2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW) | 979-8-3503-7451-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSPW62465.2024.10625840

GENERAL SPEECH RESTORATION USING TWO-STAGE GENERATIVE ADVERSARIAL NETWORKS

Qinwen Hu[1,2], *Tianyi Tan*[1,2], *Ming Tang*[3], *Yuxiang Hu*[2], *Changbao Zhu*[2], *Jing Lu*[1,2]

[1]Key Laboratory of Modern Acoustics, Nanjing University, Nanjing 210093, China
[2]NJU-Horizon Intelligent Audio Lab, Horizon Robotics, Beijing 100094, China
[3]State Grid Jiangsu Electric Power Co. Ltd, Nanjing 210024, China

{qinwen.hu, tianyi.tan}@smail.nju.edu.cn, tangming930702@163.com, {yuxiang.hu, changbao.zhu}@horizon.cc, lujing@nju.edu.cn

ABSTRACT

General speech restoration is a challenging task, which requires removing multiple types of distortions within a single system. The prevailing methods for general speech restoration largely rely on generative models, leveraging their ability to generate speech components based on prior knowledge of clean speech characteristics. Our approach adopts a two-stage processing scheme, comprising a speech restoration module and a speech enhancement module. The restoration module utilizes dilated convolutional networks and is trained using LSGAN losses. In contrast, the speech enhancement module employs a convolutional-recurrent network and is trained using metric-GAN losses. The proposed system achieves an overall opinion score (MOS) of 2.944 and a final score of 0.6805, ranking 3rd in Track 1 of the ICASSP 2024 Speech Signal Improvement Challenge (SIG-2).

Index Terms— speech signal improvement, multi-stage processing, generative adversarial networks

1. INTRODUCTION

Speech telecommunication systems contend with various scenarios that can introduce distortions into speech signals, such as background noises, reverberations, front-end distortions, codec distortions, etc. The ICASSP 2024 Speech Signal Improvement Challenge (SIG-2) [1] focuses on research to improve the speech signal quality across various distortions generally. The speech quality is measured by subjective evaluations in terms of colouration, discontinuity, loudness, and reverberation. Our entry in the challenge employs a 2-stage processing approach, utilizing two sub-modules consecutively to restore the distorted speech and enhance its quality by mitigating residual noises. Both modules are trained with adversarial losses to better exploit the latent key information within the distorted speech.

2. METHODOLOGY

Our approach comprises a restoration module and an enhancement module, as illustrated in Fig. 1. Both modules operate in the complex spectrogram domain, which encapsulates information with high resolutions and phase details. The complex spectral mapping-based generative adversarial network (CSM-GAN) [2] primarily recovers the distorted signal to an ideal state as captured by a high-fidelity microphone. Subsequently, the convolutional-recurrent metric generative adversarial network (CRMGAN)[3] further eliminates residual noises and reverberation present in the restored spectrogram. Table 1 shows the distortion types included in each training dataset. For Track 1, we downsample the full-band signal to a sampling rate of 24 kHz, process the super-wide band spectrogram with the proposed method, and upsample the recovered signal back to 48 kHz. For Track 2, we process the signal directly within the full-band range.

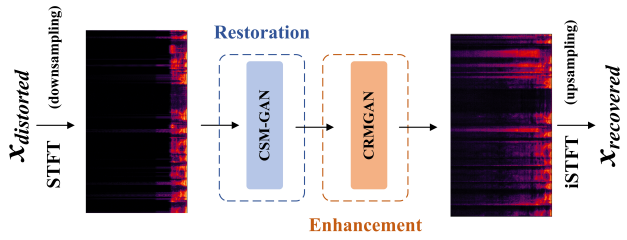

Fig. 1. The schematic diagram of the proposed system.

Table 1. The distortion types in each training set for the modules of the two stages.

	Environment	Equipment	Transmission and post-processing
CSM-GAN	far-filed, bone-conduct	equalization, front-end, clipping, rectifications, anti-AGC	resampling, codec, package-loss concealment
CRMGAN	noises, reverberations	-	-

2.1. Restoration module: CSM-GAN

The generator of CSM-GAN, serving as the restoration module, adopts a mapping-based convolutional UNet structure, similar to the generator in [2]. For Track 1, we partition the

Methods	MOS							WAcc	Score
	COL	DISC	LOUD	NOISE	REVERB	SIG	OVRL		
Noisy	3.341	3.695	3.778	3.214	3.399	3.049	2.580	0.8268	0.6509
2stageGAN-Track1	3.516	3.742	4.016	4.100	3.654	3.289	2.944	0.7949	0.6805
2stageGAN-Track2	3.484	3.737	3.984	3.973	3.635	3.263	2.907	0.7780	0.6707

Table 2. The subjective evaluations and ASR results on the SIG-2 blind set.

12-kHz-wide spectrum into 2 subbands instead of 3 as for the full-band spectrum, and concatenate them within the channel dimension.

We train the generator both regressively using spectral losses and adversarially using LSGAN losses [4]. We apply a set of discriminators, including multi-resolution discriminators [2], multi-band discriminators [2], and multi-period discriminators [5]. The multi-period discriminators can facilitate handling distortions characterized by structural features in the time domain, e.g., the rectifications.

2.2. Enhancement module: CRMGAN

The generator in the enhancement module CRMGAN [3] comprises an encoder, a time-frequency convolutional recurrent network, a mask decoder and a complex refinement decoder. The convolutional-recurrent module efficiently captures the sequential relationships along time- and frequency-axes. Training the generator involves an auxiliary adversarial loss utilizing a metric discriminator. We use perceptual evaluation of speech quality (PESQ) as the metric to guide the training.

3. EXPERIMENTS

3.1. Dataset and Training Setup

We construct our clean speech dataset with speeches from the DNS Challenge training dataset [6] and our private speech dataset, which includes English, Chinese, French, German, Italian and Russian. We simulate all distortions, except far-field and bone-conduct distortions which are real-recorded. The total training set size is approximately 500 hours.

We apply short-time Fourier transforms (STFT) to the signals with a window length of 20 ms and a hop length of 10 ms. The model sizes for Track 1 and Track 2 are 7.6 M and 9.9 M, respectively. Tested on an Intel Core i5 Quadcore CPU clocked at 2.4 GHz with a single thread, the corresponding real-time factors (RTF) are 0.47 and 0.62 for Track 1 and Track 2 respectively.

3.2. Results and Analysis

Table 2 presents the scores on the SIG-2 challenge blind set, including mean opinion scores (MOS) from subjective evaluations and the Word Accuracy Rates (WAcc) from automatic speech recognition (ASR) tests. Notably, the model in Track 1 outscores the model in Track 2 across the board, which is due to its ability to focus more on processing the low band and avoid producing unnatural high-band artefacts. Our submitted models in both tracks can improve the speech quality across all metrics, except for the WAcc.

4. CONCLUSIONS

In this paper, we introduce a two-stage GAN-based speech signal improvement model consisting of a complex spectral mapping-based LSGAN for signal restoration and a convolutional-recurrent MetricGAN for speech enhancement. Our method has demonstrated a competitive performance on the blind set of the SIG-2 challenge.

5. REFERENCES

[1] Nicolae-Catalin Ristea, Ando Saabas, Ross Cutler, Babak Naderi, Sebastian Braun, and Solomiya Branets, "ICASSP 2024 Speech Signal Improvement Challenge," in *ICASSP*, 2024.

[2] Wenzhe Liu, Yupeng Shi, Jun Chen, Wei Rao, Shulin He, Andong Li, Yannan Wang, and Zhiyong Wu, "GESPER: A Restoration-Enhancement Framework for General Speech Reconstruction," in *Proc. INTERSPEECH 2023*, 2023, pp. 4044–4048.

[3] Zhongshu Hou, Qinwen Hu, Tianchi Sun, Yuxiang Hu, Changbao Zhu, and Kai Chen, "Convolutional Recurrent MetriCGAN With Spectral Dimension Compression For Full-Band Speech Enhancement," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[4] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least Squares Generative Adversarial Networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[5] 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.

[6] Harishchandra Dubey, Ashkan Aazami, Vishak Gopal, Babak Naderi, Sebastian Braun, Ross Cutler, , et al., "ICASSP 2023 Deep Speech Enhancement Challenge," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.