



Thunder : Unified Regression-Diffusion Speech Enhancement with a Single Reverse Step using Brownian Bridge

Thanapat Trachu¹, Chawan Piansaddhayanon², Ekapol Chuangsuwanich^{1,2}

¹Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University

²Center of Excellence in Computational Molecular Biology, Chulalongkorn University

6670104921@student.chula.ac.th, schwanph@gmail.com, ekapolc@cp.eng.chula.ac.th

Abstract

Diffusion-based speech enhancement has shown promising results, but can suffer from a slower inference time. Initializing the diffusion process with the enhanced audio generated by a regression-based model can be used to reduce the computational steps required. However, these approaches often necessitate a regression model, further increasing the system's complexity. We propose *Thunder*, a unified regression-diffusion model that utilizes the Brownian bridge process which can allow the model to act in both modes. The regression mode can be accessed by setting the diffusion time step closed to 1. However, the standard score-based diffusion modeling does not perform well in this setup due to gradient instability. To mitigate this problem, we modify the diffusion model to predict the clean speech instead of the score function, achieving competitive performance with a more compact model size and fewer reverse steps.

Index Terms: speech enhancement, diffusion, Brownian bridge

1. Introduction

Speech enhancement (SE) focuses on removing noisy signals from the input speech to improve its comprehensibility and has been deployed in several real-life systems [1, 2, 3]. It could also be integrated with existing downstream tasks, such as speech recognition (ASR) [4, 5, 6] or speech verification (SV) [3, 7], to improve speech quality under adverse environments.

Speech enhancement systems, as categorized in [8], can be classified into two approaches: regressive [9, 10, 11] and generative [12, 13, 14]¹. The objective of the regression model is to learn a deterministic mapping between noisy and clean speech, whereas the generative model aims to capture the target distribution, allowing the generation of multiple valid possibilities instead of a single one. Recently, there has been a surge of interest in the diffusion model for speech enhancement [13, 14] due to its promising outcomes across various domains [15].

Despite the promising outcome, one major obstacle to the practical application of diffusion for SE is its slow inference time caused by multiple reverse diffusion steps. Thus, numerous studies have been proposed to address this issue. StoRM [8] utilized a two-stage regression-diffusion pipeline where the first model is responsible for enhancing the noisy speech in a regressive manner while the second stage is used for refining the output from the former stage using a reverse diffusion process. Since the input to the latter model is pre-cleaned, the difficulty of the reverse process decreases, requiring fewer diffusion steps. Nevertheless, this approach requires two independent models—regression and diffusion—leading to a substantial increase in the number of parameters. To address this issue, an

¹In this paper, a regressive model is a deterministic mapping between noisy and clean speech while the generative model is not.

additional head to predict both the score function and the noiseless signal was introduced in the Diffusion-based Joint Predictive and Diffusion model [16], achieving competitive outcomes while incurring fewer parameters. Nevertheless, it still requires an additional prediction head for regressive prediction.

We introduce *Thunder*, a unified regression-diffusion model capable of performing both regression and diffusion while not incurring additional parameters. We propose the use of the Brownian bridge process for diffusion-based speech enhancement which allows the model to act as both a regression and a diffusion model at the same time. Instead of modeling the score function like in typical diffusion modeling, we reparameterize the model to predict the noiseless speech to avoid the gradient instability issue and allow a single step prediction if desired (regression mode). Our method achieves competitive results on the VoiceBank + DEMAND dataset using fewer parameters and shorter inference time. Remarkably, our approach outperforms the diffusion baselines on even just one reverse diffusion step highlighting the effectiveness of the Brownian bridge process.

2. Score-based diffusion model

2.1. Forward and reverse process

Diffusion modeling comprises two essential processes: the forward process and the reverse process. In the forward process, noise is incrementally introduced into a clean speech until it becomes pure noise. Conversely, the reverse process gradually eliminates noise from noisy speech, ultimately yielding clean speech. Within the framework of score-based diffusion [15], a stochastic differential equation (SDE) is employed to represent these processes. Specifically, the forward process is represented by the following SDE:

$$dx_t = f(x_t, y)dt + g(t)dw \quad (1)$$

where x_t, y, w denotes the current state of the process at time step t , noisy speech, and a standard Wiener process, respectively. The state x_t is indexed by a continuous time variable t within the interval $[0, 1]$, in which x_t is a clean speech when $t = 0$ and a pure noise when $t = 1$. The functions $f(x_t, y)$ and $g(t)$ signify the drift coefficient and diffusion coefficient, respectively. Following [15], the reverse SDE of the Eq. 1 is:

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

There have been works [13, 14] proposed to design the SDE process for speech enhancement tasks by designing $f(x_t, y)$ and $g(t)$ that could directly transform the noisy speech into clean speech instead of Gaussian noise. For example, SGMSE+

[14] proposed the following drift and diffusion coefficient:

$$f(x_t, y) = \gamma(y - x_t) \quad (3)$$

$$g(t) = \sigma_{\min} \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)^t \sqrt{2 \log \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)} \quad (4)$$

where γ denotes the transformation speed between the clean speech and the noisy speech, and $\sigma_{\min}, \sigma_{\max}$ are the parameters controlling the variance in x_t . However, the presences of Gaussian noise still exist at $t = 1$ due to a non-zero variance. Therefore, in this paper, we have selected the subsequent drift and diffusion coefficients:

$$f(x_t, y) = \frac{y - x_t}{1 - t}; \quad g(t) = 1 \quad (5)$$

This particular SDE is referred to as the Brownian bridge process [17]. Its distinguishing feature is that it can linearly transform between the initial state (x_0) with zero variance to the noisy speech y with zero variance, offering a capability to perform as a regression model at $t = 1$ (deterministic mode).

2.2. Score-function model

As calculating $\nabla_{x_t} \log p_t(x_t)$ is intractable, following [15, 18], *denoising-score-matching* is instead performed by having the score-based model $s_\theta(x_t, y, t)$, typically a neural network, approximates $\nabla_{x_t} \log p_t(x_t|x_0)$, the value of which can be determined using the given initial state x_0 [19]:

$$p_t(x_t|x_0, y) = \mathcal{N}_{\mathbb{C}}(x_t; \mu(x_0, y, t), \sigma(t)^2 \mathbf{I}) \quad (6)$$

$$\mu(x_0, y, t) = x_0(1 - t) + yt \quad (7)$$

$$\sigma(t)^2 = t(1 - t) \quad (8)$$

where $\mathcal{N}_{\mathbb{C}}$ represents the circularly symmetric complex normal distribution, $\mu(x_0, y, t)$ denotes a mean, and $\sigma(t)$ is a standard deviation. Consequently, the training loss is defined as:

$$\mathcal{J}(\theta) = \mathbb{E}_{t, x_t, (x_0, y) \sim p_{\text{data}}} [\lambda(t) \|s_\theta(x_t, y, t) + \frac{z}{\sigma(t)}\|_2^2] \quad (9)$$

where $\mathcal{J}(\theta)$ is an objective function, and t, x_t are randomly sampled from $\mathcal{U}[0, 1]$ and $p_t(x_t|x_0)$, respectively. z is drawn from $\mathcal{N}(0, \mathbf{I})$, and $\lambda(t)$ serves as a weight function that is set to $\sigma(t)^2$ in [8, 14, 20].

2.3. Inference

To generate the predictions, the reverse SDE has to be estimated through a numerical SDE solver using the PC sampler [15] consisting of a predictor and corrector. Initially, x_1 is set to y . Then, the predictor updates the current state x_t into the next state $x_{t-\Delta t}$ by discretizing the reverse SDE using finite time steps that is subsequently fed to the corrector to refine the prediction by using only the score function. The process was iteratively repeated until $t = 0$. In this paper, we follow [15] and use the Euler-Maruyama and Langevin dynamics as a predictor and corrector, respectively.

3. Methodology

Drawing inspiration from StoRM [8] and the Joint Generative and Predictor method [16], we propose to further condense StoRM into a single model that can switch between two modes: diffusion and regression. Specifically, we train the model to predict x_0 instead of the score function and leverage the property of the Brownian bridge process to enable regressive capability.

3.1. Model parameterization

To allow the model to possess a regressive capability, the Brownian bridge process is employed. However, directly applying this process to the SDE is inappropriate since $\sigma(t)$ becomes very close to 0 when $t \rightarrow 1$ (Eq. 8), making minimizing the Eq. 9 impractical as the gradient is directly proportional to $\sigma(t)$, as shown below.

$$\nabla_{\theta} \mathcal{J}(\theta) = \nabla_{\theta} [\|\sigma(t)s_\theta(x_t, y, t) + z\|_2^2] \quad (10)$$

$$= 2\sigma(t) \nabla_{\theta} s_\theta [\sigma(t)s_\theta(x_t, y, t) + z]_2 \quad (11)$$

This hampers the model's ability to efficiently estimate the score function at $t = 1$ under one reverse step (Eq. 2). Even if the accurate score function s_θ is to be obtained, it is still infeasible to employ the regression mode at $t = 1$ due to the inability to estimate the clean speech (x_0) from the score function as shown in the following equations, derived from Eq. 6:

$$x_t \sim \mathcal{N}_{\mathbb{C}}(\mu(x_0, y, t), \sigma(t)^2 \mathbf{I}) \quad (12)$$

$$x_t = x_0(1 - t) + yt + \sqrt{t(1 - t)}z \quad (13)$$

$$x_t = x_0(1 - t) + yt - t(1 - t)s_\theta(x_t, y, t) \quad (14)$$

$$x_0 = \frac{x_t - yt + t(1 - t)s_\theta(x_t, y, t)}{1 - t} \quad (15)$$

where Eq. 13 follows the reparameterization trick from [21], and $s_\theta(x_t, y, t) \approx -z/\sigma(t)$ when optimal.

To overcome this problem, we modify the model to predict $\tilde{x}_\theta(x_t, y, t)$, an estimation of clean speech x_0 , instead of the score function, allowing our model to be used as a regression model at any t . In diffusion mode, we can perform the reverse process by first computing the score function via:

$$s_\theta(x_t, y, t) = -\frac{x_t - (\tilde{x}_\theta(x_t, y, t)(1 - t) + yt)}{t(1 - t)} \quad (16)$$

Then, the obtained score function can be used to solve the reverse SDE as described in 2.1.3. During inference, at the initial stage of the reverse process, t is set close to 1 to circumvent numerical instability. The training approach remains the same, with the training objective adjusted to:

$$\mathcal{J}(\theta) = \mathbb{E}_{t, x_t, x_0, y} [\|\tilde{x}_\theta(x_t, y, t) - x_0\|_2^2] \quad (17)$$

3.2. Justification for the Brownian bridge process

This subsection provides some analysis to justify our choice of the Brownian bridge process, the drift and diffusion coefficient of the SDE, for the speech enhancement task. The *drift coefficient* of the reverse Brownian Bridge as $t \rightarrow 1$ converges to the noise in the speech as shown in the equations below:

$$\begin{aligned} & \lim_{t \rightarrow 1} f(x_t, y) - g(t)^2 s_\theta(x_t, y, t) \quad (\text{From (2)}) \\ &= \lim_{t \rightarrow 1} \frac{t(y - x_t) + x_t - (\tilde{x}_\theta(x_t, y, t)(1 - t) + yt)}{t(1 - t)} \\ &= \lim_{t \rightarrow 1} \frac{(1 - t)x_t - \tilde{x}_\theta(x_t, y, t)(1 - t)}{t(1 - t)} \\ &= \lim_{t \rightarrow 1} \frac{x_t - \tilde{x}_\theta(x_t, y, t)}{t} \\ &= y - \tilde{x}_\theta(x_1, y, 1) = \tilde{n} \end{aligned} \quad (18)$$

The equation above suggests that the reverse process would update in the direction of noise in the input speech \tilde{n} , matching the modeling assumption of speech enhancement where the noisy

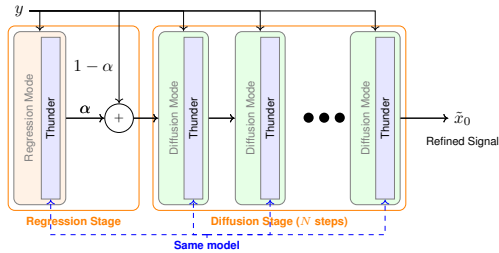


Figure 1: A summarization of Thunder during inference. The regression mode is first applied to the noisy input speech to improve the signal quality before being further refined through the diffusion mode. To reduce over-denoising artifacts caused by the regression part, the processed signal is fused with the original input to preserve its characteristic. The weights are shared across the two modes.

speech is the clean speech corrupted by the noise through an additive operation ($y = x_0 + n$). In an extreme scenario where the number of steps in the reverse process is set to 1, the model transforms into the regression mode, deterministically updating x_1 by subtracting the predicted noise. By enabling the regressive ability, the model could now enhance the speech with fewer reverse diffusion steps when initialized with the enhanced audio from the regression mode [8].

3.3. Utilizing the regression potential of the score-based model

Figure 1 summarizes the inference process. Following StoRM [8], our pipeline is a two-stage process which is a regression model followed by a generative model. The first improves the signal quality, while the latter aims to reduce the artifacts generated by the regression model. However, unlike StoRM, the two models are shared but used slightly differently in different modes. The regression mode is done by predicting x_0 directly from our model $\hat{x}_\theta(x_1 = y, y, t = 1)$. The diffusion mode is performed by acquiring the score function according to Eq. 16, and the reverse diffusion process can be performed for N steps as outlined in Section 2.3.

As suggested in [22], the input to the diffusion stage is a linear interpolation between the output from the regression mode and the noisy input speech which can help minimize the occurrence of the *over-denoising artifact*, a concept discussed and employed in [8, 13, 23]. The interpolation weight α can be chosen via grid search on a validation set.

4. Experiments

4.1. Experimental settings

We benchmarked the performance of our proposed method (Thunder) on the VoiceBank + DEMAND dataset [25, 26], consisting of 30 speakers from the Voicebank Corpus [25]. We followed the prior works [8, 14] and separated the dataset into training (26 speakers), validation (speaker “p226”, “p287”) and testing (2 speakers) sets. The training and validation sets consist of 11,572 utterances corrupted by eight recorded noise samples from DEMAND and two artificially generated noise samples (babble and speech-shaped) at SNR levels of 0, 5, 10, and 15 dB, while the testing set contains 824 utterances, each contam-

Table 1: Performance of different speech enhancement methods on the VoiceBank + DEMAND dataset. “Type” refers to the model type (“R” for regression and “G” for generative model). Numbers before and after slash refer to the performance of small (S) and large (L) NCSN++ variants, respectively. The model with the best performance in each section is underlined, and the best score in the table is bolded.

System	Type	PESQ \uparrow	ESTOI \uparrow	SI-SDR \uparrow
Noisy	-	1.97	0.79	8.4
Conv-Tasnet [9]	R	2.84	0.85	19.1
MetricGAN+ [10]	R	3.13	0.83	8.5
NCSN++M (L) [24]	R	2.82	<u>0.87</u>	<u>19.9</u>
SEGAN [12]	G	2.16	-	-
CDiffuSE [13]	G	2.46	0.79	12.6
SGMSE+ (L) [14]	G	2.93	0.87	17.3
BBED (L) [20]	G	2.95	0.87	18.7
StoRM (S) [8]	R+G	2.93	0.88	18.8
GP-Unified (L) [16]	R+G	<u>2.97</u>	0.87	18.3
Thunder (S/L)				
Regression mode	R	2.78/2.85	0.87/0.87	<u>19.6/19.7</u>
Diffusion mode	G	2.87/2.95	0.87/0.87	18.8/18.6
Mixture ($\alpha = 0.8$)	R+G	<u>2.97/3.02</u>	<u>0.87/0.87</u>	19.3/19.4

Table 2: Number of parameters in each model.

System	StoRM (S)	GP-Unified (L)	Thunder (S/L)
Parameters	55.6M	106M	27.8M/65.6M

inated with different noise samples at SNR levels of 2.5, 7.5, 12.5, and 17.5 dB. All speech data were sampled at 16 kHz.

We also followed prior works [8, 14, 16] and used the Noise Conditional Score Network (NCSN++)² [24] as a base architecture with 30 reverse diffusion steps for benchmarking with minor modifications. The model was used to predict clean speech instead of the score function, and the SDE was transformed into a Brownian bridge process. Note that the NCSN++ in StoRM [8] differs from that of SGMSE+ [14] and GP-Unified [16] since it utilizes a smaller NCSN++ variant (27.8M) for both regression and diffusion models, whereas SGMSE+ and GP-Unified employ the larger variant (65.6M). For a fair comparison, we performed evaluations on both variants.

The model was trained for 100 epochs on one Nvidia RTX4090, with Adam optimizer, a learning rate of 2×10^{-5} , and a batch size of 8. We used Perceptual Evaluation of Speech Quality (PESQ) [27], Extended Short-Time Objective Intelligibility (ESTOI) [28], Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [29], and Scale-Invariant Signal-to-Artifact Ratio (SI-SAR) [29] as evaluation metrics.

4.2. In-domain evaluation

Table 1 compares our method to other approaches. It was found that our model achieved a competitive result compared to other state-of-the-art models. Additionally, compared to other diffusion-based approaches (“StoRM”, “GP-Unified”), our method used half of the parameters (Table 2) while performing competitively compared to StoRM and GP-Unified.

²Our implementation was based on <https://github.com/sp-uhh/storm>

Table 3: The performance of Thunder (L) when varying the number of reverse time steps (N) using PC sampler. The RTF is the average time to process one second of audio. The experiments were conducted using Nvidia RTX 4090. Corrector denotes the corrector in the PC sampler.

N	Corrector	RTF[s] ↓	PESQ ↑	SI-SDR ↑
30	✗	0.538	3.02	19.4
30	✓	1.084	3.02	19.4
15	✗	0.284	3.02	19.4
15	✓	0.552	3.02	19.4
1	✗	0.038	2.99	19.6
1	✓	0.056	2.99	19.6

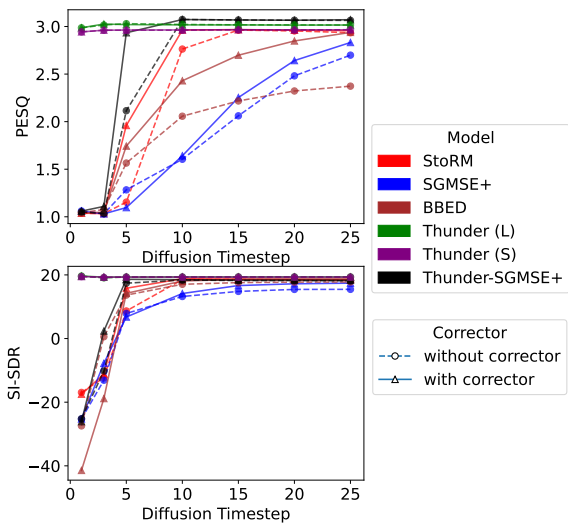


Figure 2: PESQ and SI-SDR under different numbers of diffusion time steps. Thunder performed competitively with other approaches even with just one diffusion step.

Figure 2 shows that our method maintains competitive results while requiring much fewer diffusion steps compared to the baselines (30 reverse steps), even achieving real-time inference (Table 3). To justify our design choice, we also provide comparisons against BBDE [20] and Thunder-SGMSE+. The BBDE used the same SDE as ours, but it predicted the score function instead of the clean speech. On the other hand, Thunder-SGMSE+ is our method, but the SDE is changed to be the same as SGMSE+. Note that for $N = 1$, the process requires two forward passes: regression and diffusion.

4.3. Out-of-domain evaluation

We further examine the generalizability of our method by performing an evaluation on the LibriFSD50k, the LibriSpeech dataset [30] corrupted by noise uniformly added from the FSD50k dataset [31] at SNR levels ranging from 0 to 20, without any fine-tuning. The result in Table 4 suggests that our model could still generalize under the out-of-domain setting, outperforming the other baselines (paired two-sample t-test, $p < 0.01$). Interestingly, there is only a slight degradation when reducing the number of reverse steps from thirty to one, implying that our regression mode is highly effective at eliminating the noise, requiring only one reverse step to refine.

Table 4: The performance of Thunder (L) under mismatched training conditions on the FSD50k dataset. We achieved better generalization than the MetricGAN+ because a lower relative performance change between the out-of-domain and in-domain conditions was observed.

System	Type	PESQ ↑	SI-SDR ↑	SI-SAR ↑
Noisy	-	1.92	10.0	-
MetricGAN+	R	2.18	5.8	6.1
NCSN++M (L)	R	2.03	14.7	17.0
SGMSE+ (L) (30 steps)	G	2.19	14.2	15.8
StoRM (S) (30 steps)	R+G	2.12	14.3	16.3
Thunder (L)				
Regression Mode	R	2.04	14.7	16.7
Mixture (30 steps, $\alpha = 0.8$)	R+G	2.21	14.7	17.0
Mixture (1 step, $\alpha = 0.8$)	R+G	2.21	14.7	17.0

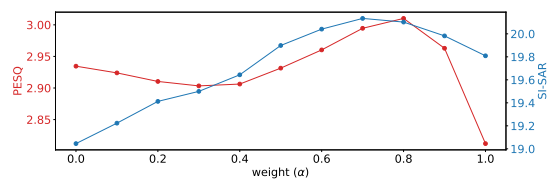


Figure 3: PESQ and SI-SAR of Thunder (L) at different interpolation weights α . At high α , performance degradation was observed due to artifacts from the regression mode, obstructing the refinement process during the diffusion mode.

4.4. Effect of regression mode

We then investigated the effect of having a regression mode as the first step by varying the interpolation weight α from 0 to 1 (no blending with the original signal) while setting t to 1. Figure 3 shows that high values of α led to audio quality degradation, as a sharp decline in PESQ and SI-SAR scores was observed when $\alpha > 0.8$ and $\alpha > 0.7$, respectively. This indicates that the regression mode generated excessive artifacts for the diffusion mode to refine. Despite this, the diffusion mode could still effectively eliminate artifacts when a sufficient degree of noisy speech y was added to reduce the artifacts, thereby enhancing PESQ and improving the performance. On the other hand, the model yielded the lowest SI-SAR when the assistance from the regression mode ($\alpha = 0$) was removed, suggesting its ability to reduce the difficulty of the reverse process, as also observed in an increase in SI-SAR and PESQ when α was around 0.5-0.8.

5. Conclusion

We proposed Thunder, a unified regression-diffusion model for speech enhancement. The model is trained to predict the clean speech instead of the score function to efficiently leverage the Brownian bridge process, allowing the model to possess both regressive and generative capabilities without incurring additional parameters. Our method achieves competitive results compared to other diffusion baselines on in-domain settings even with a single reverse diffusion step. It also outperforms other baselines in out-of-domain situations. For future work, we plan to extend Thunder to cover more general settings such as dereverberation.

6. References

- [1] G. Park, W. Cho, K.-S. Kim, and S. Lee, "Speech Enhancement for Hearing Aids with Deep Learning on Environmental Noises," *Applied Sciences*, vol. 10, no. 17, 2020.
- [2] K. Tan, X. Zhang, and D. Wang, "Real-Time Speech Enhancement for Mobile Communication Based on Dual-Channel Complex Spectral Mapping," in *ICASSP*, 2021, pp. 6134–6138.
- [3] S. Shon, H. Tang, and J. Glass, "VoiceID Loss: Speech Enhancement for Speaker Verification," in *Proc. Interspeech*, 2019, pp. 2888–2892.
- [4] Q.-S. Zhu, J. Zhang, Z.-Q. Zhang, and L.-R. Dai, "A Joint Speech Enhancement and Self-Supervised Representation Learning Framework for Noise-Robust Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1927–1939, 2023.
- [5] A. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent Neural Networks for Noise Reduction in Robust ASR," in *Proc. Interspeech*, 2012.
- [6] Y. Koizumi, S. Karita, A. Narayanan, S. Panchapagesan, and M. A. U. Bacchiani, "SNRi Target Training for Joint Speech Enhancement and Recognition," in *Proc. Interspeech*, 2022.
- [7] S. E. Eskimez, P. Soufleris, Z. Duan, and W. Heinzelman, "Front-end speech enhancement for commercial speaker verification systems," *Speech Communication*, vol. 99, pp. 101–113, 2018.
- [8] 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*, vol. 31, pp. 2724–2737, 2023.
- [9] 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*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [10] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement," in *Proc. Interspeech*, 2021, pp. 201–205.
- [11] 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. 108499, 2022.
- [12] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Proc. Interspeech*, 2017, pp. 3642–3646.
- [13] 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, pp. 7402–7406.
- [14] 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*, vol. 31, pp. 2351–2364, 2023.
- [15] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-Based Generative Modeling through Stochastic Differential Equations," in *ICLR*, 2021.
- [16] H. Shi, K. Shimada, M. Hirano, T. Shibuya, Y. Koyama, Z. Zhong, S. Takahashi, T. Kawahara, and Y. Mitsufuji, "Diffusion-based speech enhancement with joint generative and predictive decoders," in *ICASSP*, 2024, pp. 12 951–12 955.
- [17] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*. Springer, 1998.
- [18] Y. Song and S. Ermon, "Generative Modeling by Estimating Gradients of the Data Distribution," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.
- [19] S. Särkkä and A. Solin, *Applied stochastic differential equations*. Cambridge University Press, 2019, vol. 10.
- [20] B. Lay, S. Welker, J. Richter, and T. Gerkmann, "Reducing the Prior Mismatch of Stochastic Differential Equations for Diffusion-based Speech Enhancement," in *Proc. Interspeech*, 2023, pp. 3809–3813.
- [21] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *ICLR*, 2014.
- [22] K. Iwamoto, T. Ochiai, M. Delcroix, R. Ikeshita, H. Sato, S. Araki, and S. Katagiri, "How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR," in *Proc. Interspeech*, 2022, pp. 5418–5422.
- [23] S. Welker, J. Richter, and T. Gerkmann, "Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain," in *Proc. Interspeech*, 2022, pp. 2928–2932.
- [24] J.-M. Lemerrier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," in *ICASSP*, 2023, pp. 1–5.
- [25] V.-B. Cassia, W. Xin, T. Shinji, and Y. Junichi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 09 2016.
- [26] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, p. 035081, 05 2013.
- [27] P. Recommendation, "Application Guide for objective quality measurement based on recommendation," *ITUt*, vol. 862, p. 862, 2005.
- [28] 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*, vol. 24, no. 11, pp. 2009–2022, 2016.
- [29] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" in *ICASSP*, 2019, pp. 626–630.
- [30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.
- [31] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.