ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096985

CROSS-DOMAIN DIFFUSION BASED SPEECH ENHANCEMENT FOR VERY NOISY SPEECH

Heming Wang[1] *and DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA
wang.11401@osu.edu, dwang@cse.ohio-state.edu

ABSTRACT

Deep learning based speech enhancement has achieved remarkable success, but challenges remain in low signal-to-noise ratio (SNR) nonstationary noise scenarios. In this study, we propose to incorporate diffusion-based learning into an enhancement model and improve robustness in extremely noisy conditions. Specifically, a frequency-domain diffusion-based generative module is employed, and it accepts the enhanced signal obtained from a time-domain supervised enhancement module as an auxiliary input to learn to recover clean speech spectrograms. Experimental results on the TIMIT dataset demonstrate the advantage of this approach and show better enhancement performance over other strong baselines in both -5 and -10 dB SNR noisy conditions.

Index Terms— speech enhancement, diffusion model, generative model, low signal-to-noise ratio

1. INTRODUCTION

Speech signals in the real world are usually corrupted by background noise, which degrades speech quality and intelligibility. Speech enhancement aims to suppress noise interference in such environments. Conventional approaches to tackle this problem include signal processing methods, like spectral subtraction [1], and computational auditory scene analysis [2]. In recent years, major advances have been made in speech enhancement thanks to the introduction of deep neural networks (DNNs). Early DNN studies use spectral magnitude features as training targets, and train DNNs to predict the ideal binary mask [3], ideal ratio mask [4], and target magnitude spectrum [5, 6]. Recent research addresses additionally phase enhancement and attempts to recover clean speech in either the complex domain [7] or the time domain [8]. It is well established that DNN-based supervised enhancement performs well in less noisy conditions. However, it remains challenging to recover clean speech in very low signal-to-noise ratios (SNRs) conditions, where enhancement models suffer from a significant performance drop.

In this paper, we address this challenge by proposing a joint learning framework that performs cross-domain speech enhancement. In order to improve robustness for extremely noisy speech, we propose to enhance speech in a coarse-to-fine manner. Noisy speech is first enhanced by supervised learning, and we then mask the time-frequency (T-F) units in the complex spectrogram that are highly noisy and have poor estimation. Afterwards we use a diffusion model to regenerate the masked regions. Specifically, we first adopt a time-domain network to provide the coarsely enhanced speech as auxiliary information, and then mask the enhanced speech and feed it to a complex-domain diffusion-based generative model.

A related study by Hao et al. [9] proposed mask and inpainting (M&I), a two-stage method that specifically masks the noisy part of a degraded speech spectrogram, and uses an inpainting network to reconstruct the magnitude spectrogram. Different from this method, our proposed method performs end-to-end cross-domain enhancement, and addresses both magnitude and phase estimation. Furthermore, we combine generative learning and enhancement learning and gradually refine noisy speech, whereas M&I only performs magnitude inpainting on the masked noisy spectrogram. Lastly, we introduce a more powerful generation module that demonstrates better performance than the adversarial network used in M&I.

We employ a denoising diffusion probabilistic model (DDPMs) [10, 11, 12, 13], which was originally introduced for audio and image generation and demonstrated to be effective for generating high-quality samples. Diffusion-based generative models for audio generation have received considerable attention recently [14, 15, 16]. The core idea of DDPM is to use a DNN to approximate a diffusion process and progressively generate samples from a normal distribution. Its theoretical details are described in Section 2.

2. BACKGROUND

Fig. 1(a) illustrates the overall process of DDPM. It contains two processes: forward diffusion and reverse diffusion. During forward diffusion, it gradually converts the given data into a normal distribution and then learns the reverse diffusion process to generate data in the target domain from a whitened Gaussian noise. The forward diffusion process is defined as the fixed Markov chain from the target data x_0 to the latent variable x_T (a normal distribution) in T time steps, formulated as,

$$q(x_1, ..., x_T|x_0) = \prod q(x_t|x_{t-1}). \tag{1}$$

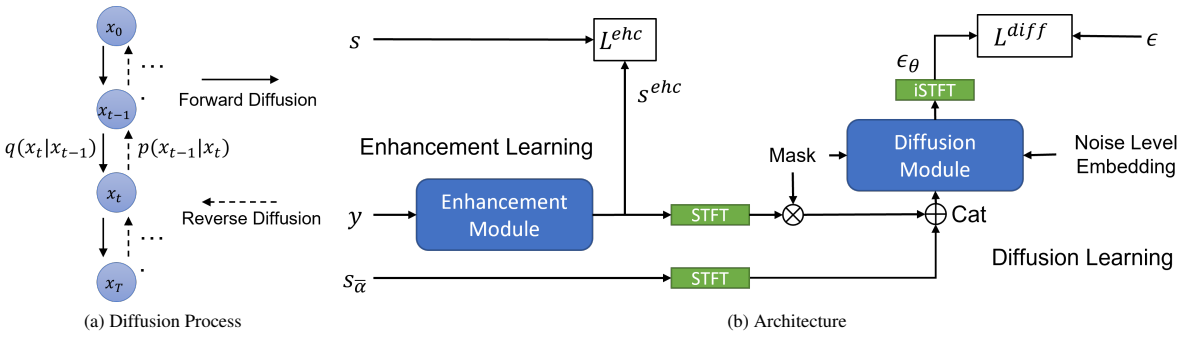
At each time step $t \in [0, T]$, a small Gaussian noise is added to x_{t-1} to obtain x_t, i.e. $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$, and the process is parameterized by the noise schedule β_t. The reverse diffusion process is also a Markov chain starts from an isotropic Gaussian x_T, and a model that is parameterized by θ is employed to learn the added Gaussian noise ϵ, defined as,

$$p_\theta(x_0|x_1, ..., x_T) = \prod p_\theta(x_{t-1}|x_t), \tag{2}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)^2 I). \tag{3}$$

In the transition probability $p_\theta(x_{t-1}|x_t)$, $\mu_\theta(x_t, t)$ and $\sigma_\theta(x_t, t)$ are the model estimated mean and variance of x_{t-1}. The goal of $p_\theta(x_{t-1}|x_t)$ is to eliminate the added Gaussian noise during the

This research was supported in part by an NIDCD (R01 DC012048) grant and the Ohio Supercomputer Center.



(a) Diffusion Process (b) Architecture

Fig. 1. (a). Diffusion process (b). The overall architecture of the proposed method.

diffusion process. Ho et al. [10] proposed to train the model by maximizing its variational low bound, and they show that it can be approximated by certain re-parameterization. They reported that optimizing the approximated objective function leads to high generation quality. With $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, the diffusion training can defined as minimizing the following objective,

$$\mathbb{E}_{t,x_0,\epsilon}[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t||_2^2]. \tag{4}$$

By optimizing Eq. 4, we can derive

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)), \tag{5}$$

$$\sigma_\theta(x_t, t) = \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t}, \tag{6}$$

and then progressively generate x_0 from x_T.

3. METHOD

We perform speech enhancement in extremely noisy conditions. Specifically, we focus on monaural speech enhancement where we have a noisy mixture y collected from a single microphone, which consists of background noise n and clean speech s. The proposed model f is used to generate an estimate $\hat{s}$ to recover the target clean speech s. It accepts as input the noisy speech and a mask M. The mask M indicates the T-F units that contain effective speech information and guides the model to regenerate the rest of the T-F units that are dominated by noises. During training, the mask is manually generated to improve the generalization capability of the diffusion model. Details about the mask generation are further described in Section 3.3. With the parameters of the model denoted as θ, the training process can be formulated as,

$$s^{ehc}, \epsilon_\theta = f(\theta, y, s_{\bar{\alpha}}, M), \tag{7}$$

where s^{ehc} and ϵ_θ are the output of the enhancement module and the diffusion module in f. For the diffusion training, as suggested by Chen et al. [17], we use a continuous noise level $\bar{\alpha}$ to generate the noise level embedding, which is sampled within adjacent discrete noise levels $\sqrt{\bar{\alpha}_{t-1}}$ and $\sqrt{\bar{\alpha}_t}$. The diffused signal $s_{\bar{\alpha}} = \sqrt{\bar{\alpha}}s + \sqrt{1-\bar{\alpha}}\epsilon$ is employed as input of the diffusion module. During inference, we first use the reverse diffusion process to derive s^{diff}, and convert outputs from both modules to the complex domain by applying short-time Fourier transforms (STFTs). The final estimation is calculated by combining the two outputs with the mask M, i.e.,

$$\hat{S} = S^{ehc}M + S^{diff}(1 - M), \tag{8}$$

where $\hat{S}$, S^{ehc} and S^{diff} are the corresponding STFTs of $\hat{s}$, s^{ehc} and s^{diff}. The estimated clean speech $\hat{s}$ is obtained by converting $\hat{S}$ back to the time domain using inverse STFT (iSTFT).

3.1. Proposed Architecture

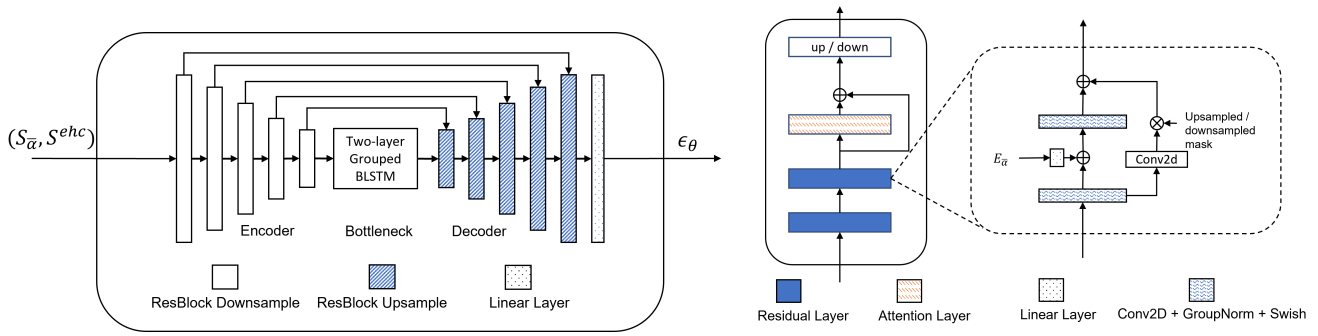
As illustrated in Fig. 1(b), our network consists of two modules. The first module is a time-domain enhancement module that performs supervised learning and provides auxiliary information for the second module. The second module is a diffusion-based generative module that operates in the complex domain and is fed with a diffused noisy speech. It also accepts the masked enhanced speech, a mask, and a noise level embedding as additional input, and learns to estimate added noise of the diffused signal.

3.1.1. Enhancement Module

We adopt the dual-path attention recurrent network (DPARN) [8] by Pandey and Wang as the enhancement module, which is an improved version of the popular dual-path recurrent neural network. In a dual-path network, the time series of the given utterances is divided into overlapping chunks, and then sequentially processed by intra-chunk and inter-chunk RNNs. This technique considerably reduces the sequence length for each RNN computation and improves the training efficiency. It also allows a relatively small frame shift for time-domain speech processing, which leads to a significant performance improvement. DPARN further incorporates inter-chunk and intra-chunk attention to improve the enhancement performance. Note that in order to reduce the computational burden, we adopt residual connections instead of dense connections between RNN modules and only use two DPARN blocks.

3.1.2. Diffusion Module

The diffusion module accepts four inputs, the diffused signal $s_{\bar{\alpha}}$, the enhanced speech s^{ehc} obtained from the enhancement module, a mask M, and the noise level embedding $E_{\bar{\alpha}}$. The noise level embedding is generated using the sinusoidal positional embedding [18] followed by two linear layers. We concatenate $S_{\bar{\alpha}}$ with S^{ehc} along the channel dimension before feeding to the diffusion module. We



(a) Diffusion Module (b) Residual Block

Fig. 2. Diagrams showing the design of the diffusion module. (a). The overall architecture of CRN. (b). The Residual Block that is used for incorporating conditioner information.

implement the diffusion module by adopting a convolutional recurrent neural network (CRN) that is based on [19]. CRN is a complex-domain encoder-decoder based architecture, and we employ a recurrent neural network bottleneck to model the temporal dependencies, which also allows us to process input of variable lengths. The technical details are depicted in Figure 2(a). The encoder of CRN is a convolutional downsampler that reduces the feature dimension along the frequency axis, and the decoder has a symmetric design that performs upsampling with transposed convolutions. The output of encoders is concatenated to the corresponding layers of the decoder for better reconstruction performance. To save computational cost and memory, we use grouped bidirectional long short-term memory (BLSTM) [20] as the bottleneck. Finally, the output of the CNN decoder passes through a linear layer to produce real and imaginary estimates. The modification we introduce to the CRN is to replace each convolution operation with a residual block that accepts noise level embedding and mask as local conditioners. The design of the residual block is based on [21], and we illustrate the detailed design in Fig. 2(b). It consists of two residual layers, an attention layer, and upsampling/downsampling operation. The noise level embedding $E_{\bar{\alpha}}$ passes through a linear layer and is then added as a bias term after the convolution layer, and the upsampled/downsampled mask is incorporated by multiplying the input of each residual layer after a pointwise convolution.

During inference, to accelerate the inference speed for the diffusion process, we use 8 iterations with a linear noise schedule adopted from [22]. In addition, instead of starting with the Gaussian white noise, we start the reverse diffusion process with the output of the enhancement module s^{ehc}. This approach is also referred to as the shallow diffusion mechanism proposed in [23]. Converting noisy enhanced speech is much easier compared with converting pure white noise into clean speech, which lightens the burden of the diffusion module and can also accelerate inference. We find in experiments it provides better enhancement objective scores.

3.2. Loss Function

We perform joint training to optimize the two modules simultaneously. The diffusion loss follows the objective proposed in [22], where a logarithm $\mathcal{L}1$ loss is computed over the estimated noise ϵ_θ and the real added noise ϵ in the time domain, i.e.,

$$\mathcal{L}^{diff} = \mathbb{E}_{\bar{\alpha}, S_{\bar{\alpha}}, S^{ehc}, \epsilon}[\log ||\epsilon - \epsilon_\theta(S_{\bar{\alpha}}, S^{ehc}, M, \sqrt{\bar{\alpha}})||], \quad (9)$$

where $S_{\bar{\alpha}}$ is the STFT of the diffused noisy signal $s_{\bar{\alpha}}$, and $\sqrt{\bar{\alpha}}$ indicates the continuous noise level. Using logarithm $\mathcal{L}1$ norm stabilizes the training as it provides a smooth scaling for different time steps.

The enhancement module aims to restore the clean speech, and we use a complex-domain loss $\mathcal{L}^{ehc}$ [24], which addresses the importance of the magnitude estimation and has shown superior enhancement performance.

$$\mathcal{L}^{ehc} = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}[||S^{ehc}(t,f)| - |S(t,f)||+$$
$$(|S_r^{ehc}(t,f) - S_r(t,f)| + |S_i^{ehc}(t,f) - S_i(t,f)|)]. \quad (10)$$

T and F denote the total number of time frames and frequency bins, which are indexed by t and f, respectively. S^{ehc} and S are the short-time Fourier transform of s^{ehc} and s. Subscripts r and i denote the real and imaginary parts of the complex vectors, respectively, and $|\cdot|$ measures the magnitude. For computing $\mathcal{L}^{ehc}$, we divide waveforms into segments with a frame size of 512 samples with a frame shift of 128 samples and then multiply these frames with a Hanning window. The STFT vectors S^{ehc} and S are calculated on windowed frames to define related terms in the complex domain. The complete objective is defined as,

$$\mathcal{L} = \lambda\mathcal{L}^{ehc} + \mathcal{L}^{diff}. \quad (11)$$

We use the coefficient λ to balance the training progress of two modules. In this paper, we set $\lambda = 0.1$ based on the performance evaluated on the validation set.

3.3. Mask Generation

To improve the generation capability of the diffusion module, we employ a random mask generation strategy during the training process, which has been demonstrated to improve the robustness against unknown masks in [25]. During training, we uniformly sample masks from rectangles of arbitrary aspect ratios (box masks) or polygonal chains dilated by a high random width (irregular masks), or masks that only keeps T-F units that contain dominant speech information (segmentation masks). The mask generation algorithm is modified based on the open-sourced repository[1]. Inspired by

[1]https://github.com/saic-mdal/lama

Table 1. Enhancement performance of baselines and the proposed method on the TIMIT dataset at different SNRs.

	Factory						Babble					
	-10 dB			-5 dB			-10 dB			-5 dB		
Model	SISNR	STOI	PESQ	SISNR	STOI	PESQ	SISNR	STOI	PESQ	SISNR	STOI	PESQ
Noisy Mixture	-9.980	0.439	1.107	-4.980	0.539	1.276	-9.989	0.428	1.080	-4.992	0.532	1.360
M&I	-2.602	0.538	1.492	2.602	0.676	1.907	-6.497	0.442	1.178	-0.099	0.626	1.654
GCRN	0.440	0.576	1.440	5.571	0.751	1.940	-4.299	0.435	1.004	3.646	0.670	1.641
DCCRN	0.364	0.566	1.421	5.863	0.751	2.013	-3.802	0.437	1.010	3.339	0.668	1.645
DPARN	2.120	0.578	1.543	7.580	0.767	2.090	-2.696	0.446	1.259	5.328	0.708	1.699
Proposed	**2.504**	**0.611**	**1.648**	**7.909**	**0.789**	**2.204**	**-2.486**	**0.465**	**1.297**	**5.415**	**0.716**	**1.845**

M&I, a binary spectral magnitude mask (BSMM) with a threshold of $\tau = 0.15$ is employed as our segmentation mask, which is computed over noisy and clean T-F units,

$$BSMM(t,f) = \begin{cases} 1 & \text{if } \frac{|S(t,f)|}{|Y(t,f)|} \geq \tau, \\ 0 & otherwise, \end{cases} \tag{12}$$

where Y is the STFT of noisy input y. During inference, the estimated BSMM is utilized to remove T-F units dominated by background noise, and the mask is predicted by a CRN-Mask network proposed in [26]. The mask prediction network is pretrained on the same dataset with a mean-squared error loss using noisy speech mixtures as input.

4. EVALUATION AND ANALYSIS

4.1. Experimental Setup

We conduct experiments on the TIMIT dataset [27], which is a corpus containing utterances from 630 speakers with a 16 kHz sampling rate. We select 4620 utterances as the training dataset, and 1153 utterances as the validation set. The core test subset that contains 192 utterances from 24 speakers is used for evaluation. We use 10,000 noises from a sound effect library[2] to simulate noisy mixtures, which have a total duration of around 126 hours. To generate training mixtures, we randomly cut a segment from the training noises, and then mix it with a randomly picked clean utterance at an SNR level that is uniformly sampled from [-5, 5] dB. During testing, we mix the testing utterances with factory and babble noises extracted from NOISEX-92 [28] at two different SNR levels, -10 dB and -5 dB.

During training, we apply mean-variance normalization (MVN) to each noisy utterance and scale the corresponding clean utterances accordingly. A window length of 32 ms with 25% overlap between adjacent frames is used in calculating STFTs, which correspond to 257-dimensional complex spectrograms. An Adam optimizer [29] is employed, and we use a batch size of 32 utterances and an initial learning rate of 1e-3 to train the model for 100 epochs. The learning rate is scheduled to be halved if the validation loss has not improved for three consecutive epochs. A gradient clipping with a maximum value of 3.0 is applied to stabilize training. We randomly cut a 4-second segment for each training utterance, and pad shorter utterances with zeros within each batch to guarantee they are of the same size.

4.2. Evaluation Results

We compare the proposed method with four other advanced speech enhancement baselines using the same experimental settings described in Section 4.1. The M&I in the table indicates the two-stage enhancement approach that performs masking and inpainting [9] on the noisy magnitude spectrogram, which has been described in Section 1. The second baseline is the gated convolution recurrent network (GCRN) [19], which is a complex-domain enhancement network based on the CRN architecture. The major difference is that the original convolution operation is replaced with gated convolution, and two decoders are employed separately to predict real and imaginary parts. For a fair comparison, we use the bidirectional LSTM (BLSTM) in the bottleneck part. Deep Complex Convolutional Recurrent network (DCCRN) [7] is adopted as the second baseline. We choose the DCCRN-E configuration with a BLSTM in this comparison. Lastly the non-causal version of DPARN [8] is employed as the fourth baseline. We use three metrics to evaluate the enhancement performance, scale-invariant SNR (SISNR), short-time objective intelligibility (STOI) [30] and perceptual evaluation of speech quality (PESQ) [31]. For all the metrics, high values indicate better enhancement performance. We provide enhanced speech signals and spectrograms at https://whmrtm.github.io/CDDSE_demo.html.

Experimental results are displayed in Table 1, and it demonstrates that our proposed approach achieves better objective scores, especially in low-SNR conditions. Other strong enhancement baselines, although already demonstrated their effectiveness in suppressing noises, suffer from severe performance degradation at -10 dB SNR. For instance, for babble noise, GCRN and DCCRN barely improve PESQ over the noisy mixture. Although M&I does not reveal superior enhancement performance as it only addresses magnitude estimation, it shows noise robustness to some extent. At -10 dB noises, we observe better PESQ scores compared with GCRN and DCCRN. This phenomenon validates our assumption that regenerating noisy regions in the spectrograms could be beneficial. The proposed model achieves better enhancement results for both SNR conditions. Specifically, compared with the enhancement baseline of DPARN, at -10 dB factory noise, STOI is improved by 3.3%, and PESQ by 0.105, which demonstrates the benefit brought by the diffusion module. Also, compared with the related study M&I, STOI is improved by 7.3%, and PESQ by 0.156.

5. CONCLUSION

We have proposed a novel architecture for low-SNR speech enhancement. Specifically, we have designed a joint framework to combine enhancement learning and generative learning. First, we employ a time-domain DPARN to enhance noisy speech. Then in the complex domain, we perform diffusion-based spectrogram inpainting for T-F units that are dominated by background noise. Experimental results on the TIMIT dataset have demonstrated the effectiveness of this approach, yielding a significant improvement over the supervised enhancement baselines, especially in very low SNR conditions. In future work, we plan to conduct training on larger corpora and extend to general audio restoration.

[2] available at https://www.soundideas.com

6. REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[2] D. L. Wang and G. J. Brown, Eds, *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley-IEEE press, 2006.

[3] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, pp. 181–197. Springer, 2005.

[4] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1849–1858, 2014.

[5] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, pp. 65–68, 2014.

[6] K. Han, Y. Wang, D. L. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.

[7] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv:2008.00264*, 2020.

[8] A. Pandey and D. L. Wang, "Dual-path self-attention RNN for real-time speech enhancement," *arXiv:2010.12713*, 2020.

[9] X. Hao, X. Su, S. Wen, Z. Wang, Y. Pan, F. Bao, and W. Chen, "Masking and inpainting: A two-stage speech enhancement approach for low SNR and non-stationary noise," in *Proceedings of ICASSP*, 2020, pp. 6959–6963.

[10] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[11] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[12] C. Luo, "Understanding diffusion models: A unified perspective," *arXiv:2208.11970*, 2022.

[13] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *arXiv:2209.00796*, 2022.

[14] Z Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," *arXiv:2009.09761*, 2020.

[15] S.-G. Lee, H. Kim, C. Shin, X. Tan, C. Liu, Q. Meng, T. Qin, W. Chen, S. Yoon, and T.-Y. Liu, "PriorGrad: Improving conditional denoising diffusion models with data-driven adaptive prior," *arXiv:2106.06406*, 2021.

[16] J. Zhang, S. Jayasuriya, and V. Berisha, "Restoring degraded speech via a modified diffusion model," *arXiv:2104.11347*, 2021.

[17] N. Chen, Y. Zhang, H. Zen, R.J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," *arXiv:2009.00713*, 2020.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[19] K. Tan and D. L. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2020.

[20] F. Gao, L. Wu, L. Zhao, T. Qin, X. Cheng, and T.-Y. Liu, "Efficient sequence learning with group recurrent networks," in *Proceedings of NAACL-HLT*, 2018, pp. 799–808.

[21] V. Popov, I. Vovk, V. Gogoryan, and M. Sadekova, T.and Kudinov, "Grad-TTS: A diffusion probabilistic model for text-to-speech," in *Proceedings of ICML*, 2021, pp. 8599–8608.

[22] J. Lee and S. Han, "NU-Wave: A diffusion probabilistic model for neural audio upsampling," in *Proceedings of INTERSPEECH*, 2021, pp. 1634–1638.

[23] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "DiffSinger: Singing voice synthesis via shallow diffusion mechanism," in *Proceedings of AAAI*, 2022, vol. 36, pp. 11020–11028.

[24] Z.-Q. Wang, P. Wang, and D. L. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1778–1787, 2020.

[25] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with Fourier convolutions," in *Proceedings of WACF*, 2022, pp. 2149–2159.

[26] H. Wang and D. L. Wang, "Neural cascade architecture with triple-domain loss for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 734–743, 2021.

[27] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STIN*, vol. 93, pp. 27403, 1993.

[28] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, 1993.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 2125–2136, 2011.

[31] 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 *Proceedings of ICASSP*, 2001, vol. 2, pp. 749–752.