TIME DOMAIN ADVERSARIAL VOICE CONVERSION FOR ADD 2022

Cheng Wen, Tingwei Guo, Xingjun Tan, Rui Yan, Shuran Zhou,
Chuandong Xie, Wei Zou, Xiangang Li

Beike, Beijing, China
{wencheng008, zouwei026,lixiangang002}@ke.com

ABSTRACT

In this paper, we describe our speech generation system for the first Audio Deep Synthesis Detection Challenge (ADD 2022). Firstly, we build an any-to-many voice conversion (VC) system to convert source speech with arbitrary language content into target speaker's fake speech. Then the converted speech generated from VC is post-processed in time-domain to improve the deception ability. The experimental results show that our system has adversarial ability against anti-spoofing detectors with a little compromise in audio quality and speaker similarity. This system ranks top in Track 3.1 in the ADD 2022, showing that our method could also gain good generalization ability against different detectors.

Index Terms— ADD 2022, time-domain, deception ability, anti-spoofing, voice conversion

1. INTRODUCTION

Thanks to the superiority of deep learning and large-scale of high-quality open source speech corpus [1] [2], computational generated speech has reached humanlike naturalness and hi-fidelity audio quality. With a small batch of recording audio samples, state-of-art synthesis systems can generate non-distinguishable speech with high similarity of the target speaker. However, these technologies threaten the anti-spoofing and automatic speaker verification (ASV) systems greatly. In a recent study [3], synthetic speech is perceptually non-distinguishable from bona fide speech, and even well trained human detectors can get only 80% in accuracy.

According to a survey [4], audio deepfake methods can divide into three subcategories: replay attack, speech synthesis and voice conversion. Replay attacks are defined as replaying the recording of a target speaker's voice. Although the method is simple and efficient, it's application is constrained by recording environment and language content. Speech synthesis (SS), also know as text to speech (TTS), is a technique that convert written language into human speech. Neural network-based SS systems can generate deepfake audios with a significant improvement in both intelligibility and naturalness, especially those with the end-to-end architectures, such as [5], [6], [7], [8], [9]. The main benefit of SS is that it can generate speech with arbitrary language content. Another benefit of SS is that it can generate any speaker's voice with the development of adaptive TTS [10] technology, such as [11] and [12]. Voice conversion (VC) is a technique that converts a source speaker's voice to a target speaker's voice without changing linguistic information. The latest out-of-standing VC systems trend to utilize Variational Autoencoder (VAEs) and Generative Adversarial Network (GAN) frameworks to improve the target speaker similarity and audio quality. Such as Cycle-VAE [13], Disentangled-VAE [14], fang's CycleGAN-based nonparallel VC [15], STARGAN-VC [16] and StarGANv2-VC [17].

Aim to accelerate and foster research on detecting deep synthesis and manipulated audios, Audio Deep Synthesis Detection Challenge [18] is held as Signal Processing Grand Challenge on ICASSP 2022. The challenge contains four tracks, among which the Track 3 is an audio fake game (FG) which includes two sub tasks: Track 3.1 Generation task (FG-G) is a generation task aims to generate fake audios that can fool the fake detection model. Track 3.2 Detection task (FG-D) is a detection task tries to detect all the generated fake audios, including results from FG-G. Our team has participated in the FG-G task and won the top rank.

This paper describes our contributions about audio deepfake anti-spoofing, especially VC-based fake speech generation method to "attack" the neural network-based detection systems. The backbone of our system is FastSpeech-VC [19] followed with HiFi-GAN [9]. Our FastSpeech- VC is designed to converts bottleneck feature (BNF) into mel- spectrograms. And then generate audio signal from the mel- spectrograms by the HiFi-GAN. In order to fool the detection systems, we further add a post-processing modification on the generated audio, which cause a slight decay in audio quality but a significant promotion in spoofing. Audio samples are available at our demo page[1]

The rest of this paper is organized as follows. Section 2 introduces our proposed method. Section 3 describes our implementation details and experimental results. Finally, the conclusion is given in section 4.

[1]https://guo-t-w.github.io/KeAI-ADD-2022/

2. SYSTEM DESCRIPTION

The goal of the Track 3.1 FG-G is to generate fake audio that can fool the detection system, which requires the distribution overlap between the generated audio and target speaker's natural audio to be as high as possible.

Inspired by [20], we propose an adversarial approach to make the distribution of the VC generated converted speech, to be closer to that of the natural audio of target speaker. As shown in Figure 1, the converted speech is sent to a residual generation network (RGN) to enable a white-box attack on a pretrained detection model.

Different from [20], we build an automatic speaker verification system for targeted speakers (ASV-TS), which is used as the anti-spoofing system for the adversarial training of the residual generation network. Given the set of all speech data $D = \{D_T, D_O\}$, where D_T is the natural audio set of target speakers and D_O denotes the set of all other speech data, the goal of ASV-TS is to distinguish D_T from D_O as accurately as possible. However, the separating hyperplane between D_T and D_O determined by the ASV-TS may not perform well in complex scenarios, since only limited audios of target speakers in AIShell-3 could be used in Track 3.1 in the ADD 2022. In order to improve the performance of the ASV-TS system, various data augmentation techniques is adopted on the available speech data in D_O. In addition, the residual signal is generated based on the time-domain signal directly, instead of the spectral amplitude, to avoid loss of information.

Our entire system involves two stages, i.e., voice conversion and time-domain adversarial post-processing, which will be introduced in detail next.

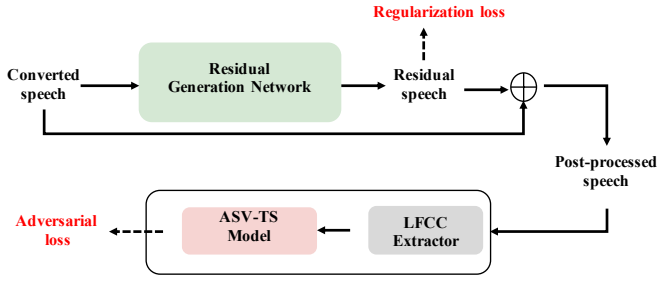

Fig. 1. Time-Domain Adversarial Post-Processing framework

2.1. Voice Conversion

As shown in Figure 2, a VC system is built to generate audio for specific timbres and content. The overall framework consists of three parts, BNF extractor, synthesizer network and vocoder.

Both phonetic posteriorgram (PPG) and BNF are widely used features in current VC tasks, which retain acoustic information while excluding speaker identities and are obtained from the ASR acoustic model. In our work, we use a DNN-based acoustic model from an ASR system trained with our own data, to perform as the BNF extractor, which consists of 7 TDNN layers and 3 unidirectional LSTM layers, followed by the 512-dim bottleneck layer.

Our synthesizer network is based on Fastspeech-VC framework [19], which originates from FastSpeech [7] TTS model. The synthesizer network can predict mel-spectrograms from BNF. Meanwhile, speaker ids are used to control the speaker identity of synthesized utterances. Therefore, our system can achieve any to many voice conversions without reliance on the training data of source speakers. In the synthesizer network, the encoder and the decoder are composed of a stack of N = 6 duplicate Feed-Forward Transformer (FFT) blocks.

Finally, HiFi-GAN vocoder is used to reconstruct audio from predicted mel-spectrograms.

2.2. Time-Domain Adversarial Post-Processing

2.2.1. *Target Speaker Verification Model*

Commonly used ResNet-34 [21] is adopted to build the ASV-TS model in our work. The loss function is binary cross-entropy. We used linear frequency cepstrum coefficients (LFCCs) with a window size of 25ms and an overlap of 10ms as input of detection model.

2.2.2. *Residual Generation Network*

The RGN in our system is a fully convolutional feed-forward network with waveform as input. It's architecture is similar to the generator of MelGAN [8]. Residual signal is generated by RGN and are then added to the input audio to obtain the post-processed speech.

2.2.3. *Adversarial Training*

Given $s \in R^T$ denote the samples of input signal, the residual signal generated by RGN can be defined as $P(s) \in R^T$. When training the RGN, the adversarial training is conducted with objectives as:

$$L_A = 1 - D_t(F(s + P(s))) \tag{1}$$

where F is the LFCC features extractor and $D_t(.)$ represents the probability of the target audio predicted by the ASV-TS model. In addition, a regularization loss, which is composed of three parts, is also designed to maintain the subjective quality of post-processed speech.

$$L_R = L_r + L_m + L_s \tag{2}$$

Among them, the L_r is the difference between the maximum and the minimum sample values in the generated residual waveform. The L_m is the mean square error (MSE) between the value of generated residual samples and a zero vector. By denoting the post-processing modification at each sample as

$$M_t = \frac{abs(P(s_t))}{abs(P(s_t)) + abs(s_t) + 0.0001} \tag{3}$$

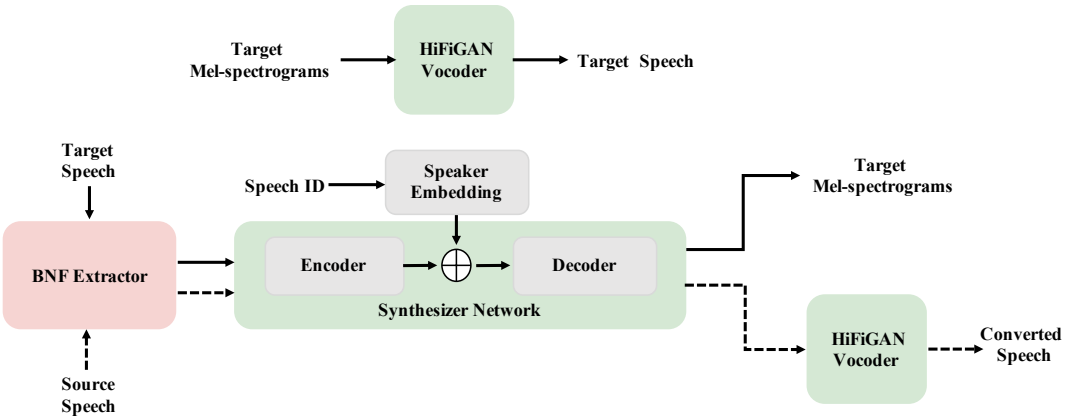

Fig. 2. Voice Conversion framework.The solid and dashed lines represent the data flow for training and inference respectively

where $P(s_t)$ and s_t are the generated residual and input values of t-th sample. The L_s is defined as

$$L_s = \frac{1}{T}\sum_{t=1}^{T} M_t \tag{4}$$

The L_s can further guarantee that the post-processing modifications are as slight as possible, especially in silence clips.

The final loss function of our post-processing model is

$$L = \lambda_A L_A + \lambda_R L_R \tag{5}$$

where λ_A and λ_R are the scale factors of adversarial and regularization loss respectively.

3. EXPERIMENTS

3.1. Datasets and Model Details

- **Voice Conversion**: For Track3.1 Generation task, participants are required to generate deepfake audio of 10 target speakers using the AIShell-3 dataset [1]. We first trained multi-speaker synthesizer network and universal vocoder based on the AIShell-3, and then fine-tuned them for all target speakers. After that, 5000 deepfake audios of 10 target speakers, denoted as VC-T5000, are generated. The deepfake waveform of each speaker is produced with 500 utterances recorded by mobile phone in quiet environment as source speech.
- **ASV-TS Model**: ASV-TS model is trained to distinguish the natural speech of target speakers from other audios. We take all utterances of 10 target speakers as positive samples. Meanwhile, the same amount of audios of other speakers randomly selected from AIShell-3 and deepfake audios generated from VC and TTS (built on AIShell-3) are used as negative audios. In addition, each of negative audio is augmented by a method randomly chosen from Table 1 to further extend the set of negative audios.
- **RGN**: For the training of RGN, all audios in VC-T5000 are used, and the ASV-TS model is employed to perform adversarial training. In the post-processing network, four upsampling layers with 8x, 8x, 2x, 2x factors respectively are used to achieve 256x upsampling. All input audios are reshaped with shape [T, 256], and the shape of output residual signal is [T*256, 1]. The values of learning rate are [0.0001, 0.00005, 0.000025, 0.0000125, 0.00000625], where decay boundaries are [5000, 10000, 30000, 50000] steps. In order to guarantee the post-processing modifications are as slight as possible, we set $\lambda_R = 20$ and $\lambda_A = 1$.

Table 1. List of Data Augmentation Methods

Approach	Methods	Description
Distortion	noise music babble	MUSAN and self-collected
	reverb	room impulse response
	volume	-10dB to 20dB
Compression	MP3 OGG AAC OPUS	Random compression ratio
	sample rate	16kHz -> 8kHz

3.2. Evaluation

The deception success rate (DSR) is chosen as the metric for generation task. DSR is defined as followed:

$$DSR = \frac{W}{A * N} \tag{6}$$

where W is the count of wrong detection samples by all the detection models on the condition of reaching each own EER performance, A is the count of all the evaluation samples, and N is the number of detection models.

The results are shown in Figure 3. Our team id is C10. It can be seen that our final result ranks first, which also shows the effectiveness of our proposed model.

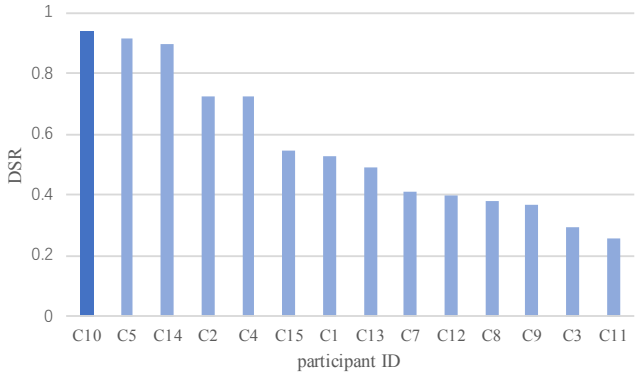

Fig. 3. The final DSR score of our submited audios

3.3. Analysis

We analysed the character error rate (CER) and the cosine similarity (COS_SIM) of speaker embedding vectors of the speech before and after post-processing. As shown in Table 2, the CER and cosine similarity have no significant change, showing that post-processing has no effect on audio quality.

In order to compare the spoofing performance of before and after post-processing, a new spoofing detection model is trained based on the data of ADD 2022 Detection Task. The model configurations are the same as ASV-TS model. We randomly select 14K utterances from the audio before and after post-processing, respectively, and combine them with 0.6K utterances from AIShell-3 as two test sets. From Table 3 we can find that, the EER changed from 24.7% to 74.0%, which means the deception ability of post-processed speech is significantly improved.

Furthermore, we analyzed the spectrum of speech. Figure 4 illustrates the spectrograms before and after post-processed by RGN model. It can be seen that the difference between the two spectrograms is very slight, especially in low frequency region. In addition, there is nearly no residual signal has been added to the silence clips in the input audio, which means RGN only perturbs where there is speech information.

4. CONCLUSION

This paper introduces our system for FG-G. It involves two stages, including voice conversion and time-domain adversarial post-processing. The voice conversion model is built using Fastspeech-VC framework and the post-processing is performed directly on the time-domain of converting audios by the RGN, which is implemented by adversarial training against the ASV-TS. The experimental results show that our system can generate audios with both high quality and adversarial ability against spoofing detectors. Our system has also achieved the highest performance, a DSR of 0.938, among all participants of the Track 3.1 of Audio Deep Synthesis Detection Challenge 2022.

Table 2. The CER and cosine similarity of speaker embedding vectors the speech before and after postprocessing

Speaker ID	Source	Before		After	
	CER(%)	CER(%)	COS_SIM	CER(%)	COS_SIM
SSB0139	_	14.39	0.93	14.76	0.91
SSB0535	_	13.24	0.91	13.26	0.90
SSB0601	_	13.70	0.90	14.46	0.88
SSB0603	_	13.63	0.90	13.86	0.87
SSB0607	_	13.27	0.89	14.26	0.85
SSB0609	_	13.96	0.89	14.75	0.86
SSB0629	_	13.83	0.94	14.04	0.91
SSB0666	_	13.76	0.91	14.68	0.89
SSB0668	_	13.13	0.93	13.49	0.91
SSB0671	_	13.54	0.87	14.02	0.84
Average	10.41	13.65	0.91	14.16	0.88

Table 3. The EER of the speech before and after postprocessing

	Before	After
EER	24.7%	74.0%

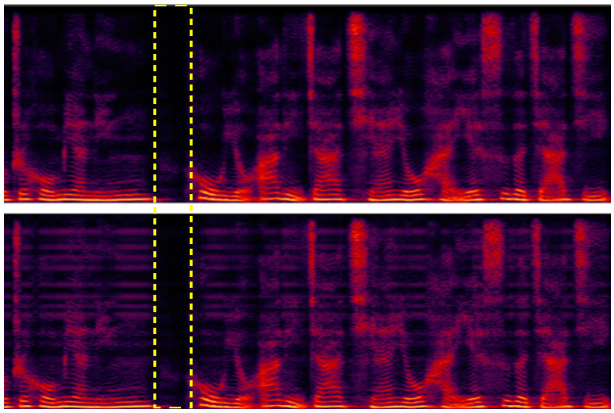

Fig. 4. The spectrograms of the before (top) and after audio (bottom) post-processing

5. REFERENCES

[1] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li, "Aishell-3: A multi-speaker mandarin tts corpus," *Proc. Interspeech 2021*, pp. 2756–2760, 2021.

[2] Tingwei Guo, Cheng Wen, Dongwei Jiang, Ne Luo, Ruixiong Zhang, Shuaijiang Zhao, Wubo Li, Cheng Gong, Wei Zou, Kun Han, et al., "Didispeech: A large scale mandarin speech corpus," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6968–6972.

[3] Nicolas M Müller, Karla Markert, and Konstantin Böttinger, "Human perception of audio deepfakes," *arXiv preprint arXiv:2107.09667*, 2021.

[4] Zahra Khanjani, Gabrielle Watson, and Vandana P Janeja, "How deep are the fakes? focusing on audio deepfake: A survey," *arXiv preprint arXiv:2111.14203*, 2021.

[5] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[6] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.

[7] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech: Fast, robust and controllable text to speech," *arXiv preprint arXiv:1905.09263*, 2019.

[8] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *arXiv preprint arXiv:1910.06711*, 2019.

[9] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *arXiv preprint arXiv:2010.05646*, 2020.

[10] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[11] Sercan O Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou, "Neural voice cloning with a few samples," *arXiv preprint arXiv:1802.06006*, 2018.

[12] Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, Sheng Zhao, and Tie-Yan Liu, "Adaspeech: Adaptive text to speech for custom voice," *arXiv preprint arXiv:2103.00993*, 2021.

[13] Patrick Lumban Tobing, Yi-Chiao Wu, Tomoki Hayashi, Kazuhiro Kobayashi, and Tomoki Toda, "Non-parallel voice conversion with cyclic variational autoencoder," *Proc. Interspeech 2019*, 2019.

[14] Manh Luong and Viet Anh Tran, "Many-to-many voice conversion based feature disentanglement using variational autoencoder," *arXiv preprint arXiv:2107.06642*, 2021.

[15] Fuming Fang, Junichi Yamagishi, Isao Echizen, and Jaime Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5279–5283.

[16] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu 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.

[17] Yinghao Aaron Li, Ali Zare, and Nima Mesgarani, "Starganv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion," *arXiv preprint arXiv:2107.10394*, 2021.

[18] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Shan Liang, Shiming Wang, Shuai Zhang, Xinrui Yan, Le Xu, and Haizhou Li, "Add 2022: the first audio deep synthesis detection challenge," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

[19] Shengkui Zhao, Hao Wang, Trung Hieu Nguyen, and Bin Ma, "Towards natural and controllable cross-lingual voice conversion based on neural tts model and phonetic posteriorgram," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5969–5973.

[20] Yi-Yang Ding, Li-Juan Liu, Yu Hu, and Zhen-Hua Ling, "Adversarial voice conversion against neural spoofing detectors," in *Proc. INTERSPEECH 2021*, 2021, pp. 816–820.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.