GESPER: A UNIFIED FRAMEWORK FOR GENERAL SPEECH RESTORATION

Jun Chen[1,2,*,†], *Yupeng Shi*[1,*], *Wenzhe Liu*[1,*], *Wei Rao*[1], *Shulin He*[1], *Andong Li*[3], *Yannan Wang*[1], *Zhiyong Wu*[2], *Shidong Shang*[1], *Chengshi Zheng*[3]

[1]Tencent Ethereal Audio Lab, Tencent, Shenzhen, China
[2]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[3]Institute of Acoustics, Chinese Academy of Sciences, Beijing, China

y-chen21@mails.tsinghua.edu.cn, {yupengshi, wenzheliu}@tencent.com, zywu@sz.tsinghua.edu.cn, cszheng@mail.ioa.ac.cn

ABSTRACT

This paper describes the legends-tencent team's real-time **Ge**neral **Spe**ech **R**estoration (Gesper) system submitted to the ICASSP 2023 Speech Signal Improvement (SSI) Challenge. This newly proposed system is a two-stage architecture, in which the speech restoration is performed, and then followed by speech enhancement. We propose a complex spectral mapping-based generative adversarial network (CSM-GAN) as the speech restoration module for the first time. For noise suppression and dereverberation, the enhancement module is presented with fullband-wideband parallel processing. On the blind test set of ICASSP 2023 SSI Challenge, the proposed Gesper system, which satisfies the real-time condition, achieves 3.27 P.804 overall mean opinion score (MOS) and 3.35 P.835 overall MOS, ranked 1st in both track 1 and track 2.

***Index Terms*—** speech signal improvement, two-stage, speech restoration, speech enhancement

1. INTRODUCTION

Real-time communication (RTC) systems such as teleconferencing systems, smartphones and telephones, have become a necessity in the life and work of individuals. However, due to the influence of acoustical capturing, noise/reveberation corruption and network congestion, the speech quality of current RTC systems is still deficient. The ICASSP 2023 SSI Challenge[1] focuses on improving the speech signal quality in RTC systems, which involves tackling the difficulties of noise, coloration, discontinuity, loudness, and reverberation of speech in a variety of complex acoustic conditions.

In this paper, we propose a unified general speech restoration two-stage framework namely Gesper which perfoms *"restoration and enhancement"* to address the complicated problems in the SSI Challenge. After considering that the excessive suppression of the degraded speech signal caused by the noise reduction methods may significantly increase the difficulty in restoring the desired speech signal, we first employ CSM-GAN as the restoration module for speech distortion restoration, narrowband bandwidth expansion (BWE) as well as preliminary denoising and dereverberation. Moreover, since there may still exist residual noise components and artifacts in the output of the restoration module, to further improve the quality of the speech signal, the enhancement module with fullband-wideband parallel processing [1] is applied in the second stage.

† Work conducted when the first author was intern at Tencent.

* These authors contributed equally to this work as first authors.

[1]https://www.microsoft.com/en-us/research/academic-program/speech-signal-improvement-challenge-icassp-2023/

2. METHODOLOGY

As shown in Fig.1, our proposed Gesper system is composed of the restoration module and the enhancement module. After applying a real-time sound level adjustment on the input time-domain audio waveform, the short-time Fourier transform (STFT) is performed on the modified audio waveform to obtain the complex spectrum. The real and imaginary parts of the complex spectrum are then fed into a two-stage architecture: 1) the restoration module first performs speech distortion restoration, and preliminary denoising and dereverberation with a generative adversarial network; 2) the enhancement module further eliminates residual noise components and artifacts based on a relatively high-quality speech complex spectrum generated by the restoration module. Each module will be described in detail in the following two parts.

2.1. Restoration Module

In speech-related fields, there are many previously available generative models in time-domain [2], mel-domain [3], and so on. Nevertheless, the poor high-frequency representation of the time-domain generative model and the inadequate utilization of phase information by the mel-domain generative model render them both inappropriate for the complex scenario of this challenge. Therefore, for the first time, we propose CSM-GAN as our restoration module by leveraging recent advances in speech enhancement and speech synthesis.

The generator of CSM-GAN is a complex spectral mapping-based UNet. Its encoder contains 2 convolution-dense layers and 3 convolution layers, while the decoder comprises the corresponding transposed convolution layers and transposed convolution-dense layers. Between the encoder and decoder, there are stacked temporal convolutional network blocks for temporal modeling. To reduce the number of parameters and computational effort, the fullband complex spectrum is divided into 3 subbands, and we then concatenate them in the channel dimension and hand them over to the generator finally. Regarding the discriminator, multi-resolution frequency discriminators [4] and our proposed multi-band discriminators are adopted together to overcome the problem of large dynamic range in different subbands.

2.2. Enhancement Module

For maintaining the performance and reducing the computational efforts, we conduct the fullband-wideband parallel processing in the enhancement module. More specifically, we divide the fullband complex spectrum into two groups of features: the complex spectrum of the wideband speech and 32 equivalent rectangular bandwidth (ERB) bands containing fullband information through band-

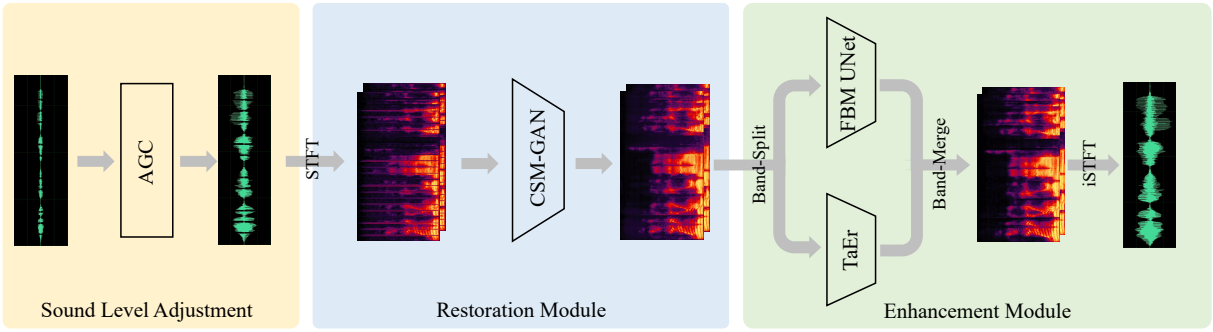

Fig. 1. The general schematic of the proposed Gesper. The "AGC" denotes auto gain control.

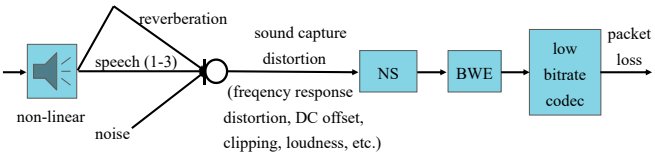

Fig. 2. The pipeline of data simulation, where "NS" refers to noise suppression and "DC" indicates direct current.

split. Subsequently, the wideband TaylorEnhancer [5] (TaEr) and fullband masking-based UNet (FBM UNet) [6] are introduced to process the wideband complex spectrum and ERB bands in parallel, respectively. TaEr has superior wideband noise suppression capability and focuses on wideband speech enhancement, while the FBM UNet provides the advantage of low complexity for fullband processing. The outputs of the two sub-networks are then integrated into the enhanced fullband complex spectrum by the band-merge operation.

3. EXPERIMENTS

3.1. Experimental Setup

We selected subsets from the DNS Challenge dataset [7] and our private dataset with different sampling rates as our clean set and the noise set. The room impulse responses (RIRs) were generated based on the image method. We analyzed the accumulated problematic audio from devset and Tencent meeting, and simulated a 1500-hour dataset according to specific cases of issues such as coloration, discontinuity, loudness, noise and reverberation, as shown in Fig.2.

We applied the Hanning window with a 20 ms window length and a 10 ms frame shift. The Gesper has a total parameter number of 12.1 M, and its real time factor (RTF) on an Intel Core i5 Quadcore CPU (clocked at 2.4 GHz) with single thread is 0.37.

Table 1. Multi-dimensional subjective test on the SSI Challenge blind test set*. "Color", "Disc", "Loud" and "Reverb" respectively denote Coloration, Discontinuity, Loudness and Reverberation.

Methods	P.804 MOS					P.835 MOS
	Overall	Color	Disc	Loud	Reverb	Overall
Noisy	2.360	3.029	4.061	2.992	3.852	2.824
Gesper	3.271	3.598	4.201	4.109	4.316	3.350

∗ More detailed results are available on this website.

3.2. Evaluation on the SSI Challenge Blind Test Set

Table 1 shows partial results of multi-dimensional subjective test on the SSI Challenge blind test set. One can see that the Gesper yields a significant improvement in all metrics relative to the noisy signals. This indicates that our proposed Gesper system efficiently alleviate the difficulties of noise, coloration, discontinuity, loudness and reverberation, which play a vital role in speech signal quality.

4. CONCLUSIONS

This paper introduces our submission to the ICASSP 2023 SSI Challenge. Our proposed two-stage framework Gesper achieves impressive results in addressing the challenges of noise, coloration, discontinuity, loudness and reverberation that reduce the speech quality. The proposed real-time system is ranked the first place in the tracks 1 and 2 of the ICASSP 2023 SII Challenge.

5. REFERENCES

[1] Shimin Zhang, Ziteng Wang, Jiayao Sun, Yihui Fu, Biao Tian, Qiang Fu, and Lei Xie, "Multi-task deep residual echo suppression with echo-aware loss," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9127–9131.

[2] Baiyun Liu, Qi Song, Mingxue Yang, Wuwen Yuan, and Tianbao Wang, "PLCNet: Real-time Packet Loss Concealment with Semi-supervised Generative Adversarial Network," *Proc. Interspeech 2022*, pp. 575–579, 2022.

[3] 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.

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

[5] Andong Li, Guochen Yu, Chengshi Zheng, Wenzhe Liu, and Xiaodong Li, "A General Deep Learning Speech Enhancement Framework Motivated by Taylor's Theorem," *arXiv preprint arXiv:2211.16764*, 2022.

[6] A Maier, AN Escalante-B, and T Rosenkranz, "DeepFilterNet2: Towards Real-Time Speech Enhancement on Embedded Devices for Full-Band Audio," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2022, pp. 1–5.

[7] 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.