Multi-stage Speech Enhancement with Cascaded SNR Domain Shifts

Xiaoran Li* and Zilu Guo* and Jun Du*†
* University of Science and Technology of China, China
E-mail: lixiaoran@mail.ustc.edu.cn

***Abstract*—This paper introduces a novel multi-stage speech enhancement (SE) algorithm that leverages cascaded signal-to-noise ratio (SNR) domain shifts to improve enhancement performance. The proposed method decomposes the SE task into multiple stages, where each stage's output SNR is optimized for the next stage's input. By training models on distinct SNR domains and using post-processing fusion, the approach effectively elevates low SNR inputs to higher, more optimal SNR domains. Experiments on CHiME-4 and DNS2020 datasets demonstrate significant improvements in objective intelligibility and perceptual quality, showcasing the effectiveness of SNR domain shifts in multi-stage enhancement. This work provides new insights into designing efficient multi-stage SE models.**

I. Introduction

Speech enhancement (SE) [1] algorithms aim to remove background noise from noisy audio while minimizing any degradation to the naturalness of the clean speech. Due to limitations such as the model's representation capability and differences between simulation and real-world environment, current speech enhancement still faces significant challenges in practical applications [2]. Multi-stage enhancement has been proven effective in improving the performance of enhancement models compared to single-stage enhancement by many researchers.

Some multi-stage enhancement algorithms can be summarized as an enhancement-restoration paradigm. Lu et al. incorporated a spectrogram refinement network at the end of a standard speech enhancement model [3]. Hao et al. decomposed the enhancement task into a two-stage structure where the first stage removes time-frequency bins heavily contaminated by noise in the speech spectrum, and the second stage repairs the spectrum holes left after removal of the first stage [4]. The design of these multi-stage enhancements benefits from the increase in model size and joint training with multiple objectives, achieving certain performance improvements [3]–[6]. However, these methods lack an analysis of the effectiveness of the multi-stage strategies designed, which weakens the credibility of the generalization of the proposed methods.

Another part of multi-stage enhancement strategies involves breaking down the enhancement task into a step-by-step approximation process based on certain rules. Progressive Learning (PL) methods transform the direct mapping from noisy speech to clean speech into multiple stages. The signal-to-noise ratio (SNR) increases progressively by guiding hidden layers to explicitly learn intermediate targets [7]–[9]. As the internal structure of the enhancement model becomes more complex, the transformation for PL framework becomes increasingly difficult, requiring fine-tuning of the SNR improvement targets for the internal hidden layers. Unlike PL frameworks, diffusion models define the noise addition process as a step-by-step diffusion transformation of clean input data into an isotropic Gaussian distribution [10], [11]. The denoising process is the reverse, where the model gradually restores the clean input data by predicting and removing the noise introduced at each step of the diffusion process [12]. The current obstacle of diffusion models lies in the slow inference speed due to the need for multiple reverse diffusion steps [13].

Inspired by the array of existing multi-stage speech enhancement methods, we propose a new principle-based foundation and analytical framework. We use speech pairs containing clean speech and noise sources to simulate a set of same-source but progressively increasing SNR test groups. By analyzing the mean gain of metrics such as perceptual evaluation of speech quality (PESQ) [14] and short-time objective intelligibility (STOI) [15] before and after enhancement, it is concluded that enhancement models have a preference for input SNR when improving objective perceptual quality. This characteristic provides a theoretical basis and design guidance for the development of multi-stage enhancement models.

Besides, based on the identified pattern of speech enhancement when facing inputs with different SNR levels mentioned above, we preliminarily propose a general paradigm for multi-stage enhancement. A multi-stage speech enhancement framework with cascaded SNR domain shifts is introduced, leveraging the tendency of enhancement models towards intermediate SNR input domains. Firstly, we propose using post-processing fusion to gradually increase the input speech SNR between sub-enhancement models in multi-stage enhancement. Through comparative experiments, we have verified that post-processing fusion can to some extent achieve the migration of the input SNR domain. Secondly, based on the characteristics of input SNR domain shifts, we adjust the training SNR domains (TSD) of the sub-enhancement models. Specifically, the training SNR domains of the sub-enhancement models are designed to have a sliding upward relationship. And we expand the union of the TSD of the sub-enhancement models to cover the shifted

†Corresponding author

979-8-3315-7206-8/25/$31.00

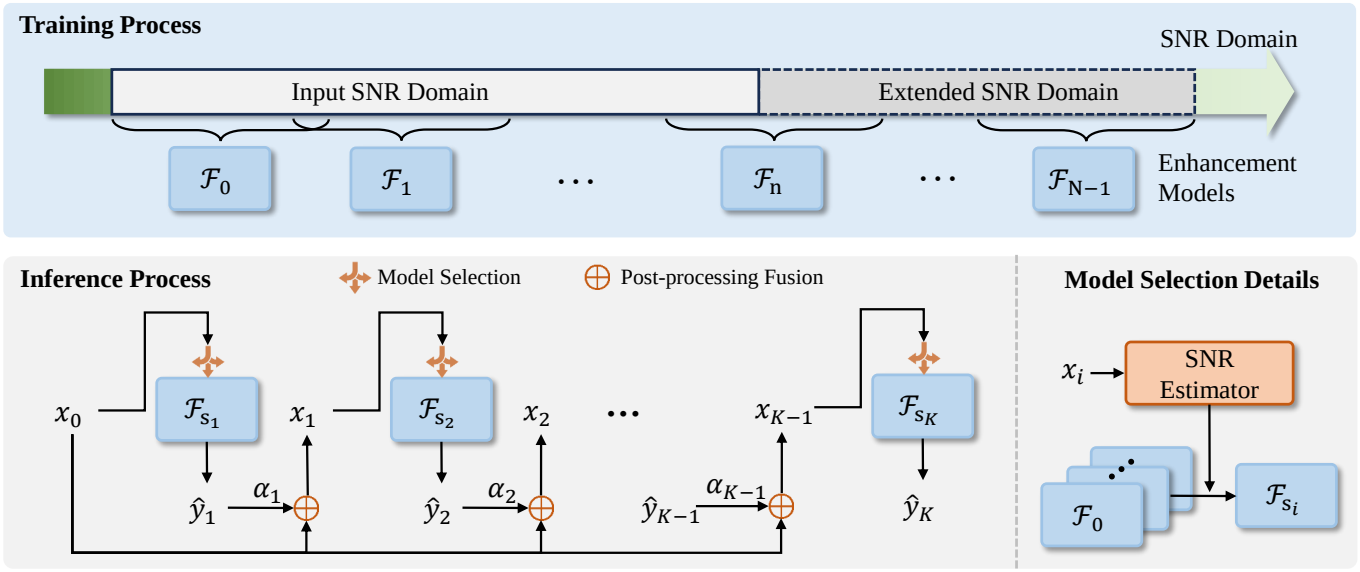

Fig. 1. The framework of the proposed multi-stage speech enhancement with cascaded SNR domain shifts algorithm

input SNR domain. Then, we segment the expanded training SNR domain to obtain N narrow TSD enhancement models to improve the performance of the enhancement models at corresponding SNRs in the cascaded enhancement process. We have validated the effectiveness of our method on the CHiME-4 dataset and demonstrated that our approach does not rely on an increase in model parameters. We also demonstrate the advantages of our proposed method for lightweight speech enhancement on the DNS-2020 synthetic dataset, which highlights the generalizability of our proposed method.

II. Proposed Methods

A. Problem Description

Given the noisy input audio, speech enhancement aims to extract the target speech. Consider the $\mathbf{x}$ is a time-domain noisy speech corrupted by additive noise, which can be defined as:

$$\mathbf{x} = \mathbf{s} + \mathbf{n} \tag{1}$$

where $\mathbf{s}$ is the target speech, $\mathbf{n}$ is the background noise [16]. Since the signals involved in this paper are all time-domain signals, the time-domain indices t are omitted for simplification.

B. Post-processing Fusion and Input SNR Domain Shifts

Given that enhancement models tend to perform better on inputs with intermediate SNRs, it would be advantageous to shift input signals with initially low SNRs to relatively higher SNR domains. Post-processing fusion is proposed to achieve this.

$$\mathbf{x}' = \alpha\hat{\mathbf{y}} + (1-\alpha)\mathbf{x} \tag{2}$$

$\hat{\mathbf{y}}$ is the clean speech prediction from SE models, and $\mathbf{x}$ is the original noisy input. By using a scalar α to adjust the weights, $\mathbf{x}'$ can be viewed as a new input at a shifted SNR domain.

C. Multi-stage Speech Enhancement with Cascaded SNR Domain Shifts (CSDS)

We further propose a general multi-stage speech enhancement with cascaded SNR domain shifts algorithm based on the post-processing fusion operation. Fig. 1 illustrates the training and inference processes.

During the training phase, we expand and segment the training SNR domain to obtain N sets of progressively shifted TSD partitioning schemes. Subsequently, we obtained N groups of simulated datasets from the same noise and clean speech pairs, with a gradually increasing SNR domain. Then we train N corresponding SE models based on these partitions individually. The CSDS algorithm shifts the input SNR domain through post-processing fusion. Therefore, the TSD during the training process is the union of the input SNR domain and the shifted input SNR domain, unlike conventional training simulations where the TSD is the same as the input SNR domain. Although it is necessary to expand the training SNR domain from the perspective of the input SNR domain shifts, Hao et al. observed that expanding the training SNR domain can lead to a decline in the performance of SE models [17]. We further divide the expanded TSD into N segments to avoid the negative impact of TSD expansion. Neighboring CSDS maintain a certain degree of overlap to mitigate the issue of decreased enhancement performance at the edges of the TSD.

During the inference phase, the CSDS strategy shifts the SNR domain through $K-1$ repeats of cascaded enhancement and post-processing fusion, followed by a final enhancement to achieve a clean prediction. Additionally, model selection ensures that the chosen model in the cascaded inference process is optimal for the given SNR input. The enhancement and post-processing fusion process at the i_{th} order is as follows:

$$\hat{\mathbf{y}}_i = \mathcal{F}_{s_i}(\mathbf{x}_{i-1}), \quad i = 1, 2, \ldots, K \tag{3}$$

$$\hat{\mathbf{x}}_i = \alpha_i \hat{\mathbf{y}}_i + (1-\alpha_i)\mathbf{x}_0, \quad i = 1, 2, \ldots, K-1 \tag{4}$$

$\mathbf{x}_i$, $\hat{\mathbf{y}}_i$ is the input and output of the i_{th} order enhancement. In the post-processing fusion process, the prior enhancement model output $\hat{\mathbf{y}}_i$ is mixed with the original noisy input $\mathbf{x}_0$. The scalar group $\{\alpha_1, \alpha_2, ..., \alpha_{K-1}\}$ controls the process of the input SNR domain shifts. The model selection process requires an SNR estimator to get the SNR of the input for each order of enhancement. Based on this estimate, the appropriate enhancement model $\mathcal{F}_{s_i}$ is identified, corresponding to the TSD whose center is closest to the estimated SNR.

In addition to the general CSDS algorithm presented above, there are special cases where the model selection step can be omitted. When $K = 2$, $N = 1$, only one enhancement model is trained and its TSD covers both the input SNR range and the expanded SNR range. In this scenario, the model selection step and the SNR estimation module shown in Fig. 1 are not required. When $K = 2$, $N = 2$, the TSD of one enhancement model can be set to the input SNR domain, while the TSD of the other model is set to the shifted SNR domain corresponding to the input SNR domain. The advantage of this approach is that during the inference process, both the original input $\mathbf{x}_0$ and $\mathbf{x}_1$ after post-processing fusion have fixed optimal enhancement models. Therefore, both model selection and SNR estimation can be omitted.

D. Loss Function

The loss function of model training employ a combination of time-domain mean absolute error (MAE) and frequency-domain multi-resolution STFT (MR-STFT) loss [18]:

$$\mathcal{L}_{\text{mae}} = \|\mathbf{y} - \hat{\mathbf{y}}\|_1 \tag{5}$$

$$\mathcal{L}_{\text{mr-stft}} = \sum_{i=1}^{m} \left(\frac{\|S_i(\mathbf{y}) - S_i(\hat{\mathbf{y}})\|_F}{\|S_i(\mathbf{y})\|_F} + \frac{1}{N} \|log \frac{S_i(\mathbf{y})}{S_i(\hat{\mathbf{y}})}\|_1 \right) \tag{6}$$

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{\text{mae}} + \lambda \mathcal{L}_{\text{mr-stft}} \tag{7}$$

$\mathbf{y}$ and $\hat{\mathbf{y}}$ is the target clean speech and enhanced speech. $\|\|_F$ and $\|\|_1$ represent the Frobenius and L_1 norms. $S_i(\mathbf{y})$ and N represent the magnitude of the linear-scale spectrogram of $\mathbf{y}$ with different STFT hyperparameters and number of elements in the magnitude. The parameter λ is used to balance the weights of different loss components.

III. Experiments

A. Datasets

We use two datasets to evaluate our proposed strategy. Both the datasets are sampled at 16kHz. The first one is the CHiME-4 dataset [19], [20]. Clean speech derived from the WSJ0 corpus is corrupted with CHiME-4 background noise to build a 30-hour training dataset. Simulated test sets are used to evaluate the effectiveness of our methods.

The second dataset is the DNS-2020 challenge database [21], [22]. It features more than 500 hours of clean speech and over 180 hours of noise. We generated 500 hours of clean and noisy speech pairs as the training set. For objective evaluation, we use the official synthetic test set without reverb.

TABLE I
Ablation study of CSDS algorithm on CHiME-4 synthetic test sets.

	K	N	PESQ	STOI(%)
Unprocessed	-	-	1.98	82.12
CRN	-	-	2.78	91.82
CRN-CSDS	2	1	2.86	92.42
CRN-CSDS	2	2	2.92	92.64
CRN-CSDS	3	2	2.91	92.60
CRN-CSDS	2	3	2.96	92.67

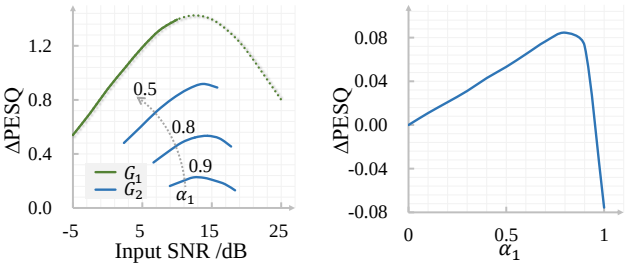

Fig. 2. Demonstration of input SNR domain shifts.

B. Implementation Details

CRN [23] is employed in ablation experiments conducted on the CHiME-4 dataset. Further comparison is conducted based on DeepFilterNet [24] at the DNS2020 dataset with other SOTA results. The CRN and DeepFilterNet configurations are the same as in the original papers. As for the SNR estimator in the model selection process, a similar approach to that proposed in DCCRN+ [25] is adopted, replacing the decoder of the DCCRN model with a single-layer LSTM, 1D convolution, and sigmoid to estimate frame-level SNR, ultimately averaging the results to predict the SNR of the input speech. The MR-STFT loss uses multi-resolution parameters with hop sizes of {50, 120, 240}, window lengths of {240, 600, 1200}, and FFT bins of {512, 1024, 2048}. Joint loss weights λ are set to 0.5. As for the training process, the learning rate is initialized as 0.001, and the Adam optimizer is used [26]. The overlap of neighboring TSDs mentioned in subsection II-C is set to 2dB.

C. Results and Analysis

Ablation results of the CSDS strategy: We conducted ablation experiments based on CRN on the CHiME-4 dataset. Table I presents the ablation results of the CSDS strategy. "Unprocessed" refers to the perceptual scores of input noisy speech. "CRN" represents the results of the baseline CRN model. "-CSDS" indicates using the CSDS strategy. "K" and "N" are the order and stage of the CSDS strategy. The improvement in objective metrics brought by speech enhancement is measured by PESQ and STOI, which respectively reflect the perceptual quality and intelligibility of speech.

It can be observed that compared to the baseline "CRN" model, the use of the CSDS strategy can effectively improve the SE performance. The improvement of "CRN-CSDS" with $K = 2, N = 1$ compared to "CRN" demonstrates the effect

of input SNR domain shifts, and the improvement of "CRN-CSDS" with $K = 2, N = 2$ compared to "CRN-CSDS" with $K = 2, N = 1$ reflects the benefits of TSD segmentation. When K exceeds 2, the performance of the CSDS strategy begins to decline, indicating that the effectiveness of the CSDS strategy lies in shifting the input SNR to an appropriate domain through post-processing fusion, rather than increasing computational complexity through cascading inference. When N exceeds 2, the performance improvement starts to diminish. This is because the benefits of TSD segmentation mainly lie in mitigating the interference caused by the coexistence of high SNR and low SNR training data on the enhancement model, and the marginal benefits of adding segmentation are limited.

SNR domain shifts: We investigated the role and mechanism of SNR domain shifts in the CSDS algorithm, starting from a special case of second-order repetition enhancement with $K = 2$ and $N = 1$. In this case, there are only two enhancement steps in total. The noisy input $\mathbf{x}_0$ is enhanced for the first time to obtain the prediction $\hat{\mathbf{y}}_1$. Then, post-processing fusion is performed with α_1 as the weighting coefficient, followed by another enhancement to obtain the final prediction output $\hat{\mathbf{y}}_2$. We first trained a CRN enhancement model, with the TSD settings configured as discussed in Section II-C, set to the union of the input SNR range (-5 to 10 dB) and the approximate shifted SNR range (0 to 20dB). Then, we utilized the simulation test set to obtain seven test sets with SNR levels increasing evenly from -5 dB to 10 dB, with each test set synthesized from the same source. This allowed us to evaluate the enhancement performance of the first and second enhancement stages, as well as determine the average optimal selection for the average alpha value considering different SNR inputs. The experimental results are shown in Fig. 2.

Assume $G(\mathbf{x}_1, \mathbf{x}_0)$ represents the ΔPESQ of noisy audio $\mathbf{x}_1$ over $\mathbf{x}_0$ corresponding to the same clean target. G_1 and G_2 in the left subplot of Fig. 2 are abbreviations of $G(\hat{\mathbf{y}}_1, \mathbf{x}_0)$ and $G(\hat{\mathbf{y}}_2, \mathbf{x}_1)$, which represents the PESQ gain of the first-order and second-order enhancement model. We have additionally included a test set for the first-order enhancement input with SNR levels ranging from 10 dB to 25 dB to demonstrate the enhancement model's performance on inputs with varying SNR levels, which we display with a green dashed line. Since this test set falls outside the SNR range used for training the enhancement model, we do not display the corresponding part in G_2. ΔPESQ in the right subplot of Fig. 2 is the average PESQ gain $G(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_0)$, which can be regarded as the benefit brought by adding the second-order enhancement.

G_1 shows an inverted U-shape, which demonstrates the enhancement model's preference for inputs with intermediate SNR levels. The performance measured by STOI shares a similar trend. The curve of G_2 follows the same trend as G_1, indicating that the enhancement response to the second-order input is similar to the original noisy speech input. As α_1 decreases, G_2 approaches G_1. This is because a smaller α_1 means a higher proportion of the original noisy signal in the second-order enhancement input, making the performance

TABLE II
PERFORMANCE ON DNS2020 NO-REVERB TEST SETS.

	#par.(M)	MACS(G/s)	PESQ	STOI(%)
Unprocessed	-	-	2.16	91.52
FullSubNet	5.60	-	3.31	96.11
DCCRN	3.70	14.36	3.27	-
GaGNet	5.94	1.63	3.56	97.13
CleanUNet	46.07	-	3.55	97.70
DeepFilterNet	1.80	0.35	3.49	97.35
-CSDS (K=2,N=1)	1.80	0.70	3.54	97.52
-CSDS (K=2,N=2)	3.60	0.70	3.66	97.83

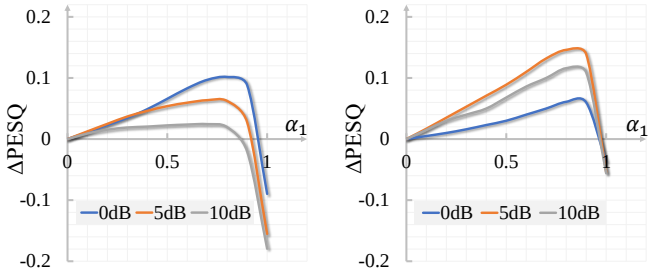

Fig. 3. Demonstration of TSD expansion.

of G_2 closer to that of G_1. Post-processing fusion should balance the improvement in the input SNR domain and the additional enhancement distortion caused by the fusion. The optimal α_1 is around 0.8 as shown in the right subplot of Fig. 2. Due to the enhancement model's preference for intermediate SNR, the higher the α_1, the closer the low SNR test data gets to the intermediate SNR domain. However, when α_1 approaches 1, the distortion components of the signal become more significant compared to the original signal components, resulting in a decrease in signal quality after the second enhancement. Fig. 2 demonstrated that post-processing fusion can achieve the shifts of the input SNR domain, thereby allowing the subsequent enhancement model to achieve better noise reduction effects.

TSD expansion: Based on the SNR domain shift mechanism in CSDS, we illustrate the importance of TSD expansion and segmentation through comparative experiments as shown in Fig. 3. The test dataset is the CHiME-4 simulation test set with an SNR range of -5 to 10 dB. Two subplots correspond to the PESQ gain of the CSDS ($K = 2, N = 1$) system (without and with TSD expansion) relative to the CRN baseline with different SNR inputs, denoted as "0dB", "5dB" and "10dB".

The horizontal axis represents the parameter α_1 used in post-processing fusion after the first-order enhancement, while the vertical axis shows the PESQ gain of the CRN-CSDS system relative to the CRN baseline. The TSD of CSDS SE model in the left subplot equals the input SNR domain (-5 to 10dB). Due to SNR domain shifts, the second-order enhancement input obtained through post-processing fusion gradually exceeds the TSD of the enhancement model, leading to performance degradation. Therefore, it can be observed that the PESQ gain peak of the CSDS enhancement system in the

left figure gradually decreases as the input SNR increases. By expanding the TSD to cover the input SNR domain and the shifted SNR domain, as shown in the right figure, the performance degradation issue at second-order inference is significantly alleviated. However, the expansion of the TSD reduces the enhancement performance of the model in the input SNR domain. It can be observed that the peak of the curve corresponding to the "0dB" input in the right figure has decreased compared to the left subplot. This explains why we incorporated the segmentation of the TSD into the CSDS algorithm.

Comparison of the CSDS algorithm and other SOTA methods: We test our CSDS strategy based on Deep-FilterNet on DNS2020 datasets. Results presented in Table II are compared with FullSubNet [27], DCCRN [28], GaGNet [29] and CleanUNet [30]. It is evident that the CSDS strategy can significantly enhance the performance of DeepFilterNet, achieving notable improvements in objective metrics such as PESQ and STOI. This demonstrates the generalizability of our proposed CSDS strategy. Comparisons with other baselines highlight the effectiveness of our proposed CSDS algorithm.

IV. Conclusions

In this paper, we introduce a novel multi-stage enhancement algorithm. Drawing on the observation that enhancement models tend to favor inputs with intermediate SNR levels, we propose a method to achieve input SNR domain shifts through post-processing fusion and further improving enhancement performance by expanding and segmenting the TSD. Based on the DNS-2020 datasets, we have demonstrated the effectiveness of our proposed algorithm. Through ablation and comparative experiments conducted on the CHiME-4 dataset, we have verified the mechanism of input SNR domain shifts within the CSDS algorithm. These analyses may offer new ideas for the design of subsequent multi-level enhancement networks. For instance, future research can be conducted on how to reduce distortion during input SNR domain shifts and whether other multi-stage enhancement algorithms, such as diffusion models, follow similar principles in terms of input SNR domain shifts.

References

[1] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[2] C. O. Mawalim, S. Okada, and M. Unoki, "Are recent deep learning-based speech enhancement methods ready to confront real-world noisy environments?" In *Proc. Interspeech*, 2024, pp. 1735–1739.

[3] H. Lu, N. Li, T. Song, *et al.*, "Speech and noise dual-stream spectrogram refine network with speech distortion loss for robust speech recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[4] X. Hao, X. Su, S. Wen, *et al.*, "Masking and inpainting: A two-stage speech enhancement approach for low snr and non-stationary noise," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6959–6963.

[5] F. Dang, H. Chen, Q. Hu, P. Zhang, and Y. Yan, "First coarse, fine afterward: A lightweight two-stage complex approach for monaural speech enhancement," *Speech Communication*, vol. 146, pp. 32–44, 2023.

[6] H. Phan, I. V. McLoughlin, L. Pham, *et al.*, "Improving gans for speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 1700–1704, 2020.

[7] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Snr-based progressive learning of deep neural network for speech enhancement.," in *Proc. Interspeech*, 2016, pp. 3713–3717.

[8] A. Li, M. Yuan, C. Zheng, and X. Li, "Speech enhancement using progressive learning-based convolutional recurrent neural network," *Applied Acoustics*, vol. 166, p. 107 347, 2020, ISSN: 0003-682X. DOI: https://doi.org/10.1016/j.apacoust.2020.107347.

[9] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Densely connected progressive learning for lstm-based speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5054–5058.

[10] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2021, pp. 659–666.

[11] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7402–7406.

[12] D. Kim, D.-H. Yang, D. Kim, *et al.*, "Guided conditioning with predictive network on score-based diffusion model for speech enhancement," in *Proc. Interspeech*, 2024, pp. 1190–1194.

[13] T. Trachu, C. Piansaddhayanon, and E. Chuangsuwanich, "Thunder: Unified regression-diffusion speech enhancement with a single reverse step using brownian bridge," in *Proc. Interspeech*, 2024, pp. 1180–1184.

[14] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, IEEE, vol. 2, 2001, pp. 749–752.

[15] 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.

[16] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[17] X. Hao, X. Su, Z. Wang, Q. Zhang, H. Xu, and G. Gao, "Snr-based teachers-student technique for speech enhancement," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2020, pp. 1–6.

[18] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6199–6203.

[19] E. Vincent, S. Watanabe, J. Barker, and R. Marxer, "The 4th chime speech separation and recognition challenge," *URL: http://spandh. dcs. shef. ac. uk/chime_challenge/(last accessed on 1 August, 2018)*, 2016.

[20] J. Du, Y.-H. Tu, L. Sun, *et al.*, "The ustc-iflytek system for chime-4 challenge," *Proc. CHiME*, vol. 4, no. 1, pp. 36–38, 2016.

[21] C. K. Reddy, V. Gopal, R. Cutler, *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *arXiv preprint arXiv:2005.13981*, 2020.

[22] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Interspeech 2020 deep noise suppression challenge: A fully convolutional recurrent network (fcrn) for joint dereverberation and denoising.," in *Proc. Interspeech*, 2020, pp. 2467–2471.

[23] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement.," in *Proc. Interspeech*, vol. 2018, 2018, pp. 3229–3233.

[24] H. Schroter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, "Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7407–7411.

[25] L. Shubo, H. Yanxin, Z. Shimin, and X. Lei, "Dccrn+: Channel-wise subband dccrn with snr estimation for speech enhancement," in *Proc. Interspeech*, 2021, pp. 2816–2820.

[26] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] 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 *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6633–6637.

[28] 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.

[29] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Applied Acoustics*, vol. 187, p. 108 499, 2022, ISSN: 0003-682X. DOI: https://doi.org/10.1016/j.apacoust.2021.108499.

[30] Z. Kong, W. Ping, A. Dantrey, and B. Catanzaro, "Speech denoising in the waveform domain with self-attention," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7867–7871.