

MDDM: A Multi-view Discriminative Enhanced Diffusion-based Model for Speech Enhancement

Nan Xu, Zhaolong Huang, Xiaonan Zhi

Alibaba Digital Media Entertainment Group, Beijing, China

{xn430658, zhaolong.hzl, dexiao.zxn}@alibaba-inc.com

Abstract

With the development of deep learning, speech enhancement has been greatly optimized in terms of speech quality. Previous methods typically focus on the discriminative supervised learning or generative modeling, which tends to introduce speech distortions or high computational cost. In this paper, we propose **MDDM**, a **M**ulti-view **D**iscriminative **E**nhanced **D**iffusion-based **M**odel. Specifically, we take the features of three domains (time, frequency and noise) as inputs of a discriminative prediction network, generating the preliminary spectrogram. Then, the discriminative output can be converted to clean speech by several inference sampling steps. Due to the intersection of the distributions between discriminative output and clean target, the smaller sampling steps can achieve the competitive performance compared to other diffusion-based methods. Experiments conducted on a public dataset and a real-world dataset validate the effectiveness of MDDM, either on subjective or objective metric.

Index Terms: speech enhancement, multi-view, discriminative model, diffusion model, noise domain

1. Introduction

In the real world, clean speech signals are always disturbed by various environmental noises and reverberations, which seriously affects speech perceptual quality and intelligibility. Therefore, speech enhancement is thus an underlying task, which aims to recover clean speech signal from noisy speech. Traditional speech enhancement methods can use the statistical properties of the noisy and clean target signals in time-frequency or spatial domain [1]. With the success of deep learning, the enhancement performance has achieved some breakthroughs in the last decade. Deep learning based methods can be divided into two different categories: discriminative approaches and generative approaches.

Discriminative approaches are dominated by supervised learning algorithms that obtain clean target speech from noisy speech, always training with labeled samples. These methods typically take time domain, time-frequency (TF) domain or both as network input to learn a deterministic mapping. Specifically, the time or time-frequency domain methods take the single-view feature as input to predict waveform or magnitude-phase related training target [2–6]. Instead, some approaches integrate time domain and time-frequency domain in the speech enhancement framework, primarily eliminating non-stationary (e.g., impulse-like noise) and stationary noises simultaneously [7–9]. In addition, several methods follow the ideas of generative models for speech enhancement, such as variational autoencoder (VAE) [10–12] or Generative adversarial networks (GANs) [13, 14]. In contrast, diffusion-based speech enhance-

ment models have recently gained attention due to the superior enhancement effects [15–18]. Diffusion-based enhancement framework usually adds noise from the Wiener process to make clean target speech into a tractable prior (e.g., standard normal distribution), which is called the forward process. For the reverse process, a trained neural network is used to generate clean speech from the aforementioned prior distribution, such as SGMSE+ [19]. Besides, the appropriate condition guided generative process can achieve more superior and stable enhancement performance [20, 21].

Although the above approaches (discriminative and generative) have achieved significantly superior performance for speech enhancement, there are still some challenges:

- For discriminative methods, a deterministic map is learned during the training process. However, training data is a finite set and cannot cover all possible noise conditions to guarantee the generalizability capacity in unseen situations. Various noise types and levels can also result in distortions, especially for complex data distributions.
- Diffusion models are always trained to learn a standard normal distribution, which requires hundreds of inference steps leading to a heavy computational cost. Additionally, the defined standard white Gaussian is also not the case of environmental noise. Although Lemerrier et al. [20] utilizes a predictive model as a guidance, the predictive results that deviate from target distributions to a certain extent are used to compute score function, resulting in a suboptimal performance.

To address the aforementioned problems, we propose a multi-view discriminative enhanced diffusion-based model, named MDDM. In this work, a multi-view discriminative network is firstly used to predict an initial result that overlaps enough with a clean target distribution. Specifically in this discriminative network, the STFT-based U-Net framework is used as a backbone network. In addition, a parallel time U-Net and a noise modulation module are integrated into the backbone, which reduces speech distortions and improves noise perception abilities. To further improve the performance, we introduce a diffusion model using the discriminative intermediate feature with multi-view information as a conditioner. Furthermore, since there is an intersection distribution between target and discriminative result, they can achieve a nearly same noisy distribution by only several forward steps. Next, this noisy distribution can be converted to a more superior clean result by only several inference sampling steps, which accelerates inference speed. Assuming that discriminative results are reliable enough, the fewer sampling steps are required. Experiment results on two datasets show that our proposed method surpasses other baseline methods in terms of enhancement performance with the sampling step of only 30.

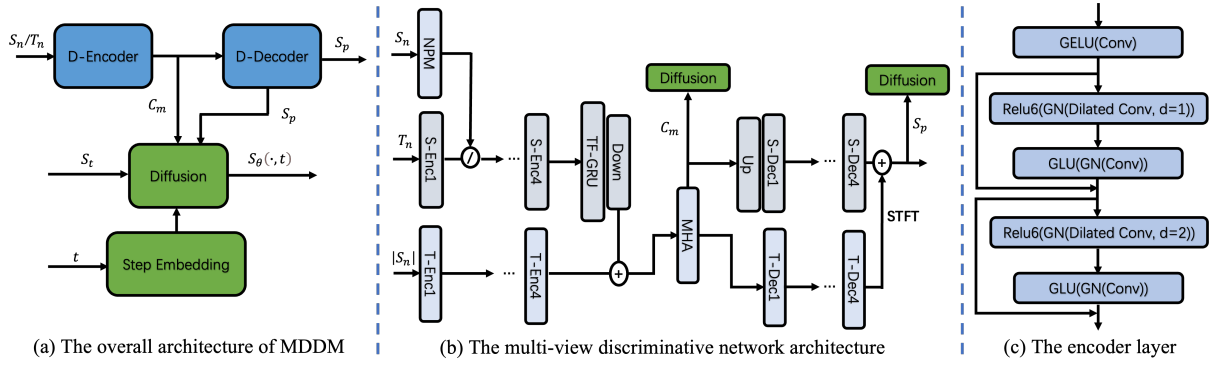


Figure 1: The overall architecture of MDDM, multi-view discriminative network and encoder layer. In subfigure (a), S_n and T_n are noisy signals in frequency and time domains, respectively. Two D-* are encoder and decoder of discriminative network and output spectrogram is S_p . S_t is the sampled spectrogram at time-step t . The intermediate multi-view feature C_m and output S_p are the conditions of diffusion model. In subfigure (b), S-* and T-* means frequency and time domains, respectively. NPM is the noise perception module and \odot means the modulation mode. Note that the skip connections are not drawn. In subfigure (c), the encoder module is shown. GELU is Gaussian Error Linear Unit and GN denotes GroupNorm. The dilation rates are 1 and 2, respectively.

2. Proposed Method

In this section, our proposed model will be introduced, as illustrated in Figure 1a. First, the discriminative network takes multi-view features as inputs to get the initial output results. Next, the output results and intermediate features with multi-view fusion information guide the diffusion model during training and inference processes. Also, the discriminative and diffusion tasks are jointly optimized by a multi-task learning scheme to improve speech enhancement performance.

2.1. Multi-view discriminative network

As illustrated in Figure 1b, the multi-view discriminative network contains three modules: STFT-based backbone network, waveform-base U-Net and noise perception module.

STFT-based backbone. Due to the superiority of U-Net in speech enhancement [9, 22], we select it as the backbone and take the noisy spectrogram S_n as input, the real and imaginary parts are considered as two channels. Specifically, the noisy spectrogram S_n is fed sequentially into four convolution down-sampling encoders, a TF-GRU module and four convolution up-sampling decoders. Particularly, the TF-GRU module learns to model time and frequency relations respectively, which is similar to [22], except that GRU is adopted in our work. Note that a downsampling convolution layer after the TF-GRU module is used to transform feature dimension to 1 for integrating time branch. The decoder is built symmetrically with skip connections and outputs the predicted spectrogram S_p .

Waveform-base U-Net. Non-stationary noises are easily distinguished in the time domain. Therefore, integrating the time domain into the backbone is more robust for various noises [7]. Besides, time and frequency are also complementary views for speech, which effectively reduces speech distortions. For the time branch, we adopt a parallel U-Net architecture that is similar to backbone. As illustrated in Figure 1b, we take the time-domain signal T_n as input. The last convolution encoder outputs of two branches are added to a multi-head self-attention layer (MHA), and then the output is sent to a diffusion model and two decoders, respectively. These encoders in time and frequency branch are identical, i.e., each encoder contains two modules as shown in Figure 1c. Note that time branch uses 1-D convolution

and 2-D for the backbone network.

Noise perception module. In real-world scenarios, noise types are various so that training data cannot cover well. Besides, supervised noise classification is also not a good choice since public noise sets are finite and environmental noises are complex [17, 23]. Inspired by [24], in this work, we design an unsupervised fashion to learn frame-level noise. Specifically, the noisy magnitude spectrogram $|S_n|$ is used as the input of the noise perception module. Next, four convolution layers and a bi-directional GRU are utilized to get frame-level representations. Then, a 8-head attention and 16 learnable noise templates are employed to get frame-level noise information. Additionally, a modulation parameter pair (γ, β) is obtained by two multi-layer perceptrons (MLP), respectively. Finally, for each frame, we can get a noise-related fusion given by:

$$N_{fusion} = E \odot \gamma \oplus \beta, \quad (1)$$

where $E \in \mathcal{R}^{F \times C}$ is the first encoder output of backbone, $\gamma, \beta \in \mathcal{R}^{1 \times F}$, F is feature dimension and C is the number of channels. \oplus, \odot are the element-wise addition and multiplication along the frequency axis. We fuse the conditional noise module at first backbone encoder output as Figure 1b shown.

2.2. Condition diffusion model

Diffusion-based speech enhancement task can be considered as a subtask of conditional generation systems. In this system, clean target speech can be generated from the noisy speech by utilizing a conditional diffusion-based model. In our paper, we also design a unified framework by incorporating the multi-view condition into the diffusion-based generative model with the forward and reverse processes.

Generally, following the work [19, 25], diffusion-based speech enhancement can define the forward stochastic diffusion processes as the general solution form to a linear SDE:

$$dx_t = f(x_t, y) dt + g(t) dw, \quad (2)$$

where x_t is the current state, y is the noisy condition signal, $t \in [0, T]$ is a continuous variable describing the current t -step in this process. w denotes a standard Wiener process. $f(x_t, y)$ is called the drift coefficient, and $g(t)$ is the diffusion coefficient, which controls the scale of the Gaussian noise injected at the

Table 1: Results of simulated (reverb and no reverb) and real-world datasets. Models sorted by the algorithm type, discriminative (D) or generative (G) are listed. SGMSE+ and StoRM use 50 reverse steps. Metrics higher are better and the best results are listed in bold.

Method	Type	Simulated reverb			Simulated no reverb			Real-world
		ESTOI	SI-SDR	MOS	ESTOI	SI-SDR	MOS	MOS
Mixture	-	0.47	8.23	2.43 \pm 0.15	0.62	12.13	3.11 \pm 0.13	2.73 \pm 0.13
HDemucs	D	0.70	15.86	3.09 \pm 0.13	0.83	16.88	3.82 \pm 0.15	3.64 \pm 0.13
BSRNN	D	0.77	16.49	3.27 \pm 0.15	0.90	17.37	3.89 \pm 0.13	3.68 \pm 0.15
MDM	D	0.83	17.58	3.40 \pm 0.16	0.91	17.83	3.93 \pm 0.14	3.77 \pm 0.13
SGMSE+	G	0.79	16.11	3.13 \pm 0.15	0.85	17.22	3.69 \pm 0.15	3.66 \pm 0.16
StoRM	G	0.82	17.25	3.31 \pm 0.12	0.91	18.09	3.94 \pm 0.13	3.84 \pm 0.17
MDDM	G	0.86	18.13	3.49 \pm 0.15	0.93	18.51	4.01 \pm 0.15	3.93 \pm 0.16

current time-step t .

Furthermore, for the reverse diffusion process, it has an associated solution of the reverse SDE according to Equation 2, which can be defined as follows:

$$dx_t = [-f(x_t, y) + g(t)^2 \nabla_{x_t} \log p_t(x_t|y)] dt + g(t) d\bar{w}, \quad (3)$$

where $d\bar{w}$ is a Brownian motion and $\nabla_{x_t} \log p_t(x_t|y)$ is the gradient term of conditional probability density distribution that can be estimated by a neural network called score model $s_\theta(x_t, y, t)$. Finally, in inference, we can obtain the reverse SDE updated as:

$$dx_t = [-f(x_t, y) + g(t)^2 s_\theta(x_t, y, t)] dt + g(t) d\bar{w}. \quad (4)$$

Following the aforementioned procedure, in our paper, some modifies are made. Firstly, we utilize the lightweight NCSN++M architecture [20] as the score network but use cross attention for multi-view condition fusion and channel concatenation for predicted spectrogram S_p fusion. In addition, to train the score model in the frequency domain, at an arbitrary time step $t \in [0, T]$, we sample to obtain the noisy spectrogram S_t from a Gaussian distribution, which can be written as follows:

$$S_t = \mu(S_c, S_p, t) + \sigma(t) z, \quad (5)$$

where z is sampled from $\mathcal{N}(z; 0, I)$, μ and $\sigma(t)$ have the identical formula as [20]. S_p is the predicted spectrogram of discriminative network and S_c is the clean spectrogram. Furthermore, the training objective can be written as:

$$\arg\min_{\theta} \mathbb{E}_{t, S_p, C_m, z, S_t} \|s_\theta([S_t, S_p], C_m, t) + \frac{z}{\sigma(t)}\|_2^2, \quad (6)$$

where C_m is the multi-view condition which comes from discriminative network. The conditions of C_m and $[S_t, S_p]$ are entered into the score model for training. Finally, after the score model is trained, we can obtain the reverse SDE according to Equation 4 for inference.

2.3. Training and inference

For the training process, we first train the multi-view discriminative network 200k steps with L1 and L2 losses. Then, we utilize the multi-task learning scheme to jointly optimize the discriminative and diffusion network until convergence. For inference, we first predict the discriminative result, and then generate the noisy sample S_k at step k through the diffusion forward process as follows:

$$S_k = S_p + \sigma(k) z. \quad (7)$$

Note that S_k is not a standard normal distribution, it has the identical distribution compared with S_t in Section 2.2. Therefore, combining the intermediate multi-view feature and from S_k , the reverse diffusion process is only performed k iterations denoising to generate the clean spectrogram, and then the inversion STFT is used to get the final waveform.

3. Experiments

3.1. Datasets

Training. We use a clean 585-hour mixture dataset for model training. Specifically, 500-hour dataset is created by clean vocal track of television drama from our intranet sites. Additional 85-hour clean dataset comes from AISHELL-3 dataset [26]. DEMAND [27] and QUT-NOISE datasets [28] are selected as noise data and are split as training and testing. In addition, 10000 room impulse responses (RIRs) with T60 between 0.1 and 1.0 seconds are randomly simulated using gpuRIR method [29]. In training, we randomly select clean data to convolve RIRs, and then mixed noisy data is obtained by mixing noise and reverb speech at a random uniform distribution signal-noise ratio (SNR) between 0 and 20 dB. Finally, the data distributions of noise-only, reverb-only and both are 40%, 30%, 30% in mixed noisy data, respectively.

Testing. In testing, we use two datasets, a simulated dataset and a real-world dataset. Specifically, for the simulated dataset, we randomly select 100 utterances from AISHELL-3 dataset for test data simulation and these utterances cannot participate in training. First, we randomly select 40% utterances to convolve extra-simulated RIRs, and then add testing noises to all utterances with a SNR uniformly sampled from $[0, 20]$ dB, which formulates the reverberation or no-reverberation test datasets. For the real-world dataset, we randomly select 20 noisy utterances of undivided track from our intranet sites, including film, drama and variety show.

3.2. Training setups and baselines

For discriminative network, (kernel 4, stride 2) along the frequency and (kernel 8, stride 4) along the time are used for all encoders in frequency and time network branches, respectively. The dilation factors are 1 and 2 in each encoder. For the backbone network, the time axis is no downsampling except for the last encoder (stride is 2). The first output channel is 32 and factor is 2 for both branches. The hidden units of TF-GRU are 256 and the convolution after it uses kernel 16 and stride 1 along the frequency axis. For noise perception module, the bi-directional GRU units are 128 and the dimension of noise templates is set

Table 2: The results of multi-view ablation experiments on the simulated "no reverb" and real-world datasets.

Method	Simulated no reverb			Real-world
	ESTOI	SI-SDR	MOS	MOS
MDDM	0.93	18.51	4.01 \pm 0.15	3.93 \pm 0.16
w/o. time	0.86	17.81	3.93 \pm 0.14	3.84 \pm 0.15
w/o. noise	0.90	18.25	3.97 \pm 0.16	3.79 \pm 0.16
w/o. both	0.81	17.52	3.88 \pm 0.15	3.74 \pm 0.16

as 256. For the diffusion, the same configuration follows [20]. The hop size and FFT length are 128 and 512, and the window length is 512. All training samples are resampled at 24k Hz. We use a max of 100 epochs for all training.

Two discriminative methods (HDemucs¹ [8] and BSRNN² [6]) and two generative methods (SGMSE+³ [19] and StoRM⁴ [20]) are used as baselines. We train all baselines using public available codes. Besides, we also compare our multi-view discriminative model (MDM) with baselines. For MDDM, the sampling step is set as 30. For evaluation, scale-invariant signal-to-distortion ratio (SI-SDR) [30] and extended short-time objective intelligibility (ESTOI) [31] are used as objective metrics and mean opinion score (MOS) [32] is subjective metric. For MOS test, 20 samples are randomly selected from each test dataset. A total of ten people participate and participants are required to evaluate each utterance once.

3.3. Results

Table 1 shows the subjective and objective experiment results of all compared methods. Over the table, "Simulated reverb", "Simulated no reverb" and "Real-world" are the simulated test datasets with and without reverberation, as well as created real unlabeled noisy samples, respectively. "Mixture" refers to the original noisy samples. Notably, generative models generally achieve better performance in perception-related MOS compared to discriminative models, e.g., higher values of StoRM than HDemucs and BSRNN. That is because generative systems can alleviate the excessive noise suppression situation to a certain extent due to the distribution learning ability. Moreover, we extend the comparison of discriminative and generative baseline methods with our proposed method. As shown in Table 1, several observations can be made. The MDDM achieves the best enhancement performance on all test datasets compared to other baselines in terms of subjective and objective metrics. Specifically, MDDM achieves the best results in terms of the ESTOI and SI-SDR on two simulated test datasets, indicating that our method can reduce speech distortions to a certain extent. In addition, MDDM obtains 3.93 MOS that is higher than other baselines on the real-world dataset, which shows powerful generalizability capacity and noise robustness. Remarkably, the multi-view discriminative model (MDM) with no diffusion process can also achieve the competitive results compared to StoRM and surpasses other discriminative methods, demonstrating the effectiveness of multi-view information fusion. Moreover, it's worth noting that MDDM only uses the sampling step of 30 to achieve impressive performance, which greatly mitigates the computational cost compared to other diffusion-based methods

¹<https://github.com/facebookresearch/demucs>

²<https://github.com/sungwon23/BSRNN>

³<https://github.com/sp-uhh/sgmse>

⁴<https://github.com/sp-uhh/storm>

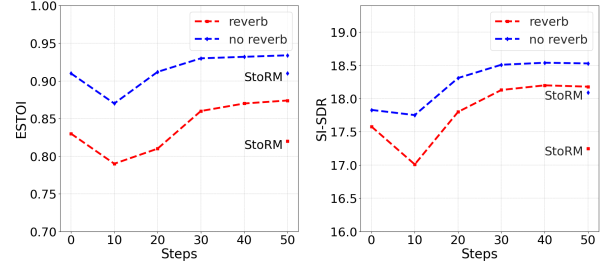


Figure 2: The objective experiment results on two simulated test datasets. The results of StoRM with the sampling step of 50 are also listed. Red and blue dots denote the simulated datasets of "with reverb" and "no reverb", respectively.

using the sampling step of 50.

As illustrated in Figure 2, we explore the influence of inference sampling step on the results. We utilized the objective evaluation metrics (ESTOI and SI-SDR) on two simulated datasets for testing. The max number of total step is 50 and step 0 denotes the results of MDM. Notably, as the number of the sampling step increases, the objective results also increase slowly. Additionally, as shown in Figure 2, although a small inference sampling step (e.g., 10) may result in the performance decline compared to MDM, the sampling step of 30 yields comparatively high quality speech, demonstrating the effectiveness of our proposed method. In contrast, StoRM [20] achieves the comparable performance with the sampling step of 50 and thus results in nearly 1.7 times slower than our proposed method. In our paper, for a trade-off between enhancement performance and computational burden, the inference sampling step is set as 30 for the experiments of our proposed method.

To verify the effectiveness of multi-view information fusion, we also perform some ablation studies in terms of conditions of the discriminative network. The time domain, noise domain and both domains are removed for testing, respectively. Table 2 shows the ablation results on simulated reverb and real-world datasets. Specifically, each of them removing can damage the performance. Furthermore, removing the time domain results in more speech distortions, reflected by lower ESTOI and SI-SDR values. In addition, removing the information of noise domain brings the limited generalizability capacity, reflected by a lower MOS value in real-world dataset. Therefore, the feature of each view is effective for speech enhancement so that multi-view feature can combine their strengths and thus achieve the superior enhancement performance.

4. Conclusion

In this paper, we propose a multi-view discriminative enhanced diffusion-based speech enhancement model. Time, frequency and noise domains are integrated into a unified framework. Multi-view feature not only generates a better discriminative output, but also can be used as a condition of diffusion-based model. Furthermore, this multi-view condition improves the generalization capabilities of enhancement model and reduces speech distortions. In addition, due to the superior enhancement performance of the discriminative network, the discriminative output has nearly identical distributions compared to the clean target. Therefore, only a small inference sampling step can be used to get final superior results, which greatly alleviates the computational burden.

5. References

- [1] T. Gerkmann and E. Vincent, "Spectral masking and filtering," *Audio Source Separation and Speech Enhancement*, pp. 65–85, 2018.
- [2] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proceedings of the 19th International Society of Music Information Retrieval Conference (ISMIR)*, 2018, pp. 334–340.
- [3] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [4] 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," in *Interspeech*, 2020, pp. 2472–2476.
- [5] H. Li and J. Yamagishi, "Noise tokens: Learning neural noise templates for environment-aware speech enhancement," in *Interspeech*, 2020, pp. 2452–2456.
- [6] J. Yu, Y. Luo, H. Chen, R. Gu, and C. Weng, "High fidelity speech enhancement with band-split rnn," in *Interspeech*, 2023, pp. 2483–2487.
- [7] K. Zhang, S. He, H. Li, and X. Zhang, "Dbnet: A dual-branch network architecture processing on spectrum and waveform for single-channel speech enhancement," in *Interspeech*, 2021, pp. 2821–2825.
- [8] A. Défossez, "Hybrid spectrogram and waveform source separation," in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation (ISMIR)*, 2021.
- [9] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [10] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2018, pp. 1–6.
- [11] J. Richter, G. Carbajal, and T. Gerkmann, "Speech enhancement with stochastic temporal convolutional networks," in *Interspeech*, 2020, pp. 4516–4520.
- [12] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 30, pp. 2993–3007, 2022.
- [13] S. Routray and Q. Mao, "Phase sensitive masking-based single channel speech enhancement using conditional generative adversarial network," *Computer Speech & Language (CSL)*, vol. 71, 2022.
- [14] H. Liu, X. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "Voicefixer: A unified framework for high-fidelity speech restoration," in *Interspeech*, 2022, pp. 4232–4236.
- [15] W. Tai, F. Zhou, G. Trajcevski, and T. Zhong, "Revisiting denoising diffusion probabilistic models for speech enhancement: Condition collapse, efficiency and refinement," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 37, 2023, pp. 13 627–13 635.
- [16] Z. Guo, J. Du, C.-H. Lee, Y. Gao, and W. Zhang, "Variance-preserving-based interpolation diffusion models for speech enhancement," in *Interspeech*, 2023, pp. 1065–1069.
- [17] Y. Hu, C. Chen, R. Li, Q. Zhu, and E. S. Chng, "Noise-aware speech enhancement using diffusion probabilistic model," in *Interspeech*, 2024, pp. 2225–2229.
- [18] W. Tai, Y. Lei, F. Zhou, G. Trajcevski, and T. Zhong, "Dose: Diffusion dropout with adaptive prior for speech enhancement," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2024, pp. 40 272–40 293.
- [19] J. Richter, S. Welker, J.-M. Lemerrier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, pp. 2351–2364, 2023.
- [20] J.-M. Lemerrier, J. Richter, S. Welker, and T. Gerkmann, "Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, pp. 2724–2737, 2023.
- [21] Y. Yang, N. Trigoni, and A. Markham, "Pre-training feature guided diffusion model for speech enhancement," in *Interspeech*, 2024, pp. 1185–1189.
- [22] S. Lv, Y. Hu, S. Zhang, and L. Xie, "Dccrn+: Channel-wise sub-band dccrn with snr estimation for speech enhancement," in *Interspeech*, 2021, pp. 2816–2820.
- [23] Y.-X. Wang, J. Du, L. Chai, C.-H. Lee, and J. Pan, "A noise-aware memory-attention network architecture for regression-based speech enhancement," in *Interspeech*, 2020, pp. 4501–4505.
- [24] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning (ICML)*, 2018, pp. 5180–5189.
- [25] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations (ICLR)*, 2021.
- [26] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," in *Interspeech*, 2021, pp. 2756–2760.
- [27] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*, 2013.
- [28] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The qut-noise-timit corpus for evaluation of voice activity detection algorithms," in *Proceedings of International Speech Communication Association (ISCA)*, 2010, pp. 3110–3113.
- [29] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpurir: A python library for room impulse response simulation with gpu acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.
- [30] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr-half-baked or well done?" in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.
- [31] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 24, no. 11, pp. 2009–2022, 2016.
- [32] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (mos) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.