Description of a Multi-Stage Audio Spoofing System in ADD Challenge 2023

Hua Hua[1], Jingze Lu[1], Peiyang Shi[1], Zengqiang Shang[1], Yuxiang Zhang[1], Xuyuan Li[1] and Pengyuan Zhang[1,†]

[1]Institude of Acoustics, Chinese Academy of Sciences, 19 N.4th Ring West Rd., Haidian Dist., Beijing, China

Abstract

This paper is a detailed description of a multi-stage audio synthesis system that participated in Track1.1 (spoofing) in ADD Challenge 2023 in which we play the role of an attacker. These stages include a fully end-to-end text-to-speech model, a fully end-to-end any-to-many voice conversion model, and an adversarial attacking model. We believe that such a design can reduce the possibility of artifact exposure at multiple levels and dimensions so that it is closer to real speech and is becoming hard to extinguish. Besides, we adopt post-processing methods to further improve our spoofing capability against detection methods for non-speech parts. Our system won 3rd place in the total score of Track1.1 in this challenge, especially, the performance in attacking black-box systems ranked 1st.

Keywords

Speech synthesis, Anti-spoofing, Deception success rate, Adversarial attacking

1. Introduction

Audio deep synthesis techniques have been able to generate high-quality speech whose authenticity is difficult for humans to recognize. Meanwhile, as the quality of deeply synthesized speech approaches human's natural voice, anti-spoofing systems have emerged and been continuously upgraded for security purposes. So audio deepfake and anti-spoofing are in a game of attacking and defending. So far, there have been several authoritative challenges for speech deepfake and anti-spoofing tasks such as ASVSpoof-2017/2019/2021[1, 2, 3] and ADD-2022[4]. This ADD-2023 Challenge also continues this theme[5].

Track 1.1 of this challenge focuses on speech spoofing attacks against anti-spoofing systems. Each participating team is required to design a speech generation system, input the given text and speaker information, and generate the corresponding fake speech. Synthesized audio will be sent to a group of black-box anti-spoofing systems (from submissions of other tracks) and a white-box anti-spoofing system (the official baseline) for detection, and the deception success rate (DSR) will be used to measure the performance of a synthesizing system. A higher DSR means stronger deception ability, in other words, it can be an indication that the audio generated by the system is judged to be more authentic.

The system designed in our paper adopts a multi-stage speech generation method as shown in Figure 1. The first stage is a state-of-art text-to-speech (TTS) module based on hierarchical feature modeling, the second stage is an end-to-end any-to-many voice conversion (VC) module leveraging intermediate features from automatic speech recognition (ASR), and the third stage is a gradient-based adversarial attacking (AA) module. In addition, we also come up with some countermeasures against non-speech part detection, which we call the post-processing (PP) module.

Why are we building such a heavy cascading pipeline with so many stages? The motivation along with related works will be discussed in Section 2. The system framework and detailed methods adopted in different components will be described in Section 3. Training settings will be described in Section 4. Experiments and our rankings in the ADD Challenge will be demonstrated in Section 5.

2. Related works and motivation

TTS and VC are techniques that convert certain types of inputs to human speech waveforms – TTS accepts raw text or phoneme sequences, while VC takes in the voice of a source speaker. With the development of deep learning, research on TTS and VC based on deep neural networks has been carried out in large numbers.

TTS technology has made remarkable progress in its goal of generating natural speech which is close to human speaking[6, 7]. Representative contributions can be found in the autoregressive Tacotron family[8, 9], highly controllable Fastspeech and its variants[10, 11], Flow-

IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023), August 19, 2023, Macao, S.A.R

† Corresponding author.

huahua@hccl.ioa.ac.cn (H. Hua); lujingze@hccl.ioa.ac.cn (J. Lu); shipeiyang@hccl.ioa.ac.cn (P. Shi); shangzengqiang@hccl.ioa.ac.cn (Z. Shang); zhangyuxiang@hccl.ioa.ac.cn (Y. Zhang); lixuyuan@hccl.ioa.ac.cn (X. Li); zhangpengyuan@hccl.ioa.ac.cn (P. Zhang)

0000-0003-4979-924X (H. Hua)

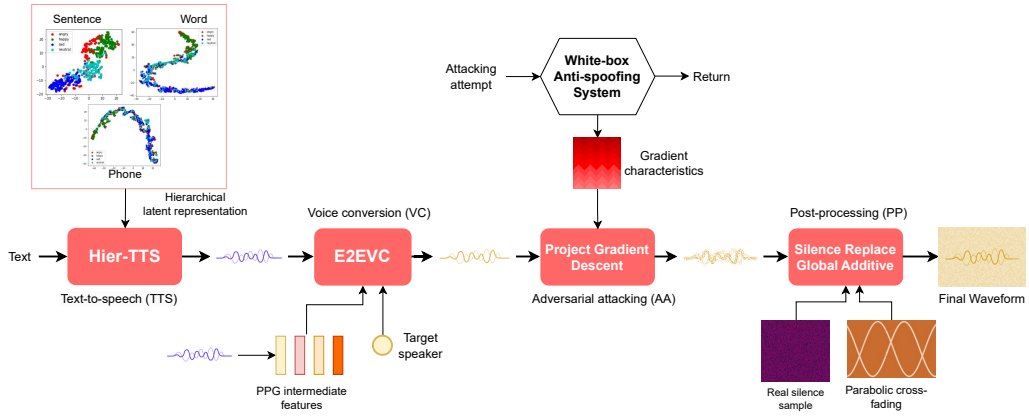

Figure 1: The overall pipeline of our proposed system.

based series[12, 13], fully end-to-end VITS[14], diffusion-based models[15, 16], etc.

Considering VC, models based on generative adversarial networks (GAN) along with their variants, such as StarGAN-VC[17] and CycleGAN-VC[18] use a generator or a conditional generator that transforms the source speaker's voice features to the target speaker's directly. Auto-encoder-based VC systems such as AutoVC[19] and VQVC[20] learn to reproduce their input as the output, exhibiting effectiveness in disentangling speaker identity information from linguistic content. Benefiting from automatic speech recognition (ASR) systems pre-trained with large corpora, models based on phonetic posteriorgrams (PPGs) are considered to have an advantage in extracting speaker-independent acoustic features from source speech[21, 22].

From the perspective of spoofing, we understand that most anti-spoofing systems do not target specific artifacts, but use more general features[23, 24, 25, 26]. Therefore, when synthesizing speech, we will try our best to move in a direction that is more authentic to human perception. But in the last year or two, the interpretability of antispoofing has become stronger and stronger, and some representative works have focused on the detection of artifacts generated by vocoders[27, 28]. Therefore, in this article, we consider making the VC stage a complete end-to-end form[29], and adopt certain strategies to alleviate artifacts such as chessboard effects. In addition, some counterfeiting studies have begun to notice the inadequacy of the acoustic model, for example, detecting some unnatural aspects of synthesized speech in long-term features that traditional anti-spoofing methods do not consider, such as prosody, style, and emotion[30, 31]. Therefore, our TTS stage adopts the latest Hier-TTS model of our team[32], which is mainly trained step by step through the hierarchical VAE structure, so that the generated speech has good long-term features at multiple levels such as sentence level, phrase level, word level, and frame level.

Additionally, machine learning models may misclassify examples that are only slightly different from correctly classified examples drawn from the data distribution[33]. In many cases, a wide variety of models with different architectures trained on different subsets of the training data misclassify the same adversarial example. This suggests that adversarial examples expose fundamental blind spots, so we can take advantage of this characteristic to create specific samples to attack the classification system, which is called an adversarial attack (AA). The representative methods include Fast Gradient Sign Method (FGSM)[33], and the Projected Gradient Descent (PGD) algorithm[34] involved in this paper is exactly a variant of FGSM. Such an AA stage may further compensate for the shortcomings of the aforementioned generative models, hiding possible exposed artifacts under adversarial perturbations especially confronting white-box anti-spoofing systems.

This is not enough. Some anti-spoofing systems even pay attention to non-speech parts like breath sounds and silent segments because generative models usually lack attention to these non-speech parts[35, 36]. Our VC system can reduce the false creation of unreal harmonic structures in breath sounds because its front end is an ASR-based phoneme classifier. Nevertheless, regarding silence, it is not that easy. In the ASVspoof 2019 and 2021 databases, the silence differences are important artifacts that influence countermeasures[2, 3]. We know that one real speech recording usually has natural silent segments at the beginning and the end, as well as between syllables, words, or phrases, which is deter-

mined by semantic or prosodic boundaries. On the other hand, synthesized speech especially through VC, tends to exhibit some anomalies different from real speech: no silence, completely zero amplitude, or unexpected noises. Although this will not cause severe problems in common application scenarios as human listeners will probably not find this problem, there is a high chance that these artifacts will be detected by an anti-spoofing system. Therefore, we adopt two non-neural post-processing (PP) algorithms that aim at implementing near-realistic silences and pauses against silence detection[29].

3. System framework and methods

The system designed is a multi-stage waveform generation pipeline. The first stage is a state-of-art text-to-speech TTS module based on hierarchical feature modeling. This module analyzes the input text to predict prosody and style features at the word level, phrase level, and sentence level, and finally generates speech waveforms with high naturalness in long-term features. The second stage is an end-to-end any-to-many VC module leveraging PPG features extracted from a pre-trained ASR. We introduce the desired voiceprint of the target speaker at this stage and transfer the aforementioned waveforms to the target speaker's timbre. The third AA stage is optional: It is difficult for us to impose this attack on the unknown black-box antispoofing systems, so we skip the stage; but confronting the white-box baseline system whose model is known, we use the adversarial attack method to further improve our spoofing ability. Finally, the generated speech waveform will also be processed in the PP stage to enhance the resistance to the detection of non-speech parts.

3.1. TTS with hierarchical variational autoencoders

We adopt Hier-TTS [32] as the very first stage of the whole model that analyzes and processes text inputs and returns raw speech waveform. Because of its powerful ability to model the unified space of text and audio, Hier-TTS can reconstruct factors such as prosody and style better than common TTS models, therefore it may achieve better performance against some recent anti-spoofing systems which take inner-speaker long-time consistency of prosody or style into consideration.

Figure 2 is the overall structure which contains a Hierarchical Audio Encoder (HAE), a Hierarchical Context Encoder (HCE), and a Hierarchical Audio Decoder (HAD). Hier-TTS introduce five latent variables at different temporal resolution, including sentence, word, subword, phoneme, and frame level. First, HAE extracts the hierarchical latent variables from the linear spectrogram in a fine-to-coarse manner, and then the HAD reconstructs the speech waveform from coarse-to-fine leveraging extracted hierarchical latent variables. To introduce text information, HCE obtains linguistic and phonological information at different scales from phoneme and character sequences and then injects them into each hierarchy of HAE and HAD. For the modeling of duration, we inject phoneme-scale durations at the phoneme level encoder and reconstruct the durations using the phoneme decoder. Thus, the duration and waveform reconstruction share part of the hierarchical hidden variables, which facilitates learning more consistent prosody.

3.2. Any-to-many VC based on phonetic posteriorgrams

The VC model features the structure proposed in [29], which can be divided into five components as shown in Figure 3. In short, it incorporates a conformer encoder from an auto-speech-recognition (ASR) model and a series of transformer blocks on one stream. On another stream, a posterior encoder upon linear spectrograms is constructed. Outputs from those two modules are constrained to be subject to the same distribution, hence reducing the gap between the generated latent variables and the distribution of the real features. Then a re-parameterization module followed by a GAN-based decoder is leveraged to convert hidden features to the target waveform. This waveform is then post-processed to become the final generated speech. For the ASR, we utilize an encoder a hybrid CTC/Attention model[37] to extract the features of phonetic posteriorgrams(PPGs) from given speech.

We use the non-causal WaveNet[38] residual blocks as the posterior encoder. Such residual block consists of several dilated convolutions with skip connection and gated activation. For multi-speaker VC, we additionally feed speaker embedding into residual blocks through the method of global conditioning.

The conversion decoder is almost essentially the HiFi-GAN[39] generator. The generator is a stack of convolution blocks, which include transpose convolution layers and a multi-receptive field fusion module (MRF). The MRF is composed of a series of residual blocks that have receptive fields of different sizes. To avoid possible checkerboard artifacts [40] caused by the transpose convolution process, we rebuild the upsampling layer using temporal nearest interpolation followed by a 1D convolution. As with GAN-based vocoders, we also add a discriminator D that attempts to distinguish audio generated by the generator G from the ground truth. Similar to the design in Hifi-GAN[39] and MelGAN[41], the discriminators include a multi-period discriminator(MPD) and a multi-scale discriminator (MSD). MPD is a mixture of window-based sub-discriminators, where different pe-

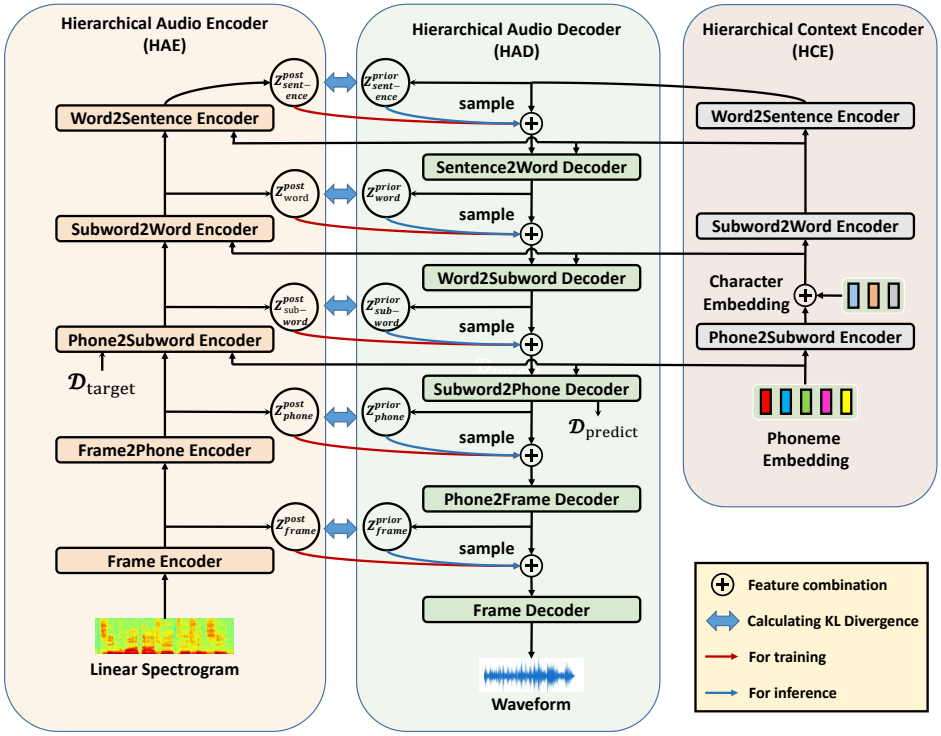

Figure 2: The architecture of Hier-TTS.

riodic patterns of waveform are operated. MSD directly functions on different scales, which consecutively evaluates audio samples at different levels that help to capture consecutive patterns and long-term dependencies.

3.3. Adversarial attacking

Projected Gradient Descent (PGD) [34] is considered to be the strongest attack method based on gradient information. When the model is robust to PGD attacks, it is robust to most gradient-based attack methods. The PGD attack uses the gradient of the model to the input to find the disturbance that maximizes the loss value. A typical PGD attack on network θ is mathematically expressed as the following recursive formula:

$$x^{t+1} = \prod_{x+S}(x^t + \alpha sgn(\nabla_x L(\theta, x, y))) \quad (1)$$

Where x represents the original sample, y denotes sample label/class, t is the number of iterations, S is the maximum perturbation range, α is the perturbation size for each iteration, and L represents the loss function of the target model. The PGD attack generates the adversarial perturbation that maximizes the sample loss value through multiple iterations. Adversarial training can be understood as solving an optimization problem with an external minimum and an internal maximum. It solves decision boundaries that are robust to maximum adversarial perturbations. For speech, adversarial attacks can target both the original audio and the acoustic features used by the model, which need to be considered at the same time during adversarial training. The optimization function for adversarial training is shown below:

$$argmin_\theta, where \rho(\theta) = E_{(x,y)}[max_{\delta \in S} L(\theta, x+\delta, y)] \quad (2)$$

Among them, $\rho(\theta)$ is the external minimum optimization function, which optimizes the network parameters for the most adversarial example of a clean sample, so that the network can be classified correctly.

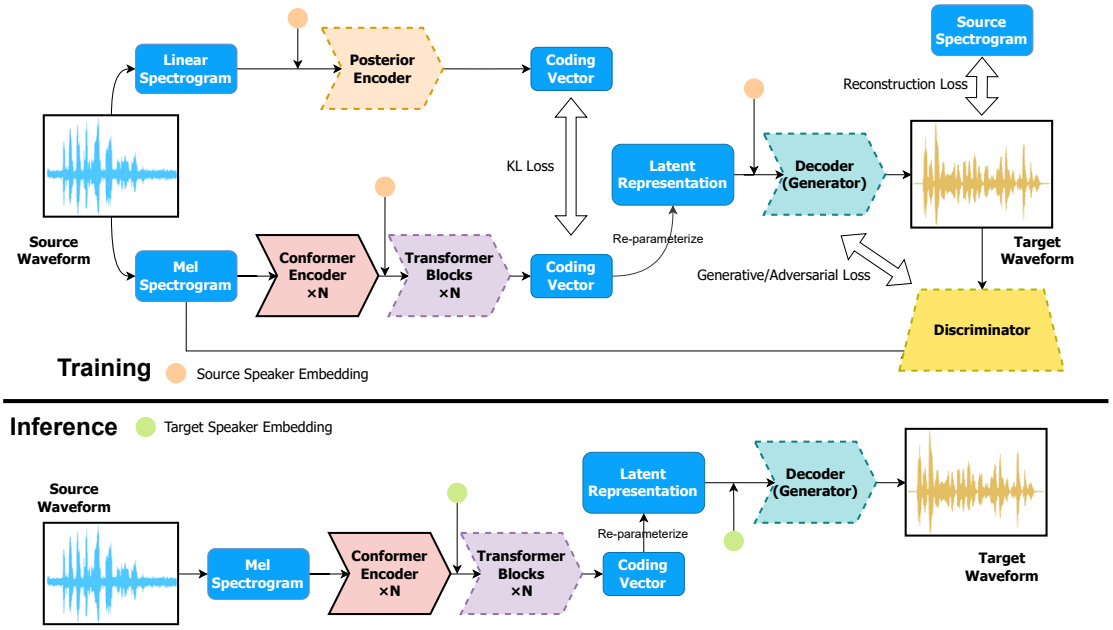

Figure 3: The architecture of E2EVC

3.4. Post-processing

Similar to what is proposed in our previous work[29], we introduce some postprocessing countermeasures against silence detection and check-out employed by anti-spoofing. We experiment with two methods of appending silence: In the first approach, we leverage a non-neural voice-activity-detecting(VAD) system that quickly finds silent segments in fake speech and gets the timestamps. Then we randomly select real silent segments from the recordings of the target speaker. We crop those real segments to the same length as calculated from the timestamps given by VAD and then simply replace them. The other is global superposition without relying on VAD. We also randomly select multiple real silent segments and normalize their amplitudes to an average level. After that, we connect them end to end through the algorithm of parabolic cross-fading till the total length surpasses the length of generated audio. Finally, we directly paste this stitched silent sample onto our synthetic speech like additive noise. The parabolic cross-fading algorithm is formulated as follows:

$$Mask_1(x,k) = \begin{cases} 1 & \\ & x < K - T, \\ \frac{1}{T^2}(K-2T-x)(x-K) & \\ & K-T \le x \le K \end{cases} \tag{3}$$

$$Mask_2(x) = \begin{cases} 1 & x > T \\ \frac{1}{T^2}(2T-x)x & 0 \le x \le T \end{cases} \tag{4}$$

for k from 0 to K_{left}:

$$s_{left}(k) \leftarrow Mask_1(k, K_{left}) s_{left}(k) \tag{5}$$

for k from 0 to K_{right}:

$$s_{right}(k) \leftarrow Mask_2(k) s_{right}(k) \tag{6}$$

for x from 0 to K_{left}+K_{right}-T:
Connect(x,s_{left},s_{right})=

$$\begin{cases} s_{left}(x) & \\ & 0 \le x < K_{left} - T, \\ s_{left}(x) + s_{right}(x + T - K_{left}) & \\ & K_{left} - T \le x < K_{left}, \\ s_{right}(x + T - K_{left}) & \\ & K_{left} \le x \le K_{left} + K_{right} - T \end{cases}$$

where s_{left} and s_{right} denote the raw silent segments to be concatenated. K_{left} and K_{right} represent the lengths of these two segments. K and x in the formulas are the subscripts of the time dimension iteration. T stands for the overlapping duration. The final result is calculated by Connect(·).

Further more, we assume that a decline in speech quality may cause the performance of some anti-spoofing systems to degrade correspondingly and thus increase the probability of a successful attack. We reduce the

Table 1
EER ablation experiments of various system components, AA± for with/without adversarial attacking, PP± for post-processing.

Settings	STFT-LCNN	FFT-SENET	SILENCE	Average
AA- PP-	20.1%	50.9%	36.4%	35.8%
AA+ PP-	23.6%	52.4%	35.0%	37.0%
AA- PP+	44.2%	72.5%	**39.4%**	52.0%
AA+ PP+	**45.8%**	**74.9%**	39.1%	**53.2%**

Table 2
Ablation experiments w/o DPM backend. Results are the average EER returning from four main-stream anti-spoofing systems.

Settings	No Diffusion Vocoder	WaveGrad	DiffWave
AA- PP-	35.8%	**36.5%**	29.7%
AA+ PP+	**53.2%**	49.4%	41.0%

speech quality of speech by adding white Gaussian noise and using the signal-to-noise ratio (SNR) to characterize our speech quality. We notice that when the SNR is too low, the intelligibility of speech drops so rapidly that it becomes nearly unintelligible and cannot pass ASR, so there is probably a trade-off and we manage to maintain the SNR at a relatively high level after perturbation.

4. Experiments and results

We tested the spoofing ability of our synthesized audio on several different mainstream anti-spoofing models along with various training features. For each anti-spoofing model, we sent 1000 pieces of the generated waveform to perform our test, and the equal error rate (EER) was considered to be the performance indicator. Results are shown below in Table 1. Judging from the results, AA and PP methods have achieved some effects in the face of a bunch of anti-spoofing systems.

In the original design that was abandoned, we also tried to incorporate a vocoder of Diffusion Probabilistic Models (DPM) after the VC module, because in diffusion models, a segment of generated signal is iterated from white noise. Unlike most vocoders that use a deconvolution-like upsampling method, diffusion models will be less likely to produce artifacts in the frequency domain. However, the actual effect is not obvious or may reduce performance, according to Table 2, which demonstrates the results of this part of the ablation experiment.

Our team's system achieved the 3rd place in the overall score in this spoofing track. It is worth noting that our synthesized speech had the highest DSR% against all black-box anti-spoofing systems in two rounds of testing. Table 3 gives the specific performance data of all teams. Our team is A03.

We believe that the reasons why our multi-stage system is more powerful against black boxes are mainly anchored in:

1. A complete end-to-end training method has been adopted–the boundary between the acoustic model and the vocoder is ambiguous so the part that acts as a vocoder in the synthesis pipeline is somewhat different from a traditional autoregressive/GAN/DPM model. This may have led to anti-spoofing models submitted by many teams that have not seen the patterns of our generated speech.

2. We have considered the existence of some long-term artifacts based on human hearing perception, and the HierTTS and E2EVC models adopted have been designed to suppress these artifacts to a certain extent.

3. AA and PP methods have increased the complexity of generating speech, masking some artifacts and causing possible bias in the feature extraction of the anti-spoofing model. Methods that use non-speech parts as the main basis for detection are also targeted.

For the given AsvSpoof baseline released by the challenge organizers, the adversarial attack method we adopt has achieved some effect. $\sim$1e-5 and 0.99 represent the average detection score before and after adding the adversarial attack respectively. We are informed that the judging threshold is 0.1, so the performance against baseline is supposed to be DSR 100%. However, according to the Challenge results, our spoofing performance against the white-box baseline remains only DSR 23%. The results are inconsistent. Our follow-up tests suggest that the PGD noise used in the AA method may cause some negative effects after being enhanced by PP approaches, but we are not confident enough.

We understand that in any case, there is still room for improvement in the attack capabilities of white-box and black-box systems, such as further dismantling the GAN-like vocoder structure of the complete end-to-end VC part, or solving the problem that PGD may play a negative role.

Table 3
ADD 2023 Track 1.1 Rankings. Round 1 and Round 2 were black-box anti-spoofing systems while Baseline was a given white-box. The results of Round 1, Round 2, and Baseline were measured by DSR (%) and the final evaluation was performed with WDSR (%). Weights of the three parts were 0.4, 0.24, 0.36 respectively.

Ranking	Team ID	Round1 DSR%	Round2 DSR%	Baseline DSR%	WDSR%
1	A01	37.91	49.60	49.80	**44.97**
2	A02	37.80	27.81	**77.05**	43.63
3 (This paper)	A03	**43.20**	**51.58**	23.45	41.48
4	A04	33.16	36.25	51.30	38.63
5	A05	36.63	38.52	36.77	37.35
6	A06	38.14	36.66	9.32	30.69
7	A07	39.68	22.79	25.45	30.18
8	A08	34.83	29.03	5.51	25.71
9	A09	0.00	35.55	24.85	18.76
10	A10	30.71	17.28	0.10	18.53
11	A11	23.58	16.00	6.71	16.80
12	A12	41.72	0.00	0.00	16.69
13	A13	40.12	0.00	0.00	16.05
Avg.					27.11

5. Conclusions

This paper is a detailed description of a multi-stage audio synthesis system that participated in Track1.1 (spoofing) in ADD Challenge 2023. The system we build takes into account the ability to hide artifacts at different feature levels. Moreover, we use adversarial sample attacks and post-processing methods to further improve our spoofing capabilities. Experiments and challenge results show that our ability to attack black-box anti-spoofing systems is relatively good, but some modules may have a negative effect on the white-box baseline. In the follow-up work, we will continue to solve these problems to strengthen the offensive and defensive capabilities in the field of speech deepfake and anti-spoofing.

References

[1] 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. doi:10.21437/Interspeech.2017-1111.

[2] 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, 2019, pp. 1008–1012. doi:10.21437/Interspeech.2019-2249.

[3] 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, 2021, pp. 47–54. doi:10.21437/ASVSPOOF.2021-8.

[4] 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, 2022, pp. 9216–9220. doi:10.1109/ICASSP43922.2022.9746939.

[5] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren, L. Xu, J. Zhou, H. Gu, Z. Wen, S. Liang, Z. Lian, H. Li, Add2023: the second audio deepfake detection challenge, in: accepted by IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023), 2023.

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

[7] 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 (2020).

[8] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, R. A. Saurous, Tacotron: Towards end-to-end speech synthesis, 2017. arXiv:1703.10135.

[9] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, Y. Wu, Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, 2018. arXiv:1712.05884.

[10] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, T.-Y. Liu, Fastspeech: fast, robust and controllable text

to speech, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019, pp. 3171–3180.

[11] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, T.-Y. Liu, Fastspeech 2: Fast and high-quality end-to-end text to speech, 2022. arXiv:2006.04558.

[12] C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, J. Xiao, Flow-tts: A non-autoregressive network for text to speech based on flow, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7209–7213. doi:10.1109/ICASSP40776.2020.9054484.

[13] J. Kim, S. Kim, J. Kong, S. Yoon, Glow-tts: A generative flow for text-to-speech via monotonic alignment search, 2020. arXiv:2005.11129.

[14] 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.

[15] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, N. S. Kim, Diff-tts: A denoising diffusion model for text-to-speech, 2021. arXiv:2104.01409.

[16] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, Grad-tts: A diffusion probabilistic model for text-to-speech, 2021. arXiv:2105.06337.

[17] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks, in: 2018 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2018, pp. 266–273.

[18] T. Kaneko, H. Kameoka, Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks, in: 2018 26th European Signal Processing Conference (EUSIPCO), IEEE, 2018, pp. 2100–2104.

[19] K. Qian, Y. Zhang, S. Chang, X. Yang, M. Hasegawa-Johnson, Autovc: Zero-shot voice style transfer with only autoencoder loss, in: International Conference on Machine Learning, PMLR, 2019, pp. 5210–5219.

[20] D.-Y. Wu, Y.-H. Chen, H.-Y. Lee, Vqvc+: One-shot voice conversion by vector quantization and u-net architecture, 2020. arXiv:2006.04154.

[21] H. Zheng, W. Cai, T. Zhou, S. Zhang, M. Li, Text-independent voice conversion using deep neural network based phonetic level features, in: 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 2872–2877. doi:10.1109/ICPR.2016.7900072.

[22] L. Sun, K. Li, H. Wang, S. Kang, H. Meng, Phonetic posteriorgrams for many-to-one voice conversion without parallel data training, 2016, pp. 1–6. doi:10.1109/ICME.2016.7552917.

[23] 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.

[24] H. Tak, M. Todisco, X. Wang, J. weon Jung, J. Yamagishi, N. Evans, Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation, 2022. arXiv:2202.12233.

[25] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, A. Larcher, End-to-end anti-spoofing with rawnet2, 2021. arXiv:2011.01108.

[26] H. Tak, J. weon Jung, J. Patino, M. Todisco, N. Evans, Graph attention networks for anti-spoofing, 2021. arXiv:2104.03654.

[27] 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, DDAM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 61–68. URL: https://doi.org/10.1145/3552466.3556525. doi:10.1145/3552466.3556525.

[28] C. Sun, S. Jia, S. Hou, E. AlBadawy, S. Lyu, Exposing ai-synthesized human voices using neural vocoder artifacts, 2023. arXiv:2302.09198.

[29] H. Hua, Z. Chen, Y. Zhang, M. Li, P. Zhang, Improving spoofing capability for end-to-end any-to-many voice conversion, in: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia, Association for Computing Machinery, 2022.

[30] C. Wang, J. Yi, J. Tao, C. Zhang, S. Zhang, X. Chen, Detection of cross-dataset fake audio based on prosodic and pronunciation features, 2023. arXiv:2305.13700.

[31] E. Conti, D. Salvi, C. Borrelli, B. Hosler, P. Bestagini, F. Antonacci, A. Sarti, M. C. Stamm, S. Tubaro, Deepfake speech detection through emotion recognition: A semantic approach, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 8962–8966. doi:10.1109/ICASSP43922.2022.9747186.

[32] Z. Shang, P. Shi, P. Zhang, L. Wang, G. Zhao, Hiertts: Expressive end-to-end text-to-waveform using a multi-scale hierarchical variational auto-encoder, Applied Sciences 13.2 (2023): 868 (2023).

[33] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, 2015. arXiv:1412.6572.

[34] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, 2019. arXiv:1706.06083.

[35] N. M. Müller, F. Dieckmann, P. Czempin, R. Canals, K. Böttinger, J. Williams, Speech is silver, silence

is golden: What do asvspoof-trained models really learn?, arXiv preprint arXiv:2106.12914 (2021).

[36] Y. Zhang, W. Wang, P. Zhang, The effect of silence and dual-band fusion in anti-spoofing system, 2021, pp. 4279–4283. doi:10.21437/Interspeech.2021-1281.

[37] H. Miao, G. Cheng, C. Gao, P. Zhang, Y. Yan, Transformer-based online ctc/attention end-to-end speech recognition architecture, 2020. arXiv:2001.08290.

[38] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, Wavenet: A generative model for raw audio, arXiv preprint arXiv:1609.03499 (2016).

[39] 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).

[40] J. Pons, S. Pascual, G. Cengarle, J. Serrà, Upsampling artifacts in neural audio synthesis, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 3005–3009.

[41] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, A. C. Courville, Melgan: Generative adversarial networks for conditional waveform synthesis, Advances in neural information processing systems 32 (2019).