This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/APSIPAASC63619.2025.10848923, 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

SDNet: Noise-Robust Bandwidth Extension under Flexible Sampling Rates

Junkang Yang*§, Hongqing Liu*, Lu Gan†, Yi Zhou*, Xing Li‡, Jie Jia‡ and Jinzhuo Yao*
* School of Communications and Information Engineering,
Chongqing University of Posts and Telecommunications, Chongqing, China
† College of Engineering, Design and Physical Science, Brunel University, London, U.K.
‡ AI Lab, vivo Mobile Communication Co,.Ltd, Nanjing, China
E-mail: s220101187@stu.cqupt.edu.cn

***Abstract*—Bandwidth extension (BWE), also known as audio super-resolution (SR), aims to predict a high resolution (HR) speech signal from its low resolution (LR) corresponding part. Most neural BWE models work at a specific sampling rate but, producing the final result in a noise-free environment by recovering the spectrogram of high-frequency part of the signal and concatenating it with the original low-frequency part. Although these methods achieve high accuracy, they become less effective when facing the real-world scenario, where unavoidable noise is present and sampling rates are flexible. To address this problem, we propose Super Denoise Net (SDNet), a neural network for a joint task of BWE and noise reduction from a flexible low sampling rate signal. To that end, we design gated convolution and lattice convolution blocks to enhance the repair capability and capture information in the time-frequency axis, respectively. The experiments show our method outperforms all current state-of-the-art (SOTA) noise-robust BWE model in Valentini-Botinhao test set. Our model also outperforms other baselines on DNS 2020 no-reverb test set with higher objective and subjective scores.**

I. Introduction

Bandwidth extension (BWE) is the task to reconstructing the high resolution (HR) part of the speech signal from its low-resolution (LR) part, which is also termed as audio super-resolution (SR). As one of the important tasks in the front-end of speech processing, BWE is widely applied in wireless communication, speech recognition [1], text-to-speech [2], to name a few.

Although BWE has achieved considerable progress in past years, few studies have focused on the task of BWE in noisy environments, i.e., bandwidth extension along with the noise suppression. Most existing frequency domain based work basically keeps the low-frequency part and predicts only the high-frequency part, and finally concatenates the two parts in series. When noise exists in the low-frequency part, this pipeline not only fails to remove the noise, but also produces the biased prediction of high-frequency part due to noise interference. To address this issue, in [3], the authors introduced a multi-stage model to respectively conduct noise reduction and bandwidth extension. In [4], authors combine UNet+AFiLM [5] and an improved DTLN [6] to form a two-stage system. The authors in [7] simultaneously estimate the missing components and the noise distribution in degraded speech signal with a DNN. For the models of MTL-MBE [8] and EP-WUN [9], the authors claimed that they achieved the SOTA performance of this task. On the other hand, there are a lot of speech signals with flexible sampling rates or known sampling rates but small effective bandwidths (the high-frequency components are missing), and so far there is no noise-robust BWE model can deal with this situation.

In this paper, we propose Super Denoise Net (SDNet), a neural network that removes noise while extending the bandwidth. To that aim, we design a generator and discriminator network, where encoder-bottleneck-decoder structure is utilized in generator and multi-discriminator structure is developed, enabling the possibility of adversarial training to make the model robust. It is of interest to note that the proposed model does not require the prior information of the sampling rate of LR speech. Our main contributions are provided as follows.

- We develop a single-stage network, where the gated convolution and lattice convolution blocks are utilized to jointly perform noise reduction and SR. Our method has improved the performance of model and outperformed the baselines in both noisy and noise-free cases.
- Our approach is one of the first noise-robust BWE model that supports to process all the speeches whose sampling rates are from 4 kHz to 16 kHz, which means that the sampling rate of LR speech is flexible in our case.
- The quality of the speeches generated by our model from 8 kHz noisy samples are even better than those of some popular models only focusing on 16 kHz to 16 kHz noise reduction.

II. Methods

A. Problem Settings

In this paper, we address the problem of recovering HR clean speech from LR noisy speech. Given a HR clear speech $y \in R^T$, where T is the length, after downsampling s times and adding noise at the same sampling rate with the same length, the LR noisy speech $x \in R^{\frac{T}{s}}$ is generated. Our goal is to design a function G that can efficiently predict y from the observation x, i.e., $\hat{y} = G(x) \approx y$. The same representation as above will be used in the formulas below.

§ Part of work was done during internship at vivo AI lab.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/APSIPAASC63619.2025.10848923, 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

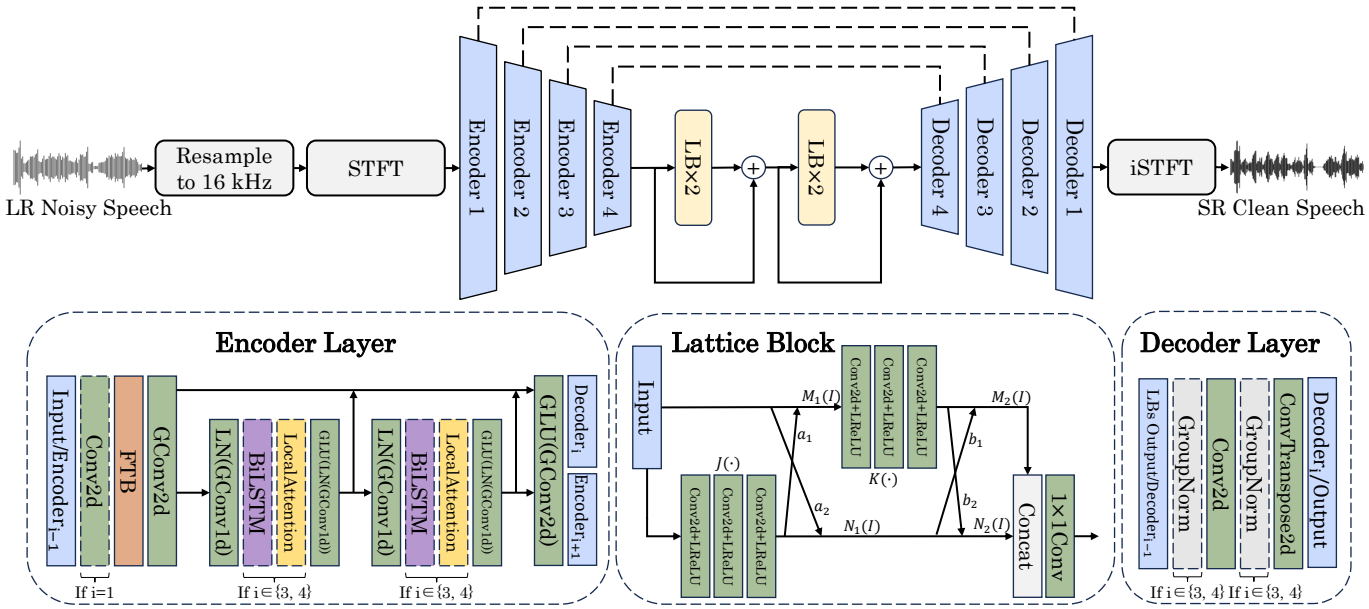

Fig. 1. The generator network architecture.

B. Network Architecture

Our model uses a U-shaped structure, containing encoders, decoders, and lattice convolution blocks (LBs) as bottleneck layers. The residual connections are set between encoders and decoders, also between the lattice convolution blocks. Model architecture is visualized in Figure 1. Due to space limit, we provide parameter settings, input size, and output size for each layer at our demo page[1].

1) Generator:

Encoder and decoder. As illustrated in Figure 1, there are 4 layers in encoder and decoder each. The input of encoder is the spectrogram of upsampled waveform and it will be reshaped at the first encoder layer with a 2D convolution, with the complex part moved to channel dimension. After that, a frequency transform block (FTB) [10] is applied to capture the non-local correlations in spectrogram along the frequency axis. Unlike previous work [11], we utilize gated convolution (GConv) instead of ordinary convolution in the later structure, which has been proved to enhance the model's generation ability by learning a dynamic feature selection mechanism for each channel and each spatial location [12]. Inside the encoder, there are two residual branches with two 1D gated convolutions at the beginning and end. In the middle, there are LSTM and temporal-based attention modules to capture long-distance relations. Followed by each encoder layer is a decoder layer that recovers the latent vectors equal to the size of spectrogram before passing the encoder. It is worth noting that there is concatenated residual connection between each encoder and decoder layer, while between two residual branches within the encoder layer, it is a summation residual connection.

[1]https://sdnetdemo.github.io/

Bottleneck layers. The bottleneck layers include 4 lattice convolution blocks (LBs), which were first proposed in image restoration task [13]. Combined with gated convolution, this structure can offer a blend of structured interpolation with adaptive and long-range context modeling. As shown in Figure 1, each LB includes paired butterfly-style structures. The input passes two branches that contain several convolution layers and LeakyReLU activation layer is followed for each layer. The two branches interact with each other through learnable combination coefficients. Specifically, given an input feature I, the first combination is

$$M_1(I) = I + a_1 J(I), \tag{1}$$

$$N_1(I) = a_2 I + J(I), \tag{2}$$

where $J(\cdot)$ denotes to the implicit non-linear function of several layers shown in Figure 1. Similarly, the second combination is

$$M_2(I) = b_1 N_1 + K(M_1(I)), \tag{3}$$

$$N_2(I) = N_1 + b_2 K(M_1(I)). \tag{4}$$

Afterwards, the output of two branches are merged in channel dimension and then compressed through a 1×1 convolution layer. The final output is

$$O = Conv(Concat(M_2(I), N_2(I))). \tag{5}$$

The combination coefficients are mainly determined in the following way. The mean and standard deviation in channel dimension are first obtained by global mean pooling in the upper branch and global standard deviation pooling in the lower branch. Then, those statistics in the two branches are passed through two fully connected layers, each followed by

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/APSIPAASC63619.2025.10848923, 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

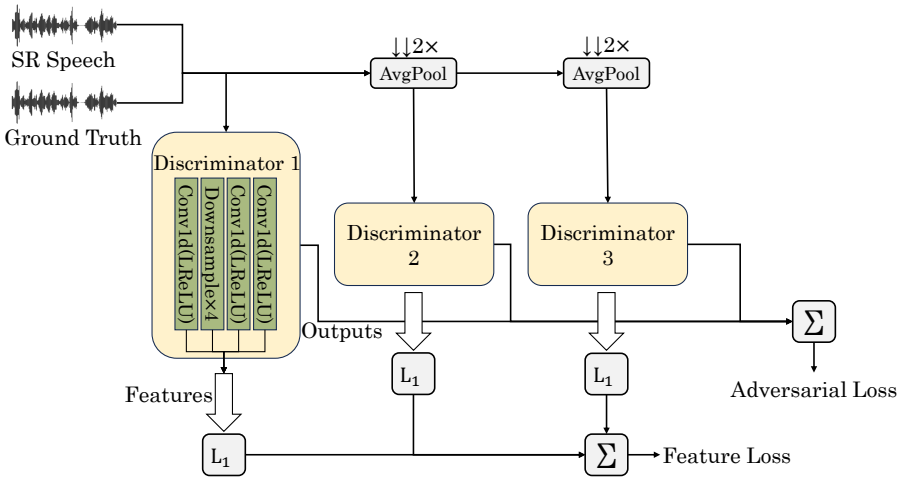

Fig. 2. The multi-scale discriminators.

ReLU and Sigmoid activation functions, respectively. Finally, the outputs of the two branches are averaged to obtain the combined coefficients.

2) Discriminators:

To implement multi-loss training in an adversarial manner, we employ multi-scale discriminators, whose structure is shown in Figure 3. The inputs to the discriminator system are super-resolution (SR) speech and high-resolution reference signals generated by the generator. It contains 3 discriminators (D_1, D_2, D_3), each with the same structure as in MelGAN. Specifically, there are 7 convolutional layers in each discriminator, 4 of which have downsampling capabilities. When the data pass through each layer, real and fake features at different scales will be produced, which are used to compute the feature loss. Also, the data passing through the discriminator will become an output, and the outputs will be used to compute the adversarial loss of the generator and discriminator. Furthermore, the inputs of D_1, D_2, and D_3 are original, 2-times down-sampled, and 4-times down-sampled waveforms, respectively. For interested readers, see more details of discriminators in [14].

C. Loss Function

The model is trained with an adversarial approach. We use a multi-scale STFT loss with FFT bins $\in \{512, 1024, 2048\}$ and hop length $\in \{50, 120, 240\}$ to form one part of the loss function. The window lengths are $\{240, 600, 1200\}$. On the other hand, the multi-scale adversarial and feature losses in time domain is added in. The total loss can be expressed as

$$\mathcal{L} = \mathcal{L}_{MSTFT} + \mathcal{L}_G^{adv} + \lambda_f \mathcal{L}_f, \tag{6}$$

where $\lambda_f = 100$, $\mathcal{L}_{MSTFT}$, $\mathcal{L}_G^{adv}$ and $\mathcal{L}_f$ are multi-scale STFT loss, adversarial loss of generator, and feature loss, respectively. Let $s(x, \theta_m)$ denote $|STFT(x)|$ with the m^{th} hyperparameters θ_m, the multi-scale STFT loss is defined as

$$\mathcal{L}_{MSTFT} = E_{(x,y)\sim p_{data}} \left[\sum_{m=1}^{3} \left(\frac{||s(y,\theta_m) - s(x,\theta_m)||_F}{||s(y,\theta_m)||_F} + \frac{1}{N} ||log \frac{s(y,\theta_m)}{s(x,\theta_m)}|| \right) \right], \tag{7}$$

where $||\cdot||_F$ and $||\cdot||_1$ are Frobenius and ℓ_1-norms, N is the number of elements in the magnitude.

As shown in Figure 3, the latter two loss functions can be depicted as

$$\mathcal{L}_G^{adv} = E_{x\sim p_{data}} \left[\frac{1}{K} \sum_k max(0, 1 - D_k(G(x))) \right], \tag{8}$$

$$\mathcal{L}_f = E_{(x,y)\sim p_{data}} \left[\frac{1}{KL} \sum_{k,l} ||D_k^l(y) - D_k^l(G(x))||_1 \right], \tag{9}$$

where $k = 1, ..., K$ is the number of discriminators, $l = 1, ..., L$ is the number of layer in one discriminator.

III. Experiments

A. Dataset

We use the dataset provided by Deep Noise Suppression (DNS) Challenge at ICASSP 2023 [16] and Valentini-Botinhao [17] to generate training data containing noise. We synthesize clean and noisy speech pairs with 500 hours by randomly mixing the speech and noise, and each sample lasts for 5 seconds. The SNR of all samples are at -5 - 20 dB and the sampling rate is 16 kHz. We then downsample all the noisy speech by a factor of $s = 2$, i.e., turning the paired data into 8 kHz noisy speech and 16 kHz clean speech. For validation, we separate %10 (50 hours) of total data as validation set. For testing, no-reverb test set of DNS Challenge 2020 [18] and Valentini-Botinhao test set are used and the noisy samples are also downsampled if necessary.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/APSIPAASC63619.2025.10848923, 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

TABLE I
TEST RESULTS OF NOISE-ROBUST BWE MODELS ON VALENTINI-BOTINHAO NOISY TEST SET DOWNSAMPLED TO 8 KHZ.

Method	PESQ-WB↑	STOI (%)↑	CSIG↑	CBAK↑	COVL↑	LSD↓
UEE [15]	2.23	93	2.27	2.39	2.17	2.72
MTL-MBE [8]	2.55	94	2.64	3.21	2.46	2.29
EP-WUN [9]	2.25	92	**3.50**	2.94	2.86	1.23
AFiLM + I-DTLN [4]	2.54	90	2.63	2.87	2.18	1.54
Ours	**2.67**	**95**	3.29	**3.32**	**2.92**	**1.16**

TABLE II
TEST RESULTS FOR DIFFERENT TASK ON DNS-CHALLENGE NO-REVERB TEST. "B" IS NOISE-FREE BWE, "D" IS DENOISE, AND "RB" IS NOISE-ROBUST BWE. "SOURCE" AND "NOISE" REPRESENT THE SAMPLING RATE OF INPUTS AND THE CASE WHETHER THE INPUTS CONTAIN NOISES.

Method	Task	Source	Noise	PESQ-NB↑	PESQ-WB	STOI(%)	CSIG	CBAK	COVL	LSD	MOS↑
WSRGlow	B	8 kHz	×	4.365	2.811	**99.4**	3.946	4.068	3.433	0.929	4.21
NU-Wave 2				4.353	2.646	**99.4**	3.663	2.869	3.209	1.328	4.08
AERO				4.369	3.295	98.5	**4.287**	4.273	3.844	0.802	4.27
Ours				**4.377**	**3.661**	98.6	4.103	**4.553**	**3.935**	**0.783**	**4.55**
DCCRN	D	16 kHz	✓	3.17	2.64	92.9	—	—	—	—	—
FullSubNet				3.28	2.72	95.3	—	—	—	—	—
DPT-FSNet				3.28	2.72	95.3	—	—	—	—	—
Ours				**3.29**	**2.80**	**96.0**	—	—	—	—	—
VoiceFixer	RB	8 kHz	✓	2.535	1.679	84.0	2.532	1.914	2.043	1.323	3.83
Ours				**3.554**	**2.777**	**97.1**	**3.313**	**3.532**	**3.063**	**1.218**	**4.38**
VoiceFixer	RB	4-16 kHz	✓	2.540	1.822	84.2	2.737	1.984	2.222	1.280	3.89
Ours				**3.550**	**3.013**	**97.3**	**3.657**	**3.726**	**3.355**	**1.112**	**4.43**

For the case of uncertain low sampling rate, we adopt same operation as above, except for downsampling. We use a filter with random parameters when doing downsampling, the types include *Chebyshev, Elliptic, Butterworth* and *Boxcar*, the order is a random integer from 2 to 10, the cutoff frequency is an integer from 2000 to 8000 Hz.

B. Training Details

Unlike the complex training policies such as multi-stage training, variable learning rate and warmming up used in previous work, our training method is single-stage and simple. We use an Adam optimizer ($\beta_1 = 0.8, \beta_2 = 0.999$) to optimize both generator and discriminator with a stable learning rate of 1×10^{-4}. We train the model on NVIDIA RTX3090 GPUs for 200 epochs and select the checkpoint which has the best performance on validation data to test. The FFT bins and hop length of the STFT operation in our network is 512 and 64 respectively.

C. Baselines

For baselines, we consider both cases of noise-robust and noise-free BWEs, and they are

1) noise-robust BWE: we compare our model with the previous SOTA methods of MTL-MBE and EP-WUN in [8], [9], in the same test set. Since the authors did not provide the source code, we have re-implemented the method proposed in [4] to produce the results. For uncertain input sampling rate case, the baseline is VoiceFixer [19].
2) noise-free BWE: we compare our model with WSRGlow [20], NU-Wave 2 [21], and AERO [11].

To further demonstrate the capability of our model, we also compare our model with 16 kHz to 16 kHz denoise-only models of DCCRN [22], FullSubNet [23], and DPT-FSNet [24]. For all the models on denoise task only, we have referred to the results provided in [25]. This comparison further showcases our model presents a high capacity of removing the noise and generating high quality speech.

D. Evaluation Metrics

To evaluate the quality of the generated speech, both objective and subjective evaluation metrics were used. The objective evaluation metrics used were PESQ [26] including narrow-band one (PESQ-NB, 0-8 kHz) and wide-band one (PESQ-WB, 8-16 kHz), STOI [27], CSIG, CBAK, COVL scores[28] and log spectral distance (LSD). The subjective evaluation metrics is overall Mean Opinion Score (MOS) [29]. We randomly selected 50 samples from the test set and asked 15 people to provide overall MOS values of each sample. The final MOS score is the average of these evaluations. For all metrics in this paper except LSD, the higher score means a better performance.

E. Results and Analysis

Table I summarizes the comparison between our model and noise-robust BWE methods. The results show that our SDNet performs best in all the metrics except for CSIG. This suggests that the audio generated with our method has a higher quality, which meets our expectation to our network and datasets

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/APSIPAASC63619.2025.10848923, 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

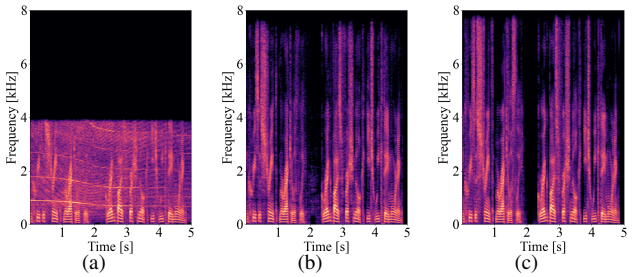

Fig. 3. Spectrograms of noise-robust BWE task results. (a) Input; (b) our method; (c) ground truth.

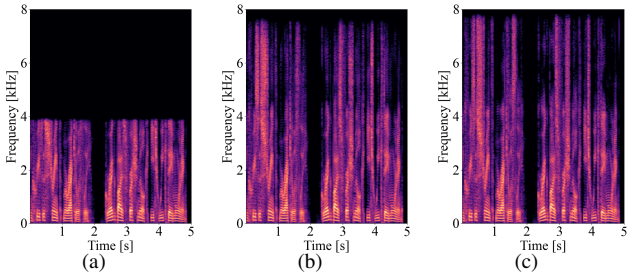

Fig. 4. Spectrograms of noise-free BWE task results. (a) Input; (b) our method; (c) ground truth.

designs. As for the mild drop in CSIG, this may because the over-suppression of noise hurts the speech component.

In 8 kHz to 16 kHz noise-free BWE task, our SDNet remains competitive in most metrics (see Table II), especially in PESQ-WB and CBAK, which shows that our generated wide-band signals have a better audibility and a more reasonable handling of background sounds. At the same time, it shows that our focus on BWE under noise does not sacrifice its performance in a noise-free environment.

In a comparison with VoiceFixer, our method outperforms it in all metrics for both 8 kHz to 16 kHz and 4-16 kHz to 16 kHz noise-robust BWE tasks. VoiceFixer aims to repair many distortions such as clipping, reverberation, and we find the speeches produced mismatch with the reference signal in terms of loudness, etc., which causes the degradation of its performance in objective metrics, but in subjective metrics, the scores of these speeches are still very high, which shows its repair is still very effective.

For the flexible sampling rate scenario (the last row in Table II), on the other hand, with a very limited bandwidth information, our model still improve the speech quality effectively, which is rare among the existing noise-robust BWE models. On 16 kHz to 16 kHz noise reduction task, our method performance better than the baseline denoise-only models, and surprisingly, the PESQ and STOI of the speeches generated by our model from 8 kHz noisy samples are even higher than the popular baseline models in Table II only focusing on 16 kHz to 16 kHz noise reduction.

For the same task, our model performs slightly better on the DNS test set than the Valentini-Botinhao's data set. This is because the speech lengths in the latter are generally very short, which prevents the model from utilizing contextual information to make more accurate predictions, resulting in a slight performance degradation.

We also tested our model with real-world data, and the results showed that the noise in these speeches was largely suppressed and the human voice was much clearer. Due to space limit, we provide the audio samples at our demo page. Figure 3 and 4 are the spectrograms of the results for different tasks.

TABLE III
ABLATION STUDY OF NETWORK STRUCTURE BASED ON 8 KHZ "RB" TASK OF DNS-CHALLENGE NO-REVERB TEST SET.

Method	PESQ-NB	PESQ-WB	LSD
SDNet	**3.554**	**2.777**	**1.218**
w/o GConv	3.445 (-0.109)	2.630 (-0.147)	1.256 (+0.038)
w/o LBs	3.442 (-0.112)	2.633 (-0.144)	1.262 (+0.044)
w/o Both	3.372 (-0.182)	2.538 (-0.239)	1.293 (+0.075)

F. Ablation Studies

In order to study the impact of network components on network performance, we conducted ablation studies by gradually removing some components of the original model. The experiments show that the original model produces the best results. When gated convolutions are replaced by general convolutions ("w/o GConv" in Table III), the performance of network declines due to the missing details at 6-8 kHz. When LBs only are removed, the performance is also degraded because of the lack of utilization of the time dimension information in the spectrogram. When both changes act together, the accuracy of the model drops dramatically.

IV. CONCLUSION & FUTURE WORK

In this paper, we proposed SDNet, a U-shaped encoder-decoder neural network to jointly handle BWE and denoise tasks in low sampling rate noisy environments. Experiments demonstrated that our model outperforms all baseline models in both objective and subjective metrics in different cases. The ablation studies have verified that our use of gated convolutions and LBs enhances the performance of the model. However, our model also have some limitations. When we try to train the model at a higher resolution such as 48 kHz, we find it is hard to deal with two tasks at one stage, and this is a common issue encountered by many models. In our future work, we will continue this task at higher resolutions and we will also consider adding music and other personalized datasets.

REFERENCES

[1] D. Haws and X. Cui, "Cyclegan bandwidth extension acoustic modeling for automatic speech recognition," in *Proc. ICASSP*, 2019, pp. 6780–6784.

[2] R. Yoneyama, R. Yamamoto, and K. Tachibana, "Non-parallel high-quality audio super resolution with domain adaptation and resampling cyclegans," in *Proc. ICASSP*, 2023, pp. 1–5.

[3] H. Seo, H.-G. Kang, and F. Soong, "A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise," in *Proc. ICASSP*, 2014, pp. 6087–6091.

[4] 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," in *2022 10th International Conference on Orange Technology (ICOT)*, 2022, pp. 1–5.

[5] N. C. Rakotonirina, "Self-attention for audio super-resolution," in *Proc. MLSP*, 2021, pp. 1–6.

[6] N. L. Westhausen and B. T. Meyer, "Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression," in *Proc. Interspeech*, 2020, pp. 2477–2481.

[7] T. Taher, N. Mamun, and M. A. Hossain, "A joint bandwidth expansion and speech enhancement approach using deep neural network," in *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2023, pp. 1–4.

[8] N. Hou, C. Xu, J. T. Zhou, E. S. Chng, and H. Li, "Multi-Task Learning for End-to-End Noise-Robust Bandwidth Extension," in *Proc. Interspeech 2020*, 2020, pp. 4069–4073.

[9] Y.-T. Lin, B.-H. Su, C.-H. Lin, S.-C. Kuo, J.-S. R. Jang, and C.-C. Lee, "Noise-Robust Bandwidth Expansion for 8K Speech Recordings," in *Proc. INTERSPEECH 2023*, 2023, pp. 5107–5111.

[10] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 9458–9465.

[11] M. Mandel, O. Tal, and Y. Adi, "Aero: Audio super resolution in the spectral domain," in *Proc. ICASSP*, 2023, pp. 1–5.

[12] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4471–4480.

[13] X. Luo, Y. Qu, Y. Xie, Y. Zhang, C. Li, and Y. Fu, "Lattice network for lightweight image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4826–4842, 2022.

[14] K. Kumar, R. Kumar, T. De Boissiere, *et al.*, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[15] B. Liu, J. Tao, and Y. Zheng, "A novel unified framework for speech enhancement and bandwidth extension based on jointly trained neural networks," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, 2018, pp. 11–15.

[16] H. Dubey, A. Aazami, V. Gopal, *et al.*, "Icassp 2023 deep noise suppression challenge," in *ICASSP*, 2023.

[17] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks," in *Proc. Interspeech 2016*, 2016, pp. 352–356.

[18] C. K. Reddy, V. Gopal, R. Cutler, *et al.*, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results," in *Proc. Interspeech 2020*, 2020, pp. 2492–2496.

[19] H. Liu, X. Liu, Q. Kong, *et al.*, "VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration," in *Proc. Interspeech*, 2022, pp. 4232–4236.

[20] K. Zhang, Y. Ren, C. Xu, and Z. Zhao, "WSRGlow: A Glow-Based Waveform Generative Model for Audio Super-Resolution," in *Proc. Interspeech*, 2021, pp. 1649–1653.

[21] S. Han and J. Lee, "Nu-wave 2: A general neural audio upsampling model for various sampling rates," in *Proc. Interspeech*, 2022, pp. 4401–4405.

[22] Y. Hu, Y. Liu, S. Lv, *et al.*, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.

[23] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *Proc. ICASSP*, 2021, pp. 6633–6637.

[24] F. Dang, H. Chen, and P. Zhang, "Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *Proc. ICASSP*, 2022, pp. 6857–6861.

[25] Y. Li, Y. Sun, W. Wang, and S. M. Naqvi, "U-shaped transformer with frequency-band aware attention for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1511–1521, 2023.

[26] I.-T. Recommendation, "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.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[28] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.

[29] I.-T. P. 835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," *ITU-T Recommendation*, 2003.