ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094845

SSI-Net: A MULTI-STAGE SPEECH SIGNAL IMPROVEMENT SYSTEM FOR ICASSP 2023 SSI CHALLENGE

Weixin Zhu[1]*, Zilin Wang*[2]*, Jiuxin Lin*[2]*, Chang Zeng*[3]*, Tao Yu*[1]

[1]Tencent, Shenzhen, China
[2]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[3]SOKENDAI, Kanagawa, Japan

wesleyzhu@tencent.com, {wangzl21, linjx21}@mails.tsinghua.edu.cn, zengchang@nii.ac.jp

ABSTRACT

The ICASSP 2023 Speech Signal Improvement (SSI) Challenge concentrates on improving the speech signal quality of real-time communication (RTC) systems. In this paper, we introduce the speech signal improvement network (SSI-Net) submitted to the ICASSP 2023 SSI Challenge, which satisfies the real-time condition. The proposed SSI-Net has a multi-stage architecture. We present the time-domain restoration generative adversarial network (TRGAN) in the first restoration stage for speech restoration. Regarding the second enhancement stage, we employ a lightweight multi-scale temporal frequency convolutional network with axial self-attention (MTFAA-Net) called MTFAA-Lite to enhance the fullband speech. In the subjective test on the SSI Challenge blind test set, our proposed SSI-Net yields a P.835 overall mean opinion score (MOS) of 3.190 and a P.804 overall MOS of 3.178, which eventually takes the 3rd place in tracks 1&2.

***Index Terms*—** speech signal improvement, multi-stage, SSI-Net

1. INTRODUCTION

Recently, RTC systems are gaining considerable popularity for widespread remote communication and collaboration. Although high-quality speech signal is essential to RTC systems, current RTC systems still suffer from issues that severely limit their speech quality, including environmental noise/reverberation, packet loss, bandwidth limitations, and attenuation, just to name a few. The ICASSP 2023 SSI Challenge [1] concentrates on resolving the problems of noise, coloration, discontinuity, loudness, and reverberation in speech, in order to improve the speech signal quality of RTC systems.

In this paper, with respect to the complicated tasks mentioned above, we propose the multi-stage framework SSI-Net. The oversuppression of the corrupted speech signal caused by the enhancement method might render the speech signal to be non-restorable. With the aim of avoiding this issue, we present the TRGAN in the first restoration stage for speech restoration and preliminary denoising/dereverberation. Besides, the output of the restoration stage may still contain residual noise and artifacts. As a consequence, to further boost the quality of the speech signal, the fullband speech enhancement model MTFAA-Lite is applied in the second enhancement stage to remove these residual noises and artifacts. Eventually, our submitted real-time system SSI-Net takes third place in tracks 1&2.

2. METHODOLOGY

Fig.1 shows that our multi-stage framework SSI-Net consists of a restoration stage and an enhancement stage. The TRGAN is responsible for speech restoration and preliminary denoising/reverberation during the restoration stage. After the processing of the raw input waveform in this stage, we will first obtain a relatively high-quality waveform. The restored waveform is transformed by the short-time Fourier transform (STFT) to the complex spectrogram which is then fed to the MTFAA-Lite. Afterward, the enhancement stage, which is mainly composed of MTFAA-Lite, will remove noise and artifacts to further improve speech quality. Eventually, the output of the MTFAA-Lite passes through the inverse STFT (iSTFT) to yield the final prediction. In the following, these parts are described in detail.

2.1. TRGAN

Previous work [2] targets the improvement of the speech signal with mel-domain generative model. However, the mel-domain model neglects the exploitation of phase information, which limits the upper bound of its performance. As another common paradigm for speech generative models, the time-domain model directly uses waveforms as input, implicitly taking the phase information into account, and has achieved excellent results in some fields [3, 4] of speech restoration. Accordingly, we propose TRGAN to perform speech signal restoration in time-domain.

The generator of the TRGAN adopts an encoder-decoder architecture. The encoder, which is made up of 1D convolution layers and residual-convolution layers [4] with a residual structure, is responsible for downsampling the speech waveform. Correspondingly, the decoder upsamples the features of the encoder output through residual-convolution layers and 1D transposed convolution layers. We utilize a pseudo quadrature mirror filter bank (PQMF) [5] to perform subband decomposition on the input waveform of the generator and signal reconstruction on its output, thus reducing the number of parameters and the computational effort. As for the discriminators, we present multi-band discriminators and integrate them with previously proposed multi-resolution frequency-domain discriminators [6] to perform well in generating different frequency components.

2.2. MTFAA-Lite

With the rapid development in recent years, speech enhancement methods are gradually extended from wideband speech enhancement to fullband speech enhancement. MTFAA-Net [7], the current state-of-the-art fullband speech enhancement approach with multi-scale

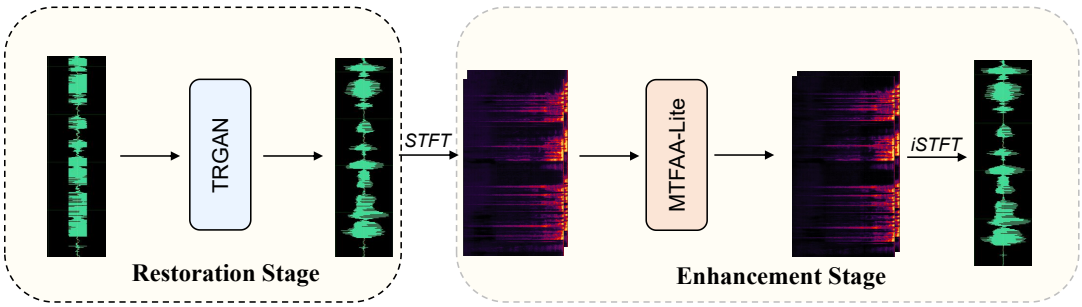

Fig. 1. The overall diagram of the multi-stage framework SSI-Net.

time-frequency processing and streaming axial attention, achieves impressive results in Deep Noise Suppression (DNS) Challenge - ICASSP 2022 [8]. To balance performance and computational complexity, we simplify the MTFAA-Net to obtain MTFAA-Lite which is applied in the enhancement stage. Specifically, we retain the frequency downsampling, frequency upsampling, and T-F convolution modules in MTFAA-Net, while dropping the T-attention with high time-complexity in axial self-attention.

3. EXPERIMENTS

3.1. Training Setup

We selected partial 48kHz audios from the DNS Challenge dataset [8] as the clean set and noise set. There were 100,000 room impulse responses (RIRs) generated by us based on the image method with RT60 [9]. After analyzing the audio from the SSI Challenge devset, we produced audio with issues such as coloration, discontinuity, loudness, noise, and reverberation in a statistical ratio for a total of 1500 hours. The training set and the validation set were divided from this 1500-hour dataset.

We ultimately worked with the SSI-Net having a total parameter number of 5.23 M. Its real-time factor (RTF) is 0.36 on an Intel Core i5 quad-core CPU clocked at 2.4 GHz.

Table 1. Subjective evaluation results based on ITU-T P.835.

Methods	ITU-T P.835 MOS		
	Overall	Signal	Background
Noisy	2.824	3.147	3.453
SSI-Net	3.190	3.471	4.073

Table 2. Subjective evaluation results based on P.804.

Methods	P.804 MOS			
	Coloration	Discontinuity	Loudness	Reverberation
Noisy	3.029	4.061	2.992	3.852
SSI-Net	3.550	4.140	4.060	4.322

3.2. Results and Analysis

Table 1 shows the results of the subjective test based on ITU-T P.835 on the SSI Challenge blind test set. It can observe that our SSI-Net not only effectively suppresses the noise but also improves the speech quality simultaneously.

We further explore the impact of our approach on several specific aspects of speech quality in Table 2. It can be seen that SSI-Net indeed efficiently tackles the problems which affect speech quality, including coloration, discontinuity, loudness, and reverberation.

4. CONCLUSIONS

This paper introduces our entry to ICASSP 2023 SSI Challenge. The submitted real-time system SSI-Net has a multi-stage architecture in which speech restoration is performed first, followed by speech enhancement. Our proposed SSI-Net finally achieved 3rd rank with impressive subjective test results based on P.835 and P.804.

5. REFERENCES

[1] Ross Cutler, Ando Saabas, Babak Naderi, Nicolae-Cătălin Ristea, Sebastian Braun, and Solomiya Branets, "ICASSP 2023 Speech Signal Improvement Challenge," *arXiv preprint arXiv:2303.06566*, 2023.

[2] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.

[3] Nan Li, Xiguang Zheng, Chen Zhang, Liang Guo, and Bing Yu, "End-to-End Multi-Loss Training for Low Delay Packet Loss Concealment," *Proc. Interspeech 2022*, pp. 585–589, 2022.

[4] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[5] TQ Nguyen, "Near-perfect-reconstruction pseudo-qmf banks," *IEEE Transactions on Signal Processing*, vol. 42, no. 1, pp. 65–76, 1994.

[6] Zhengxi Liu and Yanmin Qian, "Basis-MelGAN: Efficient neural vocoder based on audio decomposition," *arXiv preprint arXiv:2106.13419*, 2021.

[7] Guochang Zhang, Chunliang Wang, Libiao Yu, and Jianqiang Wei, "Multi-Scale Temporal Frequency Convolutional Network with Axial Attention for Multi-Channel Speech Enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9206–9210.

[8] Harishchandra Dubey, Vishak Gopal, Ross Cutler, Ashkan Aazami, Sergiy Matusevych, Sebastian Braun, Sefik Emre Eskimez, Manthan Thakker, Takuya Yoshioka, Hannes Gamper, et al., "ICASSP 2022 Deep Noise Suppression Challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9271–9275.

[9] Jont B Allen and David A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.