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

Junkang Yang
School of Communications and Information Engineering
Chongqing University of Posts and Telecommunications
Chongqing, China
s220101187@stu.cqupt.edu.cn

Hongqing Liu*
School of Communications and Information Engineering
Chongqing University of Posts and Telecommunications
Chongqing, China
* hongqingliu@outlook.com

Xing Li
AI Lab
vivo Mobile Communication Co., Ltd
Nanjing, China
li.xing2@vivo.com

Jie Jia
AI Lab
vivo Mobile Communication Co., Ltd
Nanjing, China
jiejia@vivo.com

***Abstract*—Super-resolution (SR) and noise suppression have been a hot research topic in the field of speech processing. Speech super-resolution, also named 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.**

***Keywords-speech super-resolution; noise suppression; deep learning; multi-stage network***

## I. INTRODUCTION

Channel noise can distort speech signals during transmission, causing the received signal at receivers to be a blend of the original signal and unwanted noise and leading to the loss or distortion of specific frequency components.

Speech super-resolution (SR) aims to convert low-resolution (LR) speech to high-resolution (HR) by recovering the high-frequency portion from distorted low-sampling-rate observations. In the frequency domain, it's known as bandwidth extension. Meanwhile, speech noise suppression focuses on removing background noise from recorded speech while maintaining perceptual quality and intelligibility. Both tasks have been widely studied due to their diverse applications [1].

After decades of developments, a large number of SR neural models have been proposed. Early methods based on deep neural network (DNN) [2] directly simulate the nonlinear relationship between the high frequency part and the low frequency part to improve the speech quality. The convolutional structures and Recurrent Neural Networks (RNN) have also been adopted in subsequent research, and these works extend the receptive field with less computational cost, addressing the lack of features captured by feed-forward convolutional models as well as the problem of vanishing gradients. In recent works, generative models such as generative adversarial networks (GAN) and diffusion-based approaches have achieved good results [3].

In recent years, deep learning has significantly advanced speech noise suppression, replacing traditional methods. Two main approaches involve processing the original audio waveform directly in the time domain or working with the time-frequency domain using spectrograms derived from short-time Fourier transform (STFT). Most deep learning methods are supervised-learning and categorized into DNN, RNN-LSTM, CNN [4], GAN and diffusion models based on network structure. However, these models often operate at higher sampling rates like 16 kHz or 48 kHz, presenting challenges for traditional communication systems with lower rates, such as 8 kHz for radio and wired telephones [5].

Although researches in speech super-resolution and noise suppression are well developed, very few studies have considered both at the same time, which is not consistent with real-world scenarios.

To solve this problem, in this paper we use a two-stage system to reduce the noise and extend the bandwidth of speech signals. Specifically, we use a multi-loss trained DCCRN [6] operating at the sampling rate of 8 kHz in the first stage to reduce the noise, and a WSRGlow [7] model to upsample the outputs of first stage to 16 kHz. Our experiments show that our methods largely outperform the similar model [5]. Furthermore, our method is also effective on the test data collected in real-world communication system.

## II. METHODS

As depicted in Figure 1, our system comprises two stages: the noise suppression stage and the super-resolution stage. The 8 kHz noisy signal input is processed by the noise suppression network first in spectrogram domain, and the denoised signal is then restored to the waveform using inverse STFT in this stage. This denoised waveform serves as the input for the super-resolution stage. Our super-resolution network operates in the waveform domain and directly converts the denoised 8 kHz waveform to a 16 kHz sampling rate. The spectrogram generated in the first stage serves as a condition for the super-resolution network, facilitating the enhancement of both noise suppression and bandwidth extension in a sequential manner.

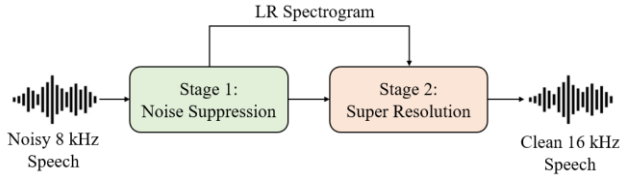


Figure 1. The architecture of our two-stage system.

### A. Stage 1: Noise Suppression

#### 1) Network Architecture

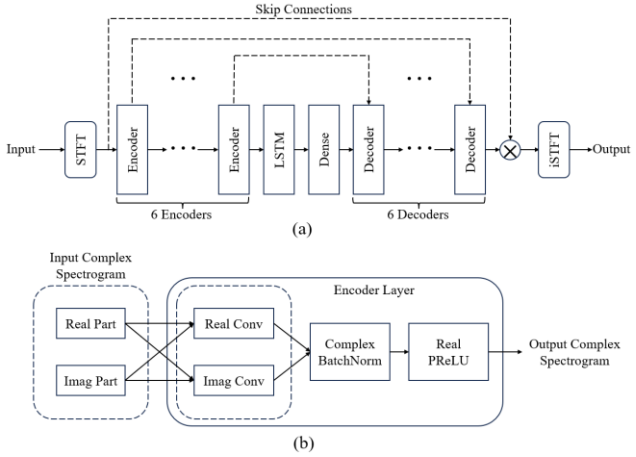


Figure 2. The architecture of the network in noise suppression stage (a), and its encoder layer (b).

In the noise suppression stage, we employ the DCCRN [6] network structure to process the complex spectrogram of 8 kHz noisy speech. As illustrated in Figure 2 (a), this network follows an encoder-decoder structure, consisting of 6 encoder and decoder layers. In the bottleneck layer, both LSTM and dense layers are incorporated. Specifically, as depicted in Figure 2 (b), within each encoder, the complex spectrogram is initially split into real part $X_r$ and imaginary part $X_i$. These components are then passed through a 2D complex convolution module, denoted as $W = W_r + jW_i$, where $W_r$ and $W_i$ represent the real and imaginary parts of a complex convolution kernel. The output of this complex convolution module can be described as:

$$S_{conv} = (X_r * W_r - X_i * W_i) + j(X_r * W_i + X_i * W_r). \quad (1)$$

After complex batch normalization and a real-value PReLU activation function, one encoder's output is formed. Skip connections exist between encoders and decoders, with the decoder structure resembling the encoder layer. In the bottleneck layer, the LSTM processes real and imaginary parts separately, enhancing the model's ability to capture complex spectrogram features.

#### 2) Loss Function

Unlike the original loss function of DCCRN, we us a multi-loss to train this network, it can be defined as

$$L_{stage-1} = L_{SI-SNR} + 0.15 L_{MSTFT}, \quad (2)$$

where SI-SNR loss is defined as

$$\begin{cases} x_{target} = (<\hat{x}, x> \cdot x) / \|x\|_2^2 \\ e_{noise} = \hat{x} - x \\ L_{SI-SNR} = 10\log 10(\dfrac{\|x_{target}\|_2^2}{\|e_{noise}\|_2^2}) \end{cases}, \quad (3)$$

$\hat{x}, x$ are denoised signal and ground truth respectively. $<\cdot,\cdot>$ is dot product and $\|\cdot\|_2$ denotes to L2 norm. As for multi-scale STFT loss $L_{MSTFT}$, we follow the settings in [1] and we do not list specific details here because of limited space.

### B. Stage 2: Super Resolution

#### 1) Inference Phase

In super resolution stage, WSRGlow [7] is used to produce high-resolution (HR) speech as following way. A sample $\boldsymbol{z}$ from a simple gaussian distribution will first be selected to undergo a series of invertible transformations through several layers, represented as $\boldsymbol{f}_i$. These layers collectively convert the initial simple distribution into the expected distribution $\boldsymbol{x}$, expressed as $\boldsymbol{x} = \boldsymbol{f}_0 \circ \boldsymbol{f}_1 \circ \cdots \circ \boldsymbol{f}_k(\boldsymbol{z})$. To summarize, the model operates in the opposite direction of the arrow in left part of Figure 3 at the inference phase.

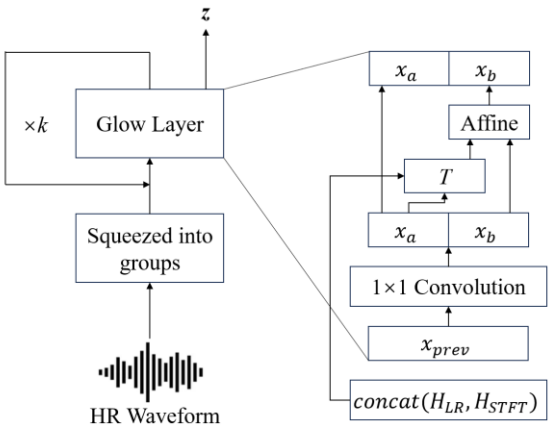


Figure 3. The architecture of the network in super resolution stage.

*2) Training Phase*

The model runs along the direction of the arrows in Figure 3 during training. Specifically, HR waveform is divided into 8 groups and it will transform into simple distribution $\boldsymbol{z}$ after passing $k$ glow layers. This process can be defined as:

$$\boldsymbol{z} = \boldsymbol{f}_k^{-1} \circ \boldsymbol{f}_{k-1}^{-1} \circ \cdots \circ \boldsymbol{f}_0^{-1}(\boldsymbol{x}). \tag{4}$$

The structure of the glow layer is as the right part of Figure 3, each of the glow layers consists of an invertible 1×1 convolution and an affine coupling layer. The feature vectors of HR waveform $\boldsymbol{x}_{prev}$ is split into $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ by this invertible convolution, and these two parts make up the output, $concat(\boldsymbol{x}_a, \boldsymbol{x}_b')$, after the T-transformation and affine layer proposed in [7]. The two encoder outputs $H_{LR}$ and $H_{STFT}$ of the LR signal are used as intermediate conditions during the T-transform. As depicted in Figure 4, $H_{LR}$ is obtained by setting $\mu$-law transformation, quantification and embedding. $H_{STFT}$ can be expressed as:

$$H_{STFT} = concat(A(\boldsymbol{x}), P(\boldsymbol{x})), \tag{5}$$

$$A(\boldsymbol{x}) = \text{real}(\mathbf{S}), \tag{6}$$

$$P(\boldsymbol{x}) = \text{Embedding}(\text{Quantize}(\text{imag}(\mathbf{S}))), \tag{7}$$

where $\mathbf{S}$ is the complex spectrogram generated in stage 1. And the network is optimized by the negative loglikelihood directly.

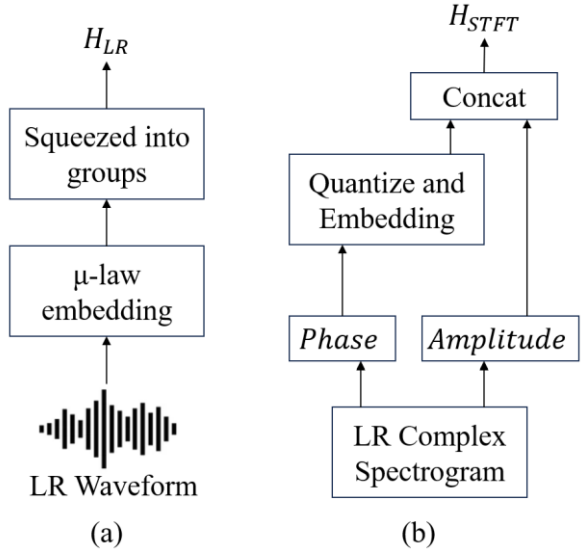


Figure 4. The LR encoder (a) and STFT encoder (b).

To recapitulate, the network uses LR audio as a condition during training and gradually achieves a reversible bi-directional mapping from HR signal to $\boldsymbol{z}$ and from $\boldsymbol{z}$ to HR signal through supervised learning.

## III. Experiments

### A. Data Preparation

To train the network of stage 1, we use the dataset provided by the ICASSP 2023 Deep Noise Suppression (DNS) challenge [8] and internal data to generate training data. We synthesize 200 hours of clean and noisy speech pairs by randomly mixing speech and noise, with each sample lasting 10 seconds. All samples have a signal-to-noise ratio of -5-10 dB and a sampling rate of 8 kHz. And we use 10% of the total data to do the validation.

After the regression of the first stage, we synthesize 300 hours of clean and noisy speech with a sampling rate of 16 kHz, where we downsample the noisy part to 8 kHz and let them pass through the first-stage network, and this processed portion of the data, along with the 16 kHz clean data, comprise the training data for the second stage. Also, 10% of the total data is picked out to do the validation.

For testing the whole system, we use the no-reverb test set of DNS Challenge 2020 [9]. The noisy speeches are downsampled to 8 kHz before putting into the system. In addition, the internal test dataset is also taken to analyze the performance of our method. This internal test dataset contains 200 English speech samples collected from cell phone calls in real-world scenario such as streets, restaurants, etc.

### B. Baselines

In our experiments, we compare our method with unprocessed input signals and a model similar to ours. The details are as follows.

- Unprocessed inputs: All inputs consist of speech signals with background noise at a sampling rate of 8 kHz. When calculating metrics, we directly resample them to 16 kHz to ensure a consistent scale with clean references.
- M-DCCRN and WSRGlow: We take the networks in each stage separately and test them as a way to verify their effectiveness and robustness.
- UNet+AFiLM and I-DTLN [5]: This is another two-stage audio super-resolution and noise cancellation system based on UNet+AFiLM and I-DTLN. The authors claim it can effectively separate noise-filled voices into clean human voices.

### C. Metrics

To assess the quality of the generated speech, we use the objective evaluation metrics including narrow-band PESQ (PESQ-NB, 0-8 kHz) and wide-band PESQ (PESQ-WB, 8-16 kHz) [10], STOI [11], CSIG, CBAK, COVL scores [12], and log spectral distance (LSD). For all metrics in this paper, except LSD, a higher score indicates better performance.

Table 1 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 |
| WSR-Glow | 2.148 | 1.434 | 0.910 | 2.430 | 2.345 | 1.900 | 1.610 |
| Base-line | 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** |

### *D. Results & Analysis*

Table 1 shows the results of our experiments on the DNS test set. "Baseline" denotes to the method in [5]. Without any processing, the input signal is low in all metrics and speech intelligibility is very poor. When the noise reduction network acts alone, the output speech has a higher PESQ in the narrowband, but is still not clear enough overall because the wideband portion is still missing. When the SR network is introduced alone, the output speech is significantly improved in terms of signal quality, but since it does not have suppression of low-frequency noise, the speech has a portion of bias distortion at both low and high frequencies, resulting in poor PESQ in both narrowband and wideband. On the other hand, our method exceeds all other methods in almost all metrics, which indicates that our method is able to generate audio with higher quality. In addition, our method substantially improves the speech quality compared to the input signal, and Figure 5 shows that the speech generated by our model not only has few noises in the low-frequency part, but also has a more accurate and fuller high-frequency part.

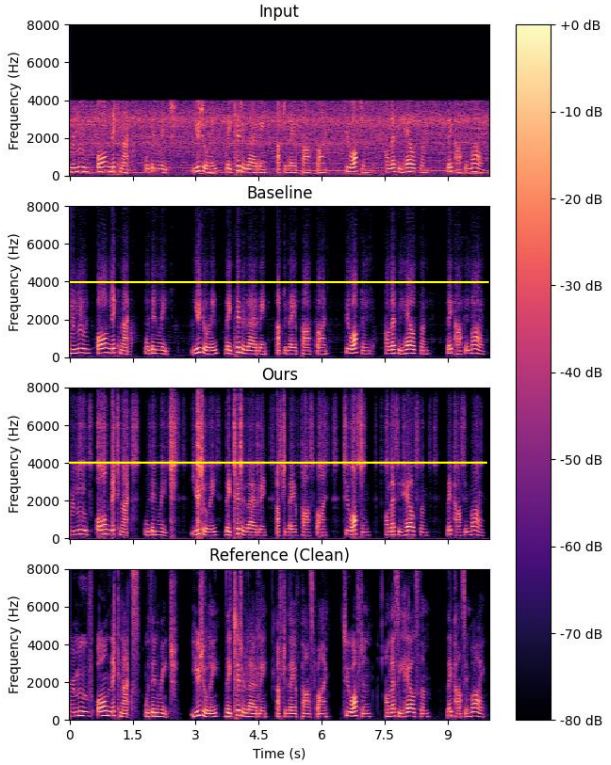


Figure 5. The spectrogram of signals generated by different methods.

As illustrated in table 2, when we use real-world data for comparison, the performance of all other methods has a degradation, which may be due to the fact that these methods are trained using synthesized data. On the contrary, our system still maintains a good performance under the test, leading in all metrics, confirming the robustness of our model in real-world communication system and the mismatch between previous methods and the real-world situation.

Table 2 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 |
| WSR-Glow | 2.799 | 1.508 | 0.916 | 2.588 | 2.398 | 2.002 | 1.666 |
| Base-line | 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** |

## IV. Conclusion

In this paper, we consider a real-world scenario, we deal with the speech super-resolution task in a noisy environment using a two-stage network. We focus on suppressing the noise in the first stage, and in the second stage we focus on complementing the missing spectral components. Experiments demonstrate that our approach achieves better results on both open-source datasets as well as internal real-world datasets. However, sometimes the speech generated by our model has some rustling and the parameter of the system is heavy, which will be something we will optimize in the future.

## References


[1] Z. Kong, W. Ping, A. Dantrey and B. Catanzaro, "Speech Denoising in the Waveform Domain with Self-Attention," Proc. ICASSP, 2022, pp. 7867-7871.

[2] K. Li and C. -H. Lee, "A deep neural network approach to speech bandwidth expansion," Proc. ICASSP, 2015, pp. 4395-4399.

[3] Han, S., Lee, J., "NU-Wave 2: A General Neural Audio Upsampling Model for Various Sampling Rates," Interspeech, 2022, 4401-4405.

[4] K. Kinoshita, T. Ochiai, M. Delcroix and T. Nakatani, "Improving Noise Robust Automatic Speech Recognition with Single-Channel Time-Domain Enhancement Network," Proc. ICASSP, 2020, pp. 7009-7013.

[5] C. -W. Chen, W. -C. Wang, Y. -Y. Ou and J. -F. Wang, "Deep Learning Audio Super Resolution and Noise Cancellation System for Low Sampling Rate Noise Environment," 2022 10th International Conference on Orange Technology (ICOT), 2022, pp. 1-5.

[6] Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., Xie, L., "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," Interspeech, 2020, 2472-2476.

[7] Zhang, K., Ren, Y., Xu, C., Zhao, Z. "WSRGlow: A Glow-Based Waveform Generative Model for Audio Super-Resolution," Interspeech, 2021, 1649-1653.

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

[9] Reddy, Chandan KA, et al. "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," arXiv preprint arXiv:2005.13981, 2020.

[10] Recommendation, ITU-T. "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Rec. ITU-T P. 862, 2001.

[11] Taal, Cees H., et al. "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE Transactions on Audio, Speech, and Language Processing 19.7, 2011, 2125-2136.

[12] Hu, Yi, and Philipos C. Loizou. "Evaluation of objective quality measures for speech enhancement." IEEE Transactions on audio, speech, and language processing 16.1, 2007, 229-238.