Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

SNR-Aligned Consistent Diffusion for Adaptive Speech Enhancement

Yonghyeon Jun[1], *Beom Jun Woo*[1], *Myeonghun Jeong*[1], *Nam Soo Kim*[1]

[1]Department of Electrical and Computer Engineering and INMC,
Seoul National University, Seoul, South Korea

{yhjeon, bjwoo, mhjeong}@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract

Generative models have shown strong performance in speech enhancement, and consistency models further improve both speed and quality. Building upon these improvements, we propose an SNR adaptation framework that dynamically aligns the diffusion timestep with the SNR of the input signal, enhancing robustness in diverse noise conditions. In our framework, the reverse process is conditioned on a diffusion timestep that is adjusted based on the estimated SNR, while the additive Gaussian noise is modulated according to the same SNR estimate. This design enables a continuous SNR-conditioning mechanism in which the diffusion timestep serves as an SNR control parameter, allowing the model to adjust its enhancement process based on the input SNR. Experimental results demonstrate that our proposed framework consistently improves perceptual quality, with even greater improvements observed under challenging SNR conditions, highlighting its effectiveness.

Index Terms: speech enhancement, diffusion models, consistency models, SNR adaptation, SNR estimation

1. Introduction

Speech enhancement is crucial to improving auditory quality [1] and supporting downstream tasks such as speech recognition and speaker identification [2, 3]. However, conventional models often struggle to effectively generalize when applied to signals with unseen noise distributions due to their dependence on a restricted training corpus [4]. This limitation underscores the need for models capable of robust performance in various levels of noise conditions.

To address the challenge of adapting to a wide range of distributions, [5–7] integrate SNR into speech enhancement tasks. Recently, PercepNet+ [6], and SNR-NAT [7] have demonstrated the potential to leverage SNR information to handle noise-adaptive tasks. PercepNet+ achieves this by designing an SNR estimator and SNR-switched post-processing to control the degree of residual noise removal. On the other hand, SNR-NAT introduces features derived from a priori and a posteriori SNRs. Despite these advancements, existing methods often treat SNR as a discrete parameter or struggle to maintain reconstruction accuracy, limiting their flexibility under varying noise conditions.

Meanwhile, generative models innovate the text, image, and audio domains, benefiting from expressive power to simulate data distributions. Inspired by these works, many existing methods leverage generative models rather than deterministic approaches to provide a robust framework for speech enhancement [8–14]. In particular, conditional diffusion models have shown strong performance in speech enhancement tasks. Diffusion models, fundamentally score-based models, have evolved into consistency models [15] that enable efficient single-step inference. We adopt SE-Bridge [14] —a method based on consistency models—as our primary baseline due to its efficiency and robust performance in generative speech enhancement. A common characteristic among diffusion-based enhancement models is the gradual change in noise levels throughout the diffusion process, which enables these models to learn from varying noise levels across different timesteps. However, most conventional approaches find it challenging to establish a consistent relationship between the SNR and the diffusion timestep, limiting the model to fully utilize this noise progression. This limitation is illustrated in the left part of Figure 1, where signals with different SNRs are mapped to the same diffusion timestep. Moreover, enhancement methods using consistency models are trained on a wide range of SNR over diffusion timesteps, but single-step inference at a fixed timestep hinders fully utilizing it.

In this study, we propose a diffusion-based speech enhancement method that effectively utilizes data learned over varying diffusion timesteps, which correspond to different SNR conditions. As illustrated in the right part of Figure 1, the method introduces an SNR-timestep alignment mechanism, aligning the SNR of waveforms with the diffusion timestep during training. This alignment allows the diffusion timestep to act as an SNR conditioning parameter, enabling continuous parameterization. During inference, the reverse diffusion process is conditioned on a timestep determined by the estimated SNR of the input.

This mechanism overcomes limitations of existing approaches, which either apply SNR discretely or degrade reconstruction accuracy for perceptual quality. By assigning the diffusion timestep an additional role as an SNR conditioning parameter, our method fully exploits the SNR variations learned across timesteps. Additionally, leveraging the alignment between input SNR and diffusion timesteps, the model adjusts Gaussian noise levels based on the input SNR. This helps balance noise suppression and speech preservation [16], further improving speech enhancement performance.

Our results demonstrate consistent performance at various noise levels, with an average PESQ increase of 0.12 over the baseline. Moreover, the model shows particularly higher PESQ improvements in challenging SNR conditions and demonstrates the improved ability to maintain reconstruction accuracy under varying noise levels compared to the existing approach [7], highlighting its robustness in various noise scenarios.

2. Background

2.1. Consistency models for speech enhancement

Consistency models [15] are a generative framework designed to transform noise into structured data through forward and reverse processes while significantly reducing computational cost

10.21437/Interspeech.2025-1721

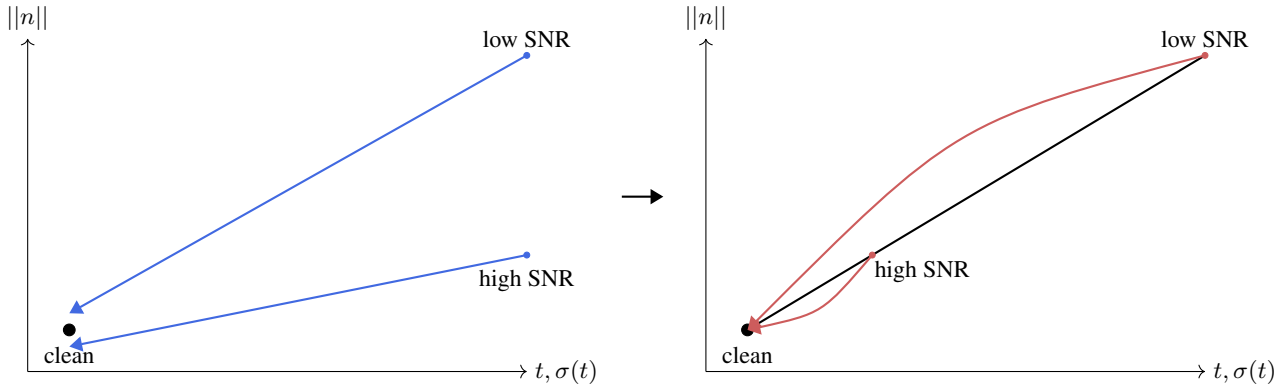

Figure 1: Illustration of the proposed SNR-adaptive diffusion path. The left diagram shows the conventional model using a fixed diffusion timestep. The right diagram represents the proposed approach, where the timestep is aligned to the SNR of the input signal and determines the appropriate Gaussian noise level $\sigma(t)$ for the waveform's SNR. $||n||$ denotes the magnitude of the acoustic noise, which is related to the SNR.

compared to conventional diffusion models [17, 18]. They achieve this by enabling single-step inference without compromising generative performance, making them well-suited for speech enhancement tasks.

In speech enhancement, consistency models incorporate noisy speech y as a conditioning input during the reverse process, guiding the reconstruction of clean speech x_0 [14]. The model employs short-time Fourier transform (STFT) coefficients of speech signals that capture the characteristics of the time-frequency domain. To bring out frequency components with lower energy, a transformation $H(\cdot)$ is applied to the STFT coefficients:

$$\tilde{c} = \frac{|c|^{\alpha}}{\beta} e^{i\angle c} = H(c) \;\Leftrightarrow\; c = \beta|\tilde{c}|^{1/\alpha} e^{i\angle \tilde{c}} = H^{-1}(\tilde{c}), \quad (1)$$

where c represents the STFT coefficients, and α and β are empirically chosen parameters. In this work, the values of α and β follow those in the baseline [12]. We assume that the signals are processed in the transformed STFT domain, where the STFT coefficients are modified using the transformation defined above.

The forward process is represented using a perturbation kernel that samples x_t at an arbitrary timestep t. The state distribution is given as:

$$p_{0t}(x_t|x_0, y) = \mathcal{N}\left(x_t; \mu_{x_0,y}(t), \sigma(t)^2\mathbf{I}\right), \quad (2)$$

where $\mu_{x_0,y}(t)$ is the mean, and $\sigma(t)^2$ represents the variance at timestep t.

In the reverse process, consistency models reconstruct the clean speech directly from the noisy input x_t guided by y, predicting the clean signal x_0 in a single step:

$$f_\theta(x_t, t, y) = \hat{x_0}. \quad (3)$$

3. Proposed method

This study proposes a generative speech enhancement model that continuously conditions the SNR by using the diffusion timestep as an SNR conditioning parameter. The model builds on the SE-Bridge [14] framework, a consistency model-based approach for speech enhancement, integrating the following techniques to improve performance and adaptability. As illustrated in Figure 2, we first introduce a non-linear perturbation kernel to align the SNR of noisy signals with the corresponding diffusion timesteps. Second, we modulate the Gaussian noise variance based on the input SNR to match the noise level. During inference, the model determines the appropriate timestep based on the estimated SNR of the input signal, allowing the enhancement process to adapt to varying noise levels. For SNR estimation, we employed the PESQNet [19] architecture, minimizing computational cost while maintaining accuracy. For the enhancement model f_θ, we adopt the NCSN++ architecture [18], following its successful use in SGMSE+ [12] and SE-Bridge [14]. The model is trained using the consistency training algorithm [15, Algorithm 3], which enables efficient sam-

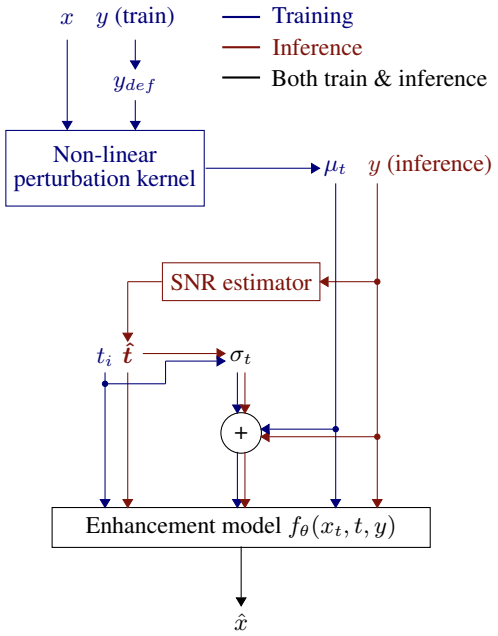

Figure 2: The architecture of the proposed model. The blue line represents the training flow, the red line represents the inference flow, and the black block indicates the components used in both.

pling while preserving high-quality speech enhancement. In our proposed work, the SNR of the signal represents the ratio between clean speech and acoustic noise, excluding Gaussian noise added during the diffusion process.

3.1. Forward process configured for SNR conditioning

To utilize t as a continuous SNR conditioning parameter, we modify the diffusion process to align the SNR of x_t and y with the diffusion timestep t. This alignment involves the following steps:

3.1.1. Refining noisy signal to a single-SNR level

During training, the input noisy signal y_0 is processed to have a fixed SNR η. This is achieved by adjusting it using the corresponding clean signal x_0 as:

$$y_\eta = H^{-1}(x_0) + 10^{\frac{\text{SNR}(y_0)-\eta}{20}}\{H^{-1}(y_0) - H^{-1}(x_0)\}. \quad (4)$$

Here, $H^{-1}(\cdot)$ is defined in equation (1). To maintain consistency in the power of the acoustic noise term, y_η is normalized by scaling it with its maximum value, yielding $y_\eta^{\text{norm}} = y_\eta/\max(y_\eta)$. After normalization, the transformation $H(\cdot)$ (Eq.1) is applied, producing $y_{def} = H(y_\eta^{\text{norm}})$, which serves as the basis for generating x_t and y during the training process.

3.1.2. Generating x_t using the proposed perturbation kernel

The perturbation kernel is modified to ensure proper alignment between the SNR of x_t and diffusion timestep t. Linear kernels fail to account for the non-linear transformations, leading to mismatches between t and the SNR properties of x_t. To address this, the proposed method uses a non-linear perturbation kernel:

$$p'_{0t}(x_t|x_0, y_{def}) = \mathcal{N}\left(x_t; \mu'_{x_0,y_{def}}(t), \sigma(t)^2\mathbf{I}\right), \quad (5)$$

where

$$\mu'_{x_0,y_{def}}(t) = H\{(1-t)H^{-1}(x_0) + tH^{-1}(y_{def})\}. \quad (6)$$

Finally, x_t is sampled from the perturbation kernel:

$$x_t = p'_{0t}(x_t|x_0, y_{def}). \quad (7)$$

3.1.3. Aligning the SNR of y with t

In the conventional training process, y is set to y_{def}, which results in a fixed SNR for y and limits the model's capabilities to handle diverse SNR conditions. To address this, y is replaced with $\mu'_{x_0,y_{def}}(t)$ to maintain alignment between the SNR of y and the diffusion timestep. This adjustment allows the model to better adapt to the varying SNR values in real-world scenarios.

3.2. Gaussian noise modulation

In diffusion-based speech enhancement models, the variance of Gaussian noise controls the trade-off between noise suppression and speech preservation [16]. Higher variance improves noise reduction but can distort speech, while lower variance retains speech quality but may reduce noise less effectively.

To allow a trade-off between noise suppression and speech preservation based on the SNR, the SNR of x_t and the magnitude of Gaussian noise included in x_t are matched. Specifically, the variance of the Gaussian noise, $\sigma(t)$, is set as:

$$\sigma(t) = \sigma_0 t. \quad (8)$$

3.3. SNR-adaptive inference

To adapt the inference process to the noise level of the input signal, we estimate the corresponding timestep $\hat{t}$ based on the noise level during inference. This is achieved using an auxiliary SNR estimator model $f_{\text{SNR};\phi}$, which predicts the SNR of the input noisy signal as:

$$\hat{\text{SNR}}(y_0) = f_{\text{SNR};\phi}(y_0). \quad (9)$$

To train the SNR estimator, we normalize the SNR values into the range $[0, 1]$ so that they align with the model's output range. The transformation is defined as:

$$\xi = \frac{10^{-\text{SNR}/20}}{1 + 10^{-\text{SNR}/20}}. \quad (10)$$

Both the actual and estimated SNR values are transformed into ξ and $\hat{\xi}$, respectively, and the model is trained by minimizing the L2 loss between them.

Once the SNR is estimated, it is mapped to the corresponding diffusion timestep using:

$$\hat{t} = \min(1, 10^{\frac{-\hat{\text{SNR}}(y_0)+\eta}{20}}). \quad (11)$$

Using the estimated SNR, the input noisy signal is normalized according to the following equation:

$$y_{\text{norm}} = H\left(\frac{H^{-1}(y_0)}{\max(H^{-1}(y_0))}\sqrt{\frac{1 + 10^{-\hat{\text{SNR}}(y_0)/10}}{1 + 10^{-\eta/10}}}\right). \quad (12)$$

Finally, the initial state $x_{\hat{t}}$ is constructed as:

$$x_{\hat{t}} = y_{\text{norm}} + \sigma(\hat{t})z, \quad z \sim \mathcal{N}(0, I). \quad (13)$$

These values, $\hat{t}$ and $x_{\hat{t}}$, are then used as the timestep and input for the reverse diffusion process during inference.

4. Experiments

In this section, we evaluate our proposed method in two different settings: (1) realistic speech enhancement evaluated on the VoiceBank-DEMAND test set and (2) an SNR-specific experiment to examine the robustness of the model across different SNR levels. The source code for our implementation is accessible on GitHub[1].

4.1. Dataset

The datasets used in this study consist of three parts: the main training dataset, the auxiliary training dataset, and the test dataset. The main training dataset, derived from the VoiceBank-DEMAND [20] training set, was used to train the enhancement model and consists of clean and noisy signal pairs. The noisy signals were augmented to have single SNR values, specifically $\eta = [0, 5, 10]$, to cover varying noise conditions. The clean signals remained unchanged. The auxiliary training dataset, also based on the VoiceBank-DEMAND training set, was designed to train the SNR estimation model. Noisy signals were created by augmenting the data to cover a wide range of SNR values, sampled between -60 dB and infinity (∞). These SNR values were normalized to the range [0, 1] to ensure compatibility with the loss function. For testing, two test sets were used: the original VoiceBank-DEMAND test set, used without modification

[1] https://github.com/yh-jun/SNR-Aligned_diffSE

for real-world noise conditions, and an augmented version with SNR values ranging from -5 to 35 dB in increments of 5 dB, allowing for a performance evaluation across varying noise levels.

4.2. Baselines and evaluation metrics

Various baselines were considered to evaluate the relative performance of the proposed model. SE-Bridge [14] was selected as the primary baseline due to its relevance to the proposed framework and its competitive performance. Additional baselines include SGMSE+ [12] and StoRM [13], both based on score-based generative models, providing a comparison among various approaches.

The evaluation focused mainly on WB-PESQ [21], a perceptual quality metric closely aligned with human auditory perception. PESQ scores were computed across various SNR levels to evaluate the model's ability to generalize under various noise conditions. Additionally, SI-SDR [22] and ESTOI [23] metrics were used to further evaluate the model's performance.

4.3. Implementation details

The enhancement model was trained with the diffusion timestep t ranging from $\epsilon = 0.001$ to $T = 1$. Timesteps $t_i \in \{t_1, t_2, \ldots, t_N\}$ were determined following the same setting as used in the consistency models [15], with $N = 30$ as the total number of timesteps. For the SNR estimation model f_{SNR}, we modified the PESQNet architecture [19] for efficiency by reducing the convolutional channels from 384 to 32 and merging three fully connected layers into a single layer. Both models were trained separately.

During inference, $\hat{t}$ is selected as the value in the t_i array that is closest to the value estimated by the SNR estimator. This ensures that the reverse diffusion process begins at an appropriate timestep while adhering to the trained set of timesteps.

5. Results

5.1. Evaluation on the original VoiceBank-DEMAND test set

To evaluate overall performance, the proposed method was tested on the original VoiceBank-DEMAND dataset. Six configurations, **M1–M6**, were used for evaluation, defined based on the chosen value η, and the approach of estimating SNR: whether an SNR estimator or an SNR oracle was used.

Table 1 presents the performance for all configurations and the baseline models. The results show that the proposed method outperforms all baselines for PESQ in configuration **M6**. In this evaluation, the number of NCSN++ function evaluations (NFEs) is used as a measure of computational cost during inference. Notably, **M6** with just a single NFE achieves superior performance compared to models evaluated with up to 30 NFEs.

The SNR estimator achieves an average error of 1.42 dB on the test set, and our enhancement model maintains stable performance for estimation errors within 5 dB. Performance degradation is observed only under larger errors, which are highly uncommon in real scenarios. This, along with the minimal performance gap between **M1–M3** and **M4–M6**, supports the practicality of the SNR estimator.

5.2. SNR-specific test results

The performance of the models **M4-M6** was further evaluated on the augmented version of the dataset, with SNR values ranging from -5 to 35 dB in increments of 5 dB. Table 2 presents the PESQ scores for each configuration across these SNR levels. The results show that the proposed method performs consistently well at all SNR levels, with **M6** achieving the best scores at most SNR values. Notably, the proposed method maintained high performance under low SNR conditions despite not being trained at those noise levels. This shows that the model generalizes effectively to unseen noise levels.

Model	NFEs	η	SNR	PESQ	ESTOI	SI-SDR
Mixture	30	-	-	1.97	0.79	8.4
SGMSE+ [12]	30	-	-	2.93	0.87	17.30
StoRM [13]	31	-	-	2.93	**0.88**	18.80
SE-Bridge [14]	**1**	-	-	2.97	0.87	18.95
M1	1	0	oracle	3.02	0.85	18.79
M2	1	5	oracle	3.07	0.86	19.22
M3	1	10	oracle	3.09	0.87	19.02
M4	**1**	0	estimated	3.02	0.85	18.83
M5	**1**	5	estimated	3.07	0.86	**19.23**
M6	**1**	10	estimated	**3.09**	0.87	19.01

Table 1: Average performance of our method and baselines on the VoiceBank-DEMAND test set, comparing their output to the noisy speech mixture. The best values in each column are bold.

Table 3 presents the SI-SDR scores for each SNR level. Existing approaches using SNR-related features, such as [7], tend to improve perceptual quality at the expense of reconstruction accuracy, leading to a decline in metrics such as SegSSNR [24]. In contrast, the proposed method not only improves perceptual scores but also achieves enhancements in reconstruction accuracy, as reflected by SI-SDR, at certain SNR levels. In outlier conditions, increased uncertainty makes it inevitable that distortion increases in order to maintain higher perceptual quality [25], especially in unseen low SNR ranges.

Model	-5	0	5	10	15	20	25	30	35
SE-Bridge [14]	1.95	2.33	2.72	3.09	3.41	3.68	3.90	4.07	4.19
M4	**2.11**	2.50	2.83	3.13	3.42	3.70	3.94	4.13	**4.27**
M5	2.07	**2.51**	2.89	3.20	3.46	3.72	**3.96**	**4.14**	4.26
M6	2.01	2.48	**2.90**	**3.22**	**3.51**	**3.75**	**3.96**	4.13	4.24

Table 2: PESQ scores by SNR level

Model	-5	0	5	10	15	20	25	30	35
SE-Bridge [14]	12.9	16.0	18.0	19.5	21.1	22.8	24.2	25.0	25.3
M4	**13.7**	**16.8**	18.3	19.2	20.2	21.2	22.5	23.6	24.4
M5	12.2	16.4	**18.6**	19.9	20.9	22.0	23.3	24.4	25.1
M6	8.5	14.3	18.0	**20.2**	**21.7**	**23.1**	**24.6**	**25.9**	**26.6**

Table 3: SI-SDR scores by SNR level

6. Conclusion

This study presented a novel SNR-adaptive framework for the consistency model-based speech enhancement model. By aligning the diffusion timestep with the input SNR and incorporating SNR-adaptive Gaussian noise modulation, the model enhances robustness across diverse noise conditions. The experimental results showed an average PESQ increase of 0.12 over the baseline, with further improvements in challenging SNR scenarios. The proposed approach can be extended to other diffusion-based models that incorporate timestep conditioning and additive Gaussian noise, highlighting its potential to improve generalization and robustness in speech enhancement tasks.

7. Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-00554289).

8. References

[1] P. C. Loizou. *Speech enhancement: theory and practice*. CRC press, 2007.

[2] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In *Latent Variable Analysis and Signal Separation: 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings 12*, pages 91–99. Springer, 2015.

[3] R. Ahmad, S. Zubair, and H. Alquhayz. Speech enhancement for multimodal speaker diarization system. *IEEE Access*, 8:126671–126680, 2020.

[4] P. Wang, K. Tan, et al. Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:39–48, 2019.

[5] S.-W. Fu, Y. Tsao, and X. Lu. SNR-aware convolutional neural network modeling for speech enhancement. In *Interspeech*, pages 3768–3772, 2016.

[6] X. Ge, J. Han, Y. Long, and H. Guan. Percepnet+: A phase and SNR aware percepnet for real-time speech enhancement. In *Interspeech 2022*, pages 916–920, 2022.

[7] R. Rehr and T. Gerkmann. SNR-based features and diverse training data for robust DNN-based speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1937–1949, 2021.

[8] S. Pascual, A. Bonafonte, and J. Serrà. SEGAN: Speech enhancement generative adversarial network. In *Interspeech 2017*, pages 3642–3646, 2017.

[9] 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)*, pages 659–666. IEEE, 2021.

[10] 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)*, pages 7402–7406. IEEE, 2022.

[11] S. Welker, J. Richter, and T. Gerkmann. Speech enhancement with score-based generative models in the complex STFT domain. In *Interspeech 2022*, pages 2928–2932, 2022.

[12] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2351–2364, 2023.

[13] J.-M. Lemercier, 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*, 2023.

[14] Z. Qiu, M. Fu, F. Sun, G. Altenbek, and H. Huang. Sebridge: Speech enhancement with consistent brownian bridge. *arXiv preprint arXiv:2305.13796*, 2023.

[15] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32211–32252, 2023.

[16] B. Lay and T. Gerkmann. An analysis of the variance of diffusion-based speech enhancement. In *Interspeech 2024*, pages 2205–2209, 2024.

[17] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[18] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[19] Z. Xu, M. Strake, and T. Fingscheidt. Deep noise suppression maximizing non-differentiable PESQ mediated by a non-intrusive PESQNet. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1572–1585, 2022.

[20] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi. Investigating RNN-based speech enhancement methods for noise-robust text-to-speech. In *Proceedings of the ISCA Speech Synthesis Workshop*, pages 146–152, 2016.

[21] 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 the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 749–752, 2001.

[22] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey. SDR–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019.

[23] 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*, 24(11):2009–2022, 2016.

[24] T. Lotter and P. Vary. Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model. *EURASIP Journal on Advances in Signal Processing*, 2005:1–17, 2005.

[25] R. Cohen, I. Kligvasser, E. Rivlin, and D. Freedman. Looks too good to be true: An information-theoretic analysis of hallucinations in generative restoration models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.