The USTC-NERCSLIP System for the Track 1.2 of Audio Deepfake Detection (ADD 2023) Challenge*

Haochen Wu[1,†], Zhuhai Li[1,†], Luzhen Xu[1], Zhentao Zhang[2], Wenting Zhao[2], Bin Gu[1], Yang Ai[1], Yexin Lu[1], Jie Zhang[1,*], Zhenhua Ling[1] and Wu Guo[1]

[1]University of Science and Technology of China, Hefei, 230027, China
[2]China Merchants Bank, Shenzhen, 518057, China

Abstract
This paper describes the system of USTC-NERCSLIP submitted to the track 1.2 of the second Audio Deepfake Detection Challenge (ADD 2023). Our system consists of a wav2vec2.0-based front-end feature extractor and an AASIST-based back-end classifier. To further solve the problem of the gap in the noise and synthesis algorithms between the training and evaluation sets, we propose a multi-level data augmentation method. Specifically, we add a variety of noises to the training set to simulate the noise environment of the evaluation set. Besides, we use several vocoders to synthesize fake audio based on the genuine audio in the training set to enrich the synthesis algorithms. Results show that the proposed method achieves a WEER of 12.45% on the two-round evaluation with a single system, which ranks the top among all submissions.

Keywords
ADD 2023, data augmentation, wav2vec 2.0, speech synthesis, vocoder.

1. Introduction

Over the last decades, the development of artificial intelligence (AI) has in turn contributed to rapid advances in speech synthesis and voice conversion applications. Deep learning models can generate realistic and human-like speech, which has a wide range of applications in human-computer interaction, smart home, entertainment, education, etc [1]. Nevertheless, they also bring a potential to pose a serious threat to the society if someone misuses it, e.g., using fake audio to commit fraud or mislead public opinions. Therefore, in order to improve the speech security, detecting fake audio is essential to reduce the threat posed by the disinformation embedded in speech. Also, deepfake audio detection can help to address the serious vulnerability of automated speaker verification systems against various malicious spoofing attacks [2].

Guarding against such abuse and misuse, deepfake audio detection is therefore an interesting emerging topic in the AI community. The ASVspoof Challenge [3, 4, 5, 6] is held every two years dedicated to spoofing speech detection, including text to speech synthesis (TTS) [7, 8], voice conversion (VC) [9], speech replay and impersonation. To further address diversified and challenging attack situations in realistic applications, the first Audio Deepfake Detection Challenge (ADD 2022) [10] extends the attack situations of fake audio detection. Different from ADD 2022, the second ADD Challenge (ADD 2023) [11] focuses on surpassing the constraints of binary real/fake classification, localizing the manipulated intervals in a partially fake speech as well as pinpointing the source algorithm used to generate the fake audio. The ADD 2023 Challenge contains three tracks, among which the Track 1 is an audio fake game (FG) consisting of an audio generation task (Track 1.1) and a fake audio detection task (Track 1.2). For Track 1.1, participants aim to generate fake audio that can spoof the detection systems of Track 1.2. For Track 1.2, participants aim to detect fake utterances, especially the fake samples generated from Track 1.1. The two tracks represent a more realistic situation of what anti-spoofing researchers need to deal with day to day.

Recently, self-supervised learning (SSL) has achieved significant advances in the fields of natural language processing (NLP) [12], automatic speech recognition [13] as well as speaker verification [14]. It was shown that building a general pre-trained model based on the exploitation of a large mount of unlabeled data can be quite essential to boost the performance of many downstream tasks, reduce data labeling efforts and lower entry barriers for individual tasks. To our knowledge, only a few works have used self-supervised pre-trained models for fake audio detection. In this work, we thus make efforts to use an open-sourced self-supervised pre-trained model as the feature extractor to help build a robust ADD system.

IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023), August 19, 2023, Macao, S.A.R
*This work was supported by the USTC-CMB Joint Laboratory of Artificial Intelligence, the National Natural Science Foundation of China (62101523), Hefei Municipal Natural Science Foundation (2022012) and USTC Research Funds of the Double First-Class Initiative (YD2100002008).
*Corresponding author.
†These authors contributed equally.
whc1414858026@mail.ustc.edu.cn (H. Wu); snowsea@mail.ustc.edu.cn (Z. Li); jzhang6@ustc.edu.cn (J. Zhang)

This paper presents our submitted system to the Track 1.2 of the ADD 2023 Challenge. Due to the diversified noise and rich synthesis algorithms in the evaluation set, it is difficult to obtain the desired performance using the training set directly. We thus propose a multi-level data augmentation (DA) method to address this problem. The first-level DA aims to develop a robust system to noise, reverberation and channel variation. We use the public noise, reverberation datasets and RawBoost DA tool [15] to conduct online DA. On the other hand, we adopt the speed perturbation [16] and compression coding [17] methods for offline DA. We find that these techniques have a little contribution to fix the gap in the number of synthesis algorithms between the training and evaluation sets. Therefore, the second level of our DA method aims to increase the variety of the synthesis algorithms in the training set. Specifically, we use several vocoders to synthesize fake audio based on the genuine audio in the training set. Experimental results show that the proposed method outperforms all other submissions with a weighted equal error rate (WEER) of 12.45% in Track 1.2.

The remainder of this paper is organized as follows. Section 2 and 3 describe the proposed multi-level DA method and model architecture in detail, respectively. Section 4 introduces the experimental setup, followed by experimental results in Section 5. Finally, Section 6 concludes this work.

2. Data Augmentation

In this section, we will show a detailed description of the proposed multi-level DA method from two aspects, i.e., for diverse noise and synthesized speech, respectively.

2.1. DA for Diversified Noise

To reduce over-fitting and bias caused by diversified noise in real scenes, we apply augmentation methods from three aspects: noise, reverberation and channel variation.

For the noise, on one hand, we add recorded noises and distortion from MUSAN [18] dataset to the clean audio, which is commonly used in other speech-related fields. On the other hand, by utilizing the RawBoost DA technique, we add different nuisance noises dependent on the corresponding raw waveform inputs. Based on a variety of convolutional and additive noises, RawBoost models nuisance variability stemming such as encoding, transmission, microphones and amplifiers, as well as linear and nonlinear distortions. RawBoost consists of three independent noises [15]: 1) linear and non-linear convolutive noise; 2) impulsive signal-dependent additive noise; 3) stationary signal-independent additive noise. The details are available in [15]. Finally, the noises are mixed at a random signal-to-noise ratio (SNR) ranging from 10 dB to 30 dB.

For the reverberation, we add distortion to the clean training set using the public room impulse response (RIR) [19]. It should be pointed out that the reverberation and noise are added to the audio sequentially. First, we add reverberation at a certain probability p, and then noise is added at a certain probability q. Besides, unlike traditional DA, which enlarge the training set, we add nuisance variability to the existing training data online.

For the channel variation, we apply a variety of codec algorithms [20], including MP3, OGG, AAC OPUS, a-law and μ-law. Besides, to mock the telephony transmission loss, audio samples are first downsampled to 8kHz and then upsampled back to 16kHz. To add more spices, we also consider speed perturbation to help improve the performance. Considering the high computational costs of compression coding and speed perturbation, we use them in an offline manner. The amount of the resulting training set is increased by three times compared to the original training data.

2.2. DA for Synthesis Algorithms

Most DA methods focus on improving the generalization of the system in real scenes, such as the methods in Section 2.1, which, however, cannot cover the speech synthesis algorithms in the training set. To tackle this issue, we use several vocoders to synthesize fake audio, including traditional and neural vocoders. The traditional vocoder can directly synthesize fake audio after determining the parameters of the algorithm, while the neural vocoder can only synthesize audio with the generator trained on the training set. In this work, we choose three traditional vocoders (TV) and one neural vocoder (NV). The details are given as follows.

- ***Griffin-Lim***[21]: This is a traditional vocoder, which synthesizes audio using the true mel-spectrum. It uses a phase reconstruction method based on the redundancy of the short-time Fourier transform and promotes the consistency of a spectrogram. The mel-spectrum is first used to estimate the amplitude spectrum, which is then used to estimate the audio waveform.
- ***WORLD*** [22]: World shows a superiority in not only the sound quality, but also the complexity (thus being appropriate for real-time cases). It includes three parameter estimation modules (for estimating F0, spectral envelop, aperiodic parameter extraction, respectively), followed by a synthesis module to generate the speech-like signals.
- ***RAAR*** [23]: RAAR is extensively used in optics, and its effectiveness has been validated in many applications. In [23], Tomoki et al. applied a phase reconstruction algorithm to acoustic scenarios, where the evaluated acoustical metrics

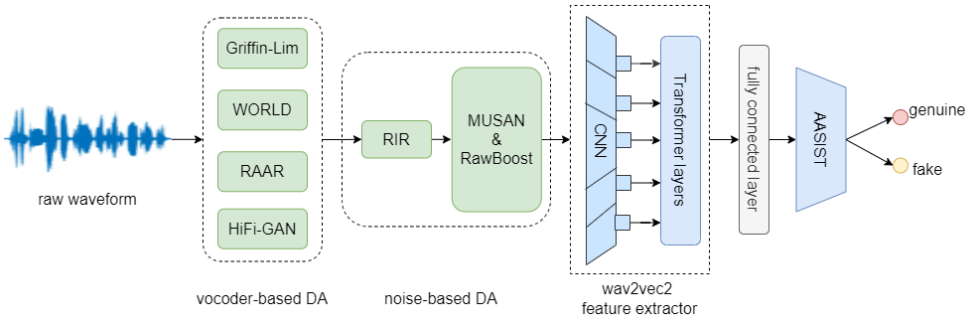

Figure 1: The overall flow diagram of our system, which includes the vocoder-based and noise-based DA modules, the wav2vec2 based feature extractor and the AASIST-based classifier.

show that RAAR is robust against noise and performs well for both small and large number of iterations.

- ***HiFi-GAN*** [8]: As one of the state-of-the-art neural vocoders, HiFi-GAN generates audio based on generative adversarial networks (GANs) using the true mel-spectrum. It includes one generator and two discriminators: multi-scale and multi-period discriminators, which can achieve efficient and high-fidelity speech synthesis. The generator and discriminators are trained adversarially by incorporating two additional losses to improve the training stability and model performance.

3. Methodology

The diagram of our system for the ADD 2023 Challenge is illustrated in Figure 1, which follows a fully automated end-to-end pipeline consisting of two modules: feature extractor and classifier. The wav2vec2 based feature extractor aims to extract a high-level representation of the speech. The classification module is a modified version of AASIST [24] with the removal of the sinc convolutional layer based front-end.

3.1. Wav2vec2 Front-End

Wav2vec2 [25] is a self-supervised pre-trained model, which can extract speech representations or embeddings from raw waveform. It has shown an impressive performance on many downstream tasks, particularly on automatic speaker recognition. As a variant, XLS-R [26] is a new self-supervised cross-lingual speech representation model based on wav2vec2, which scales the number of languages, the amount of training data as well as the model size. XLS-R is pre-trained on 436K hours of unannotated speech in 128 languages. Although wav2vec2 and XLS-R follows the same model architecture, the features extracted from XLS-R can be more general due to the more fruitful data resource, leading to a stronger feature reliability and domain robustness. Therefore, we use XLS-R as the feature extractor for the ADD task.

The XLS-R model mainly includes three stages. Firstly, the raw waveform is sent into a feature encoder composed of several convolutional layers (CNN). The feature encoder extracts vector representations of size 1024 every 20ms and the receptive field is 25ms. Secondly, these encoder embeddings are fed into the context encoder, which contains 24 transformer block layers and is used to explore the contextual information contained in the input speech. At the third stage, the feature encoder representation is processed by a quantization module to obtain a quantized representation. Then, the model is trained in a self-supervised manner with a contrastive loss by using the contextual representations to predict the masked counterparts at certain positions.

3.2. AASIST Back-End

AASIST is an end-to-end **A**udio **A**nti-**S**poofing system using **I**ntegrated **S**pectro-**T**emporal graph attention networks, which won the top rank in ASVspoof 2019 logical access (LA). It is an extension of RawGAT-ST [27] with three modifications: 1) a novel heterogeneous stacking graph attention layer, which models artefacts spanning heterogeneous temporal and spectral domains with a heterogeneous attention mechanism and a stack node, 2) a max graph operation that involves a competitive selection of artefacts, and 3) a modified readout scheme. AASIST uses a sinc convolutional layer based front-end, and thus can extract representations directly from raw waveform inputs.

In [28] Tak et al. tried to improve the generalization and domain robustness using a pre-trained, self-supervised model with fine-tuning. Specifically, the sinc convolution layer of AASIST is replaced by the afore-

Table 1
The details of the four vocoders.

vocoder	feature dim	hop	win	fft
Griffin-Lim	160	80	400	2048
WORLD	-	80	-	1024
RAAR	512	80	320	1024
HiFi-GAN	80	256	1024	1024

mentioned wav2vec2 model. Besides, a fully connected layer after the pre-trained model is used to reduce the representation dimension from 1024 to 128. In this work, we adopt the same model architecture as in [28].

4. Experimental Setup

This section introduces the dataset used in experiments, parameter configurations for our multi-level DA method and implementation details of our system.

4.1. Dataset

The Track 1.2 of ADD 2023 aims to distinguish fake audios from genuine ones, which are fully fake utterances generated by text-to-speech or voice-conversion algorithms. The training set consists of 3012 genuine utterances and 24072 fake utterances. The development set consists of 2307 genuine utterances and 26017 fake utterances. Besides, there are 111976 and 118477 utterances in the first and second round evaluation sets, respectively, where the second contains noise and fake audio generated by the teams participated in the Track 1.1 using unknown synthesis algorithms.

We find that the audio in the training and development sets are quite clean and have identical data distribution, making the trained model performs well on the development set. To be more general, we thus choose to re-partition the training and development sets. First, we combine the training set and development set as a larger dataset. Then, we randomly select 50% of the fake audio and all the genuine audio from the dataset and combine them into the new training set. Furthermore, we apply our DA method on the new training set to enable the final system with a high robustness and performance.

4.2. Vocoders

The three traditional vocoders synthesize fake audio from different input features. The mel-spectrum and Fourier amplitude spectrum are used for the Griffin-Lim and RAAR, respectively, while fundamental frequency, spectral envelope and aperiodic parameter estimated by WORLD are used to synthesize audio. More details about the vocoders are summarized in Table 1. For the neural vocoder HiFi-GAN, we only use all of the 5319 genuine audios for training, without using any other external data. We directly use the trained model to synthesize fake audio based on the genuine audio used for training. Note that HiFi-GAN takes the mel-spectrum as input.

Each vocoder can synthesize 5319 fake audios from all the genuine audios. The resulting 21276 fake audios are incorporated altogether in the training set.

4.3. Implementation Details

In experiments, the audio streams are truncated or repeated to a duration of 6 seconds during the train stage. The probabilities p and q in Section 2.1 are set to $\frac{1}{3}$ and $\frac{1}{5}$, respectively. During fine-tuning, the pre-trained wav2vec2 model is optimized jointly with the AASIST via the back-propagation. We use the standard Adam optimizer [29], which adopts a mini-batch size of 16 and a learning rate of 10^{-5} with a weight decay of 10^{-4} to avoid over-fitting. Since the result after two epochs of training is always worse than that obtained from only one epoch, all models are fine-tuned for only one epoch on two RTX 3090 GPUs. Considering the imbalance between the genuine and fake audios in the training set, we use the weighted cross entropy to minimize the training loss. The weights associated with the genuine and fake categories are set to 0.9 and 0.1, respectively.

5. Results

In this section, we show the performance of the proposed multi-level DA method based ADD system. The results depend on the equal error rate (EER) and the final score is weighted equal error rate (WEER), which is defined as

$$\mathrm{WEER} = \alpha * \mathrm{EER}_{\mathrm{R1}} + \beta * \mathrm{EER}_{\mathrm{R2}} \tag{1}$$

where $\alpha = 0.4$, $\beta = 0.6$, $\mathrm{EER}_{\mathrm{R1}}$ and $\mathrm{EER}_{\mathrm{R2}}$ denotes the EERs obtained in the two rounds of Track 1.2.

Model Comparison: As our system adopts the wav2vec 2.0 front-end and AASIST back-end, in order to show the respective function we conduct several comparisons. First, we compare the AASIST with the Light Convolutional Neural Networks (LCNN), which adopts the same architecture as [30] in the ADD 2022 Track 3.2. Besides, the LCNN takes the STFT as input features instead of the raw waveform. The results for the AASIST with the sinc-layer front-end or the wav2vec 2.0 front-end are presented in Table 2. For the training set, DA for diversified noise and the three traditional vocoders are used. Besides, we only show the results on the first round of the evaluation set.

Since the LCNN and AASIST with the sinc-layer front-end do not use the pre-trained model, we train them differently. Specifically, the Adam optimizer is adopted

Table 2
Comparison of different models.

Model	Front-end	EER(%)↓
LCNN	STFT	36.85
AASIST	sinc-layer	32.19
AASIST	wav2vec 2.0	25.45

Table 3
Ablation study for vocoder-based DA.

Model	TV set	NV set	EER(%)↓
wav2vec2	✗	✗	40.53
& AASIST	✓	✗	25.45
	✓	✓	11.56

Table 4
Ablation study for MUSAN and RawBoost

Model	MUSAN	RawBoost	EER(%)↓
	✗	✗	19.15
wav2vec2	✗	✓	17.87
& AASIST	✓	✗	13.55
	✓	✓	11.56

with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and weight decay 10^{-4}. The batch size is set to 32. The learning rate is initially set to 0.0003 with 50% decay for every 10 epochs. We train the network for 100 epochs and select the model with the lowest EER on the development set as the final model for evaluation.

From Table 2, we can see that AASIST with the sinc-layer front-end outperforms the LCNN. However, the EER is still unacceptably high with a poor robustness. Replacing the sinc-layer with the wav2vec 2.0 front-end can reduce the EER by 7.26%, which validates the benefit of using pre-trained model for the ADD task.

Comparison of Vocoders: Then, we conduct ablation studies on the proposed vocoder-based DA method. According to the types of vocoders, we combine the audios generated by the three traditional vocoders (TV) into the TV set, while the audios generated by the neural vocoder (NV) are regarded as the NV set. We train our system on different datasets by combining the training set with the TV set or NV set. Besides, the noise-based DA is used, but we only show the results on the first round of the evaluation set in Table 3. It is clear that the EER on the evaluation set is very high (e.g., 40.53%) even using the DA for diversified noise and the wav2vec 2.0 front-end. Applying the traditional vocoders, the EER drops to 25.45%, which can be further reduced to 11.56% in case of training on both the TV and NV sets. This reveals that the vocoders in combination with DA techniques are rather helpful to improve the ADD performance.

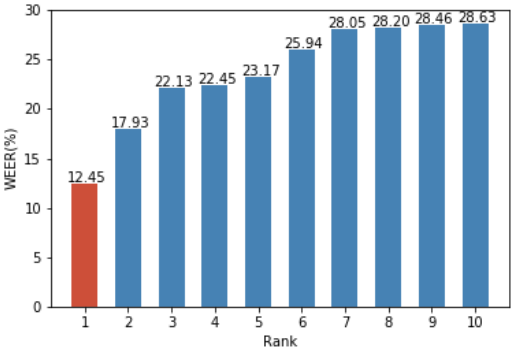

Figure 2: Summary of the top 10 submissions to Track 1.2.

Comparison of MUSAN and RawBoost: Apart from synthesis algorithms, the differences in the background noise, reverberation and channel variety between the training and evaluation sets also play an important role. However, due to the tight challenge schedule, we ignore the effect of the RIR, speed perturbation and compression coding. Here, we compare the influence of the MUSAN and RawBoost used for the online DA in Table 4, where note that vocoders are incorporated. We can see that using online noises leads to a significant EER decrease from 19.15% to 11.56%. It also shows that MUSAN is more beneficial than RawBoost to increase the diversity of noise in the training set.

Finally, it should be noticed that the performance slightly drops on the Round 2 evaluation (from 11.56% to 13.05%). This is probability due to the fact that the fake audio synthesized by participants in Track 1.1 are added into the evaluation set, which further increases the data diversity. More importantly, our system still shows its superiority and ranks the top in Round 2. In Figure 2, we summarize the overall WEER of the top 10 participants, where the teams are anonymized accordingly.

6. Conclusions

This paper presents the detailed system description of USTC-NERCSLIP submitted to the ADD Challenge 2023, which involves the wav2vec2-based feature extractor and the AASIST-based classifier. In addition, we proposed a multi-level DA method for the diversified noise and synthesis algorithms in the evaluation set, which was shown to largely improve the performance and robustness. Due to the new data source in the second-round evaluation, the performance slightly drops, but our system still ranks the first place in Track 1.2.

References

[1] N. M. Müller, K. Pizzi, J. Williams, Human perception of audio deepfakes, in: *Proc. DDAM 2022*, October 2022, pp. 85–91.

[2] C. Zhang, S. Ranjan, J. H. Hansen, An analysis of transfer learning for domain mismatched text-independent speaker verification, in: *Proc. Odyssey Workshop*, June 2018, pp. 182–186.

[3] Z. Wu, T. Kinnunen, N. Evans, et al., ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge, in: *Proc. INTERSPEECH 2015*, September 2015, pp. 2037–2041.

[4] T. Kinnunen, M. Sahidullah, H. Delgado, et al., The ASVspoof 2017 Challenge: assessing the limits of replay spoofing attack detection, in: *Proc. INTERSPEECH 2017*, August 2017, pp. 2–6.

[5] M. Todisco, X. Wang, V. Vestman, et al., ASVspoof 2019: Future horizons in spoofed and fake audio detection, in: *Proc. INTERSPEECH 2019*, September 2019, pp. 1008–1012.

[6] J. Yamagishi, X. Wang, M. Todisco, et al., ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection, in: *Proc. ASVspoof2021 Workshop*, 2021.

[7] A. v. d. Oord, S. Dieleman, H. Zen, et al., Wavenet: A generative model for raw audio, in: *Proc. ISCA*, 2016.

[8] J. Kong, J. Kim, J. Bae, Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, *Advances in Neural Information Processing Systems* 33 (2020) 17022–17033.

[9] T. Kaneko, H. Kameoka, Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks, in: *Proc. EUSIPCO*, 2018, pp. 2100–2104.

[10] J. Yi, R. Fu, J. Tao, et al., ADD 2022: the first audio deep synthesis detection challenge, in: *Proc. ICASSP*, May 2022, pp. 9216–9220.

[11] J. Yi, J. Tao, R. Fu, et al., ADD 2023: the second audio deepfake detection challenge, in: *Proc. IJCAI Workshop on DADA 2023*, August 2023.

[12] T. Brown, B. Mann, N. Ryder, et al., Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.

[13] C. Wang, Y. Wu, Y. Qian, et al., Unispeech: Unified speech representation learning with labeled and unlabeled data, in: *Proc. ICML*, 2021, pp. 10937–10947.

[14] Z. Fan, M. Li, S. Zhou, et al., Exploring wav2vec 2.0 on speaker verification and language identification, in: *Proc. INTERSPEECH*, 2021, pp. 1509–1513.

[15] H. Tak, M. Kamble, J. Patino, et al., Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing, in: *Proc. ICASSP*, May 2022, pp. 6382–6386.

[16] T. Ko, V. Peddinti, D. Povey, et al., Audio augmentation for speech recognition, in: *Proc. INTERSPEECH 2015*, September 2015, p. 3586–3589.

[17] T.-L. Vu, Z. Zeng, H. Xu, et al., Audio codec simulation based data augmentation for telephony speech recognition, in: *Proc. APSIPA ASC*, September 2019, pp. 198–203.

[18] D. Snyder, G. Chen, D. Povey, MUSAN: A music, speech, and noise corpus, in: *arXiv preprint arXiv:1510.08484*, 2015.

[19] T. Ko, V. Peddinti, D. Povey, et al., A study on data augmentation of reverberant speech for robust speech recognition, in: *Proc. ICASSP*, March 2017, pp. 5220–5224.

[20] R. K. Das, Known-unknown data augmentation strategies for detection of logical access, physical access and speech deepfake attacks: Asvspoof 2021, *Proc. ASVspoof2021 Workshop* (2021) 29–36.

[21] N. Perraudin, P. Balazs, P. L. Søndergaard, A fast griffin-lim algorithm, in: *Proc. WASPAA*, October 2013, pp. 1–4.

[22] 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.

[23] T. Kobayashi, T. Tanaka, K. Yatabe, et al., Acoustic application of phase reconstruction algorithms in optics, in: *Proc. ICASSP*, May 2022, pp. 6212–6216.

[24] J.-w. Jung, H.-S. Heo, H. Tak, et al., AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks, in: *Proc. ICASSP*, May 2022, pp. 6367–6371.

[25] A. Baevski, Y. Zhou, A. Mohamed, et al., Wav2vec 2.0: A framework for self-supervised learning of speech representations, *Advances in neural information processing systems* 33 (2020) 12449–12460.

[26] A. Babu, C. Wang, A. Tjandra, et al., XLS-R: Self-supervised cross-lingual speech representation learning at scale, in: *Proc. INTERSPEECH 2022*, 2022, pp. 2278–2282.

[27] H. Tak, J.-w. Jung, J. Patino, et al., End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection, in: *Proc. ASVspoof Workshop*, 2021.

[28] H. Tak, M. Todisco, X. Wang, et al., Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation, in: *The Speaker and Language Recognition Workshop*, June 2022.

[29] D. P. Kingma, J. L. Ba, Adam: A method for stochastic optimization, in: *Proc. ICLR*, 2015, pp. 1–15.

[30] H. Tak, M. Todisco, X. Wang, othets, Deepfake detection system for the add challenge track 3.2 based on score fusion, in: *Proc. DDAM 2022*, October 2022, pp. 43–52.