Speech Super Resolution and Noise Suppression System Using a Two-stage Neural Network

Junkang Yang[1], Hongqing Liu[1,*], Xing Li[2], Jie Jia[2]

[1] School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China

[2] AI Lab, vivo Mobile Communication Co., Ltd, Nanjing, China

ABSTRACT & INTRODUCTION

Super-resolution (SR) and noise suppression have been a hot research topic in the field of speech processing. Speech super-resolution, also named as bandwidth extension (BWE), aims to extend the bandwidth of narrowband speech and improve its clarity, and noise suppression focuses on removing background noise from speech. Most of the past researches have performed these two tasks separately, while in the real world, bandwidth loss and noise exist almost simultaneously. Therefore, it is important to treat super-resolution and noise suppression jointly.

In recent years, deep learning has made a big splash in the field of speech processing, and we see the promise of using deep learning techniques to solve this problem. In this paper, we propose a joint speech super-resolution and noise suppression method based on deep learning methods. Specifically, we use two networks to handle these two single tasks separately and then cascade them, we use multiple loss functions as well as multiplexing of the spectrogram, and the experiments show that our method outperforms the baselines.

SYSTEM OVERVIEW

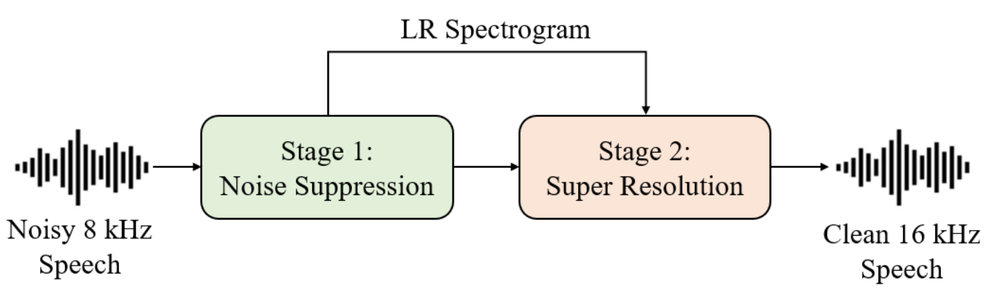

Figure 1. The architecture of our two-stage system.

As depicted in Figure 1, our system comprises two stages - the noise suppression stage and the super-resolution stage. The first stage handle the denoise task for 8 kHz noisy speech and the second stage focuses on the SR task for denoised speech. The "LR" means low sampling rate speech signal.

STAGE 1: NOISE SUPPRESSION

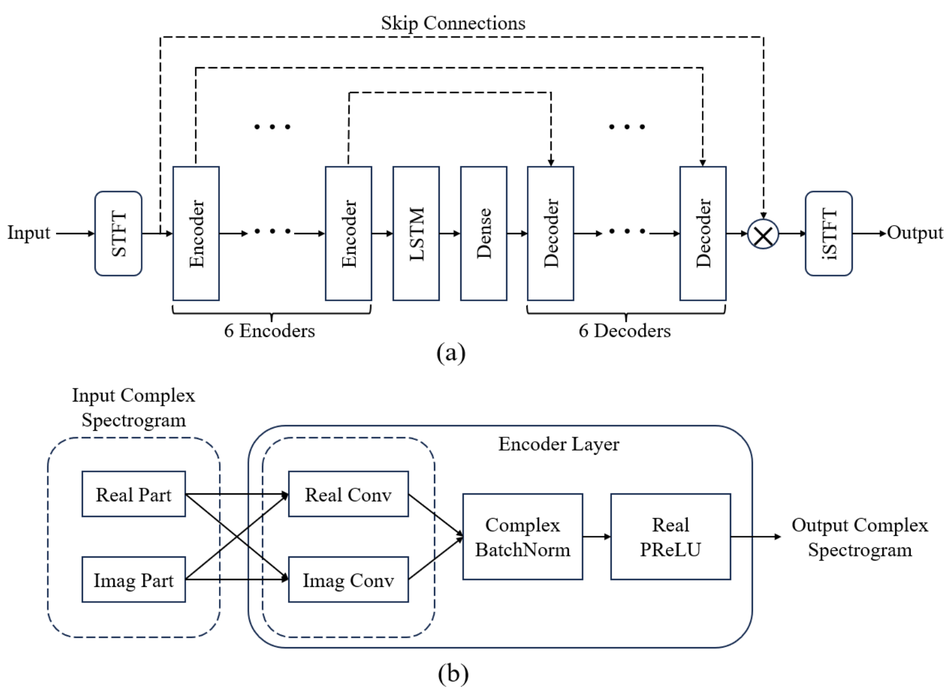

Figure 2. The architecture of the network in noise suppression stage (a), and its encoder layer (b).

As illustrated in Figure 2, in the noise suppression stage, we employ the DCCRN [1] network structure to process the complex spectrogram of 8 kHz noisy speech.

The loss function is as below:

$$L_{stage-1} = L_{SI-SNR} + 0.15L_{MSTFT}, \quad (1)$$

and the details of SI-SNR loss is defined in [1]. As for multi-scale STFT loss , we follow the settings in [2] and we do not list specific details here because of limited space.

STAGE 2: SUPER RESOLUTION

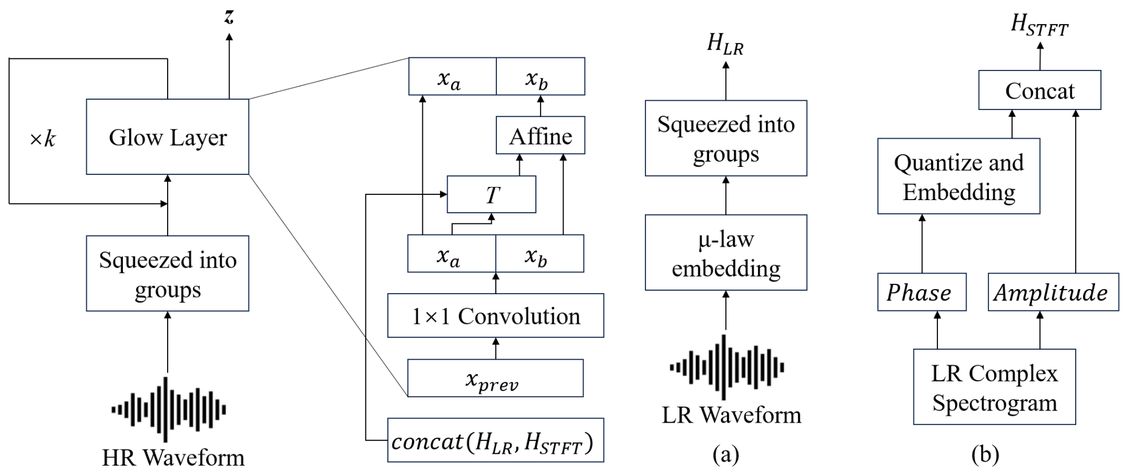

Figure 3. The architecture of the network in super resolution stage.

Figure 4. The LR encoder (a) and STFT encoder (b).

In stage 2 we use WSRGlow [3] to do the SR. The model operates in the opposite direction of the arrow in left part of Figure 3 at the inference phase and along the direction of the arrows during training. H_{LR} and H_{STFT} are generated by LR encoder and STFT encoder in Figure 4 to be the conditions of glow layers.

TRAINING DATA

Stage 1: The dataset provided by the ICASSP 2023 Deep Noise Suppression (DNS) challenge [4] and internal data are used to generate training data (200h).

Stage 2: The noisy 8 kHz speech passing through the first-stage network and its 16 kHz clean version (300h).

RESULTS

TABLE I. TEST RESULTS ON DNS NO-REVERB TEST SET

	PESQ-NB	PESQ-WB	STOI	CSIG	CBAK	COVL	LSD
Input	2.104	1.513	0.898	1.000	2.406	1.015	3.090
M-DCCRN	2.543	1.646	0.836	1.001	2.024	1.037	3.012
WSRGlow	2.148	1.434	0.910	2.430	2.345	1.900	1.610
Baseline	2.776	**1.734**	0.913	2.167	2.526	2.077	1.572
Ours	**2.784**	1.711	**0.919**	**2.647**	**2.769**	**2.321**	**1.425**

TABLE II. TEST RESULTS ON INTERNAL REAL-WORLD TEST SET

	PESQ-NB	PESQ-WB	STOI	CSIG	CBAK	COVL	LSD
Input	2.862	1.589	0.912	1.008	2.274	1.155	2.837
M-DCCRN	2.898	1.595	0.853	1.009	1.968	1.127	2.851
WSRGlow	2.799	1.508	0.916	2.588	2.398	2.002	1.666
Baseline	2.573	1.762	0.915	2.533	2.399	2.089	1.616
Ours	**2.957**	**1.901**	**0.925**	**2.811**	**2.434**	**2.314**	**1.368**

REFERENCES

[1] Z. Kong., et al., "Speech Denoising in the Waveform Domain With Self-Attention," in Proc. ICASSP, 2022, pp. 7867-7871.

[2] Hu, Y., et al., "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," Interspeech, 2020, 2472-2476.

[3] Zhang., et al., "WSRGlow: A Glow-Based Waveform Generative Model for Audio Super-Resolution," Interspeech, 2021, 1649-1653.

[4] Dubey, Harishchandra, et al. "ICASSP 2023 deep speech enhancement challenge," arXiv preprint arXiv:2303.11510, 2023.

CONTACT

E-Mail: s220101187@stu.cqupt.edu.cn