06	56.63	12	C12	43.50		Avg.	48.82

Table 8
ADD 2023 Track 3 Rankings. The results were performed by F_1-score (%).

#	ID	F_1	#	ID	F_1	#	ID	F_1
1	D01	89.63	6	D06	73.50	11	**S05**	53.50
2	D02	83.12	7	D07	72.05	12	D10	21.10
3	D03	75.41	8	D08	68.15	13	D11	11.73
4	D04	73.55	9	D09	63.78		Avg.	62.87
5	D05	73.52	10	**S06**	54.16			

ture of ResNet based on [26] was adopted from ResNet-18. Baseline S05 used a simple method-softmax with threshold, a thresholding procedure on probability to identify whether a speech signal is generated by a known or an unknown deepfake algorithm. Baseline S06 used the traditional open-set recognition method of OpenMax [27].

4.2. Results and analysis

The four tracks of ADD 2023 have all received sufficient submissions, and the summary data of the rankings are shown in Tables 5, 6, 7, and 8. The ID number of each participating team is determined by their ranking order.

Track 2 and 3 are the first subchallenge of fake region location and algorithm recognition in the field of deepfake audio detection. For track 1.1, we received 14 submissions. The average WDSR of all submissions was 27.11%, and the two-round combined performance of the best team was 44.97%. Track 1.2 received a total of 49 submissions, with 11 WEER below the best baseline S01, and the best team had a WEER of 12.45%. The average WEER of all submissions was 49.94%. For Track 2, 11 teams scored higher than the baseline S04, with the highest score of 67.13%. The average score of the 16 submissions was 48.82%. The results show that it is challenging for fake region location.

For Track 3, Nine teams performed better than the baseline systems S06 and S07. Although the best team achieved an F_1-score of 89.63%, the average F_1-score of Track 3 is still low. We hope that the challenge data and evaluation results of track 3 will further promote researchers to explore new deepfake audio algorithm recognition methods.

5. Conclusions

This paper provides an overview of the ADD 2023 Challenge, which consists of four distinct subchallenges. In order to better simulate real-world challenges, the challenge introduces two new tasks and more difficult datasets. The results indicate that the fake region location task and the algorithm recognition task are still challenging, especially for fake region location track. The solutions of participants and further analysis will be presented at the ADD 2023 workshop. In future competitions, we plan to optimize the datasets and competition rules, aiming to promote more advanced research in the deepfake audio community.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) (No. 61831022, No. U21B2010, No. 62101553, No. 61971419, No. 62006223, No. 62276259, No. 62201572, No. 62206278), Beijing Municipal Science&Technology Commission, Administrative Commission of Zhongguancun Science Park No. Z211100004821013, Open Research Projects of Zhejiang Lab (NO. 2021KH0AB06). Thanks to AISHELL [4] for providing the open source dataset for this challenge.

References

[1] X. Tan, T. Qin, F. Soong, T.-Y. Liu, A survey on neural speech synthesis, arXiv preprint arXiv:2106.15561 (2021).

[2] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, Grad-tts: A diffusion probabilistic model for text-to-speech, in: International Conference on Machine Learning, PMLR, 2021, pp. 8599–8608.

[3] J. Kim, J. Kong, J. Son, Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech, in: International Conference on Machine Learning, PMLR, 2021, pp. 5530–5540.

[4] B. Sisman, J. Yamagishi, S. King, H. Li, An overview of voice conversion and its challenges: From statistical modeling to deep learning, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2020) 132–157.

[5] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, Z. Ling, The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods , in: Proc.

[4] https://www.aishelltech.com

The Speaker and Language Recognition Workshop (Odyssey 2018), 2018, pp. 195–202. doi:10.21437/Odyssey.2018-28.

[6] Z. Yi, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z.-H. Ling, T. Toda, Voice Conversion Challenge 2020 — Intra-lingual semi-parallel and cross-lingual voice conversion —, in: Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, 2020, pp. 80–98. URL: http://dx.doi.org/10.21437/VCC_BC.2020-14. doi:10.21437/VCC_BC.2020-14.

[7] D. Harwell, Remember the 'deepfake cheerleader mom'? prosecutors now admit they can't prove fake-video claims, March 14 (2021) 2021.

[8] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilc¸i, et al., Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge, in: Proc. of INTERSPEECH, 2015.

[9] T. Kinnunen, M. Sahidullah, H. Delgado, N. E. M. Todisco, et al., The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection, in: Proc. of INTERSPEECH, 2017.

[10] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, K. Lee, Asvspoof 2019: Future horizons in spoofed and fake audio detection, in: Proc. of INTERSPEECH, 2019.

[11] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection, 2021.

[12] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, et al., Add 2022: the first audio deep synthesis detection challenge, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 9216–9220.

[13] J. Yi, Y. Bai, J. Tao, H. Ma, Z. Tian, C. Wang, T. Wang, R. Fu, Half-truth: A partially fake audio detection dataset, in: Proc. of INTERSPEECH, 2021.

[14] H. Ma, J. Yi, J. Tao, Y. Bai, Z. Tian, C. Wang, Continual learning for fake audio detection, in: Proc. of INTERSPEECH, 2021.

[15] B. Peng, H. Fan, W. Wang, J. Dong, Y. Li, S. Lyu, Q. Li, Z. Sun, H. Chen, B. Chen, et al., Dfgc 2021: A deepfake game competition, in: 2021 IEEE International Joint Conference on Biometrics (IJCB), IEEE, 2021, pp. 1–8.

[16] B. Peng, W. Xiang, Y. Jiang, W. Wang, J. Dong, Z. Sun, Z. Lei, S. Lyu, Dfgc 2022: The second deepfake game competition, in: 2022 IEEE International Joint Conference on Biometrics (IJCB), IEEE, 2022, pp. 1–10.

[17] X. Yan, J. Yi, J. Tao, C. Wang, H. Ma, T. Wang, S. Wang, R. Fu, An initial investigation for detecting vocoder fingerprints of fake audio, in: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia, 2022, pp. 61–68.

[18] X. Yan, J. Yi, J. Tao, C. Wang, H. Ma, Z. Tian, R. Fu, System fingerprints detection for deepfake audio: An initial dataset and investigation, arXiv preprint arXiv:2208.10489 (2022).

[19] Y. Shi, H. Bu, X. Xu, S. Zhang, M. Li, Aishell-3: A multi-speaker mandarin tts corpus and the baselines, in: arXiv preprint arXiv:2010.11567, 2020.

[20] H. Bu, J. Du, X. Na, B. Wu, H. Zheng, Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline, in: 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), 2017, pp. 1–5. doi:10.1109/ICSDA.2017.8384449.

[21] D. Wang, X. Zhang, Thchs-30: A free chinese speech corpus, arXiv preprint arXiv:1512.01882 (2015).

[22] C. Geng, S.-j. Huang, S. Chen, Recent advances in open set recognition: A survey, IEEE transactions on pattern analysis and machine intelligence 43 (2020) 3614–3631.

[23] M. Sahidullah, T. Kinnunen, C. Hanilçi, A comparison of features for synthetic speech detection, in: Proc. of INTERSPEECH, 2015.

[24] Z. Wu, R. K. Das1, J. Yang, H. Li, Light convolutional neural network with feature genuinization for detection of synthetic speech attacks, in: Proc. of INTERSPEECH, 2020.

[25] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[26] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[27] A. Bendale, T. E. Boult, Towards open set deep networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1563–1572.