# LDMSE: Low Computational Cost Generative Diffusion Model for Speech Enhancement

Yuki Nishi*,Koichi Shinoda† and Koji Iwano‡
* Tokyo Institute of Technology, Tokyo
E-mail: y-nishi@ks.c.titech.ac.jp
† Tokyo Institute of Technology, Tokyo
E-mail: shinoda@c.titech.ac.jp
‡ Tokyo City University, Tokyo
E-mail: iwano@tcu.ac.jp

*Abstract*—**Recently, a generative model called diffusion model has attracted attention. Compared to GANs, it can be trained stably but has a high computational cost in the generation stage. This paper proposes a method called Low computational cost Generative Diffusion Model for Speech Enhancement (LDMSE). It reduces its computational cost with comparable quality by compressing speech signals to a latent space using an autoencoder and removing noise in the latent space using the diffusion model. In our evaluation using VOICBANK-DEMAND and WSJ0-CHiME3 datasets, the proposed method reduced the generation time by more than 35% without any degradation in speech quality.**

## I. Introduction

Speech enhancement technology for noise reduction is needed in various situations. These include speech assistance with hearing aids, noise reduction in speech recognition, and processing for speech-containing content on the Web. In recent years, neural network-based technologies have been widely used to enhance speech with higher quality. Among neural network (NN) techniques, the Diffusion Model has attracted much attention because of its high quality in generation tasks and stable training. It has also been studied in the field of speech enhancement. However, this method has a drawback in that it is computationally expensive due to the need to pass data through the network repeatedly during the generation phase.

In this study, we propose a method called Low computational cost generative Diffusion Model for Speech Enhancement(LDMSE), aiming to reduce the computational cost of the generation stage in the diffusion model without sacrificing enhancement quality. The input speech data is dimensionally reduced into a latent space using an Encoder and Decoder. Then, the computational cost of each generation step is reduced accordingly. In our evaluation, we proved that LDMSE reduces the total computational cost by more than 35% without any degradation in speech quality.

## II. Previous studies

### A. Speech Enhancement

Speech enhancement has been utilized for various situations. In this study, we estimate clean speech data $x$ given noisy speech $y = x + n$, a mixture of environmental noise $n$ and $x$. Additionally, we only consider single-channel data. Recently, NN-based methods have achieved higher quality than those without using NN, such as the method of suppressing certain frequency regions [1] and the method of modeling human voices with hidden Markov models [2]. Some methods directly estimate a mask in the frequency domain [3], [4]. For example, Kingma et al. [5] utilize Variational Autoencoder (VAE) for this purpose. Moreover, Generative Adversarial Networks (GAN) have achieved high quality in speech enhancement [6]. For instance, MetricGAN+ [7] applies GAN to speech enhancement. However, GAN is unstable in its training. The following diffusion model solves this problem and achieves better quality.

### B. Diffusion Model

One of the most popular generative models is the diffusion model. It models the diffusion process of the target data to which Gaussian noise is added and regenerates the original clean data from the target noisy data reversely. Diffusion models are divided into two categories: score-independent and score-based diffusion models.

The score-independent diffusion model, the so-called Denoising Diffusion Probabilistic Model (DDPM) [8], estimates a Gaussian noise added to the input data. It first defines the noise addition formula with $N_{\mathrm{DDPM}}$ steps. The transition of data from $n$ to $n+1$ $(0 \leq n \leq N_{\mathrm{DDPM}} - 1)$, which we call it forward direction, is defined as $x_{n+1} = \sqrt{1-\beta_n} x_n + \sqrt{\beta_n} z$, where $\beta_n$ is a hyperparameter and $z$ is sampled from normal Gaussian distribution. The transition from $x_{n+1}$ to $x_n$ is also computed explicitly, which is the same as the inference process for, for example, generating images and sounds.

The score-based diffusion model [9] is formulated based on the stochastic differential equation (SDE). First, Gaussian noise diffusion is formulated as:

$$\mathrm{d}x_t = f(x_t)\mathrm{d}t + g(t)\mathrm{w}(t), \quad 0 \leq t \leq T \tag{1}$$

where $x_t$ is the state of the data at $t$ and $x_0$ corresponds to the clean data, and $\mathrm{w}(t)$ is a Wiener process. The functions $f, g$ are defined in various ways depending on the experimental setup. The SDE in the reverse direction is formulated as [9], [10]:

$$\mathrm{d}x_t = \left[-f(x_t) + g(t)^2 \nabla_{x_t} \log p_t(x_t)\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathrm{w}}(t) \quad (2)$$

where $0 \leq t \leq T$ and $\nabla_{x_t} \log p_t(x_t)$ is called a score, $p_t(x_t)$ is the probability distribution of $x_t$, and $\bar{\mathrm{w}}(t)$ is the standard Wiener process when $t$ is traced backward from $T$ to 0. The score is computed from $f$ and $g$.

Several score-based diffusion models have been proposed to obtain $x_0$ by simulating backward transitions from noisy data $x_T$ [10]. The simplest method is Euler-Maruyama method [11], which discretizes the range from 0 to $T$ into $N$ steps. Another high-quality method is the Predictor-Corrector (PC) samplers [9] where the process is divided into a Predictor and a Corrector. The Predictor solves the SDE, and the Corrector refines it to achieve high quality.

### *C. Reduction of Computational Cost of Diffusion Model*

In the diffusion model, the process of gradually removing noise by passing the data through NN multiple times (e.g., 30 times for SGMSE+ [12], which is the baseline for this study) during the generation stage during the generation stage results in high computational cost. Several studies solve this problem by reducing the number of steps in the generation phase. As for DDPM, Denoising Diffusion Implicit Models (DDIM) [13] generalizes the definition formula of the diffusion process so that the inverse process is designed with arbitrary step intervals.

This approach that reduces the number of steps cannot be applied to the score-based diffusion model because its reverse process simulates SDEs during inference. Instead, Fast sampler [14] trains an NN sampler, which is applied to various diffusion models. Denoising MCMC (DMCMC) [15] introduces a relatively small NN that estimates the noise magnitude at each step. These two methods can generate with a small number of steps, though they slightly degrade the quality.

### *D. Speech Enhancement with Diffusion Model*

Several studies have applied the diffusion model to speech enhancement. However, the widely used DDPM [8] removes only Gaussian noise, which limits its applicability to speech enhancement tasks where the noise types are various, such as the crowds at train stations and birdsong [16]. CDiffuSE [17] addresses this limitation by incorporating not only Gaussian noise but also non-Gaussian noise (e.g., environmental noise) in the forward step.

A study of speech enhancement using the score-based diffusion model is SGMSE+ [12], which is used as the baseline in this study. In this method, the functions $f$ and $g$ in Eq. 2 are defined as

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

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

$$\sigma_r = \sigma_{\max}/\sigma_{\min}. \quad (5)$$

Then, the score in Eq.2 can be derived as follows.

$$\nabla_{x_t} \log p_t(x_t) = -\frac{x_t - \mu(x_0, y, t)}{\sigma(t)^2} \quad (6)$$

$$\mu(x_0, y, t) = e^{-\gamma t} x_0 + (1 - e^{-\gamma t} y) \quad (7)$$

$$\sigma(t)^2 = \frac{\sigma_{\min}^2 \left(\sigma_r^{2t} - e^{-2\gamma t}\right) \log\left(\sigma_r\right)}{\gamma + \log(\sigma_r)}, \quad (8)$$

where $y$ is the sound mixed with speech and environmental noise, $\sigma_{\max}$ and $\sigma_{\min}$ are hyperparameters set to 0.5 and 0.05 respectively. $\gamma$ is a hyperparameter to control the strength of the transition ("hardness") from $x_0$ to $y$ and is set to 1.5. By defining the hyperparameters in this way, we obtain

$$p_{0t}(x_t | x_0, y) = N_C(x_t; \mu(x_0, y, t), \sigma(t)^2 \mathrm{I}) \quad (9)$$

where the center of $x_t$ asymptotically moves toward $y$ as $t$ increases. See [12] and [18] for a more detailed explanation of these formulas.

This SGMSE+ [12] achieves higher quality than CDiffuSE [17]. Authors in SGMSE+ [12] suggest its reason is the architecture and fine hyperparameter tuning.

### *E. Diffusion model with latent space*

The term "latent space" in AutoEncoder [19] and VAE [5] refers to the space formed by the Encoder's mapping of data. Latent Diffusion Model (LDM) [20] is a diffusion model that applies the diffusion process to the latent space formed by a VAE for image generation. It assumes that the image generation process is divided into Perceptual stage with a low compression ratio and a Semantic stage with a high compression ratio. The dimension of the latent space is compressed in the spatial direction and expanded in the channel direction. The total dimension becomes smaller because the reduction rate of the spatial dimension is greater than the expansion rate of the channel dimension.

AudioLDM [21] uses LDM to process audio data. This model is designed for Text To Audio (TTA) tasks. In this method, the diffusion model is executed in the latent space defined by VAE. It achieved State-Of-Art in the TTA. The architecture of Encoder and Decoder used in AudioLDM is shown in Fig.1. Another study [22] utilizes LDM for text-to-speech tasks and achieved state-of-the-art performance. LSGM [23] combines a score-based diffusion model with VAE for image generation tasks. Our study differs from these studies in that it is developed for speech enhancement, in which speech quality is directly evaluated by some metrics such as PESQ and ESTOI.

## III. LDMSE

### *A. Use of Encoder-Decoder*

We propose a method to reduce the computational cost of the diffusion model using a latent space with a smaller dimension than that of the original input data. We call this method Low computational cost Generative Diffusion Model for Speech Enhancement (LDMSE). This method has an encoder before the diffusion process and a decoder after it (hereafter, these two

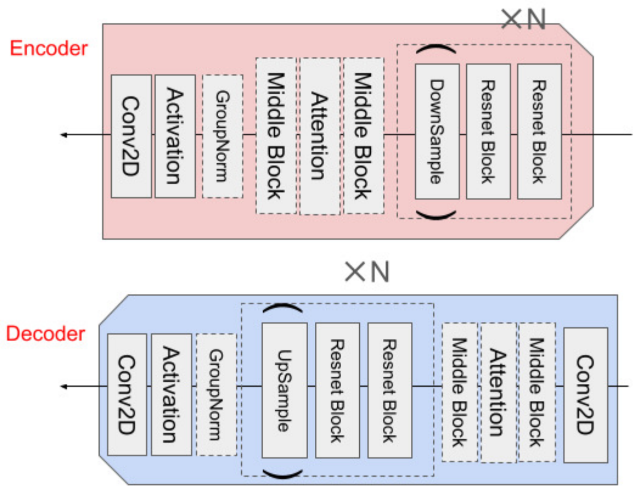

Fig. 1. Encoder-Decoder Architecture. Resnet Block is the same as AudioLDM [21]. Blocks with dashed lines mean they have been removed. Dashed rectangles denote processing groups that are executed N times. The Up/Down Sample layers in parentheses with dashed rectangles are not executed in the final N times loop.

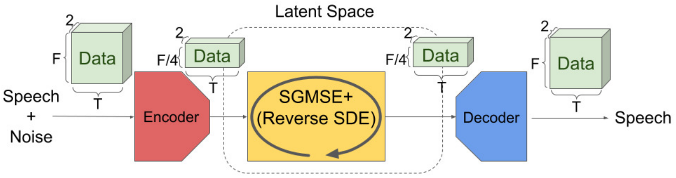

Fig. 2. Proposed process when $\xi = 1/4$

are abbreviated as EncDec). If the dimensionality of the data handled in the inverse diffusion is smaller, the computational cost will be reduced, which is the purpose of the processing process.

Let $\xi$ represent the compression rate. Fig. 2 shows the proposed process flow when $\xi$ is $1/4$. Also, $F$ and $T$ represent the number of frequency and time dimensions after STFT, respectively. The theoretical computational cost required for reverse diffusion is $\xi$ $(< 1)$ times lower than the Baseline.

In this study, EncDec is not a VAE but an ordinary AutoEncoder. We applied a tanh function to the Encoder's output because we found empirically that the variation of the data in the latent space would be very different depending on the seed value at the time of training if we did not do this.

The training process of LDMSE is divided into two phases. The first phase is to train an encoder to project data into the latent space and a decoder to restore it from the latent space to the real space. Only clean speech data were used for training. And loss function was L2 loss. The second stage is to fix the parameters of the EncDec and train the Diffusion Model to execute the inverse diffusion for the inference.

The EncDec was trained with 100 epochs, 8 minibatch size, and the Diffusion part was done with 300 epochs, 16 minibatch size. The learning rate in both trainings is $8.0 \times 10^{-6}$.

## B. Architecture

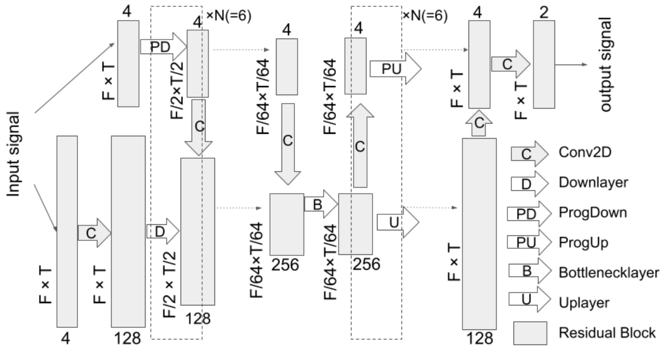

Fig. 3. Model architecture for the Diffusion part. The dimension of the input signal is $4 \times F \times T$. The numbers above or below the rectangles representing the Residual Blocks indicate the channel dimension. Four on the axis of the channel in the input dimension means the real and imaginary parts of the clean and noisy data, respectively. Its details are described in [12].

*1) Diffusion part:* We adopt the same architecture NCSN++ that is proposed in SGMSE+ [12] (Fig. 3). It has a similar structure to U-Net [24]. The components in Fig. 3, Residual Block, Uplayer, Downlayer, Bottlenecklayer, ProgUp, ProgDown, and Bottlenecklayer, are the same as used in NCSN++. The input and output signals are almost the same as the signals processed by STFT, but with scaling described in Sec. 3.2 in [18].

Depending on $\xi$, dimension-related issues may arise during the processing of NCSN++. For example, if $\xi = 1/8$, the dimension of the input data of the diffusion model becomes $F/8 \times T$, where $F \times T$ is the dimension before processing by the Encoder. Since $F = 256$ and $N = 6$ in the original SGMSE+, the dimension of the data processed in the last iteration of the iterative structure on the left side in the Fig. 3 becomes $(256/8)/(2^6) < 1$, indicating that the vector dimension becomes zero. In such cases, we refrain from further dimension reduction.

*2) AutoEncoder part:* The architecture of the EncDec module is adapted from AudioLDM [21]. It is shown in Fig. 1. However, the EncDec architecture utilized in LDMSE incorporates three modifications compared to AudioLDM. First, there are alterations in the architecture within the EncDec module itself. The middle block and normalization layer used in AudioLDM outside the Resnet block loop have been removed. The middle block calculates Attention for the time and frequency dimensions. This is because the computational cost of this block was very high when we used STFT as input, whereas AudioLDM used a mel-spectrogram. In our preliminary experiments, we found that EncDec's accuracy without the normalization layer was better.

Second, UpSample and DownSample are applied only to the frequency dimension and not to the time dimension. NCSN++ used in the diffusion process also compresses the time dimension. In the case of $\xi = 1/4$, the time dimension is compressed up to $1/256$ at minimum. For example, if we input data with a time dimension of 65, the time dimension becomes 32 after the first compression, as shown in Fig. 3, and 64 after the restoration. Then, for $\xi = 1/4$, the time dimension in the output is a multiple of 256, corresponding to 2 seconds. This is too long to have a desirable time resolution in the output.

Third, the number of loops in the ResNet Block varies with $\xi$. For $\xi = 1/2, 1/4, 1/8$, $N = 2, 3, 4$ in Fig. 1. This is to make this architecture compatible with multiple $\xi$. This architecture consists of a stack of layers, each of which is a set of Down/Up Sample, Resnet Block, and Conv2D in the frequency dimension. The data is down/up sampled with each layer. Therefore, to support multiple $\xi$, the number of layers is changed.

## IV. Experiment

### A. Dataset

For our evaluation, we employ two datasets: VOICEBANK-DEMAND and WSJ0-CHiME3.

VOICEBANK-DEMAND serves as a benchmark, aligning with previous studies of speech enhancement. This is a mixture of VOICEBANK [25] and DEMAND [26]. VOICEBANK has two sets, one with 28 speakers and the other with 56 speakers, of which 28 speakers are selected. Each speaker has about 400 utterances. There are a total of 11572 utterances for training and 824 for evaluation. In the training data, each noise is mixed with one noise per utterance, with the signal-to-noise ratios of 15 dB, 10 dB, 5 dB, and 0 dB for the training data and 17.5 dB, 12.5 dB, 7.5 dB, and 2.5 dB for the evaluation data, respectively. Each utterance is used exactly once.

WSJ0-CHiME3 is utilized for comparison to SGMSE+. This is a mixture of WSJ0 [27], comprising readings from the Wall Street Journal, and CHiME3 [28], prepared in the 3rd CHiME Challenge, a speech recognition competition and hereafter abbreviated as WSJ-C3. In WSJ0, the subset si_tr_s with a total of 12,776 utterances was used for training, and a subset of si_dt_05 with 1,026 utterances was used for the valid split. The signal-to-noise ratios are uniformly distributed from 0 dB to 20 dB. Each utterance is used only once. WSJ0-CHiME3 is divided into train, valid, and test sets. This paper presents the results of the evaluation of the Valication set.

We use the same consistent datasets for training and evaluation, not changing the dataset in these two stages.

### B. Metrics

The Baseline for this research is SGMSE+ [12]. We mainly compare the quality with it. For the indicators of how close the emphasized speech is to the clean speech, we measured PESQ [29], ESTOI [30], SI-SDR [31]. All of these mean that the higher the value, the better the quality.

- PESQ [29]: The Perceptual Evaluation of Speech Quality. recommended by ITU-T P. 862 in 2001. The value range is from 1.0 to 4.5.
- ESTOI [30]: The Extended Short-Time Objective. The Extended Short-Time Objective Intelligibility. It takes a value range from 0.0 to 1.0.
- SI-SDR [31]: Scale-Invariant Signal-to-Distortion Ratio. It is a measure of sound distortion calculated in a way that is invariant to the volume of the two data being compared.

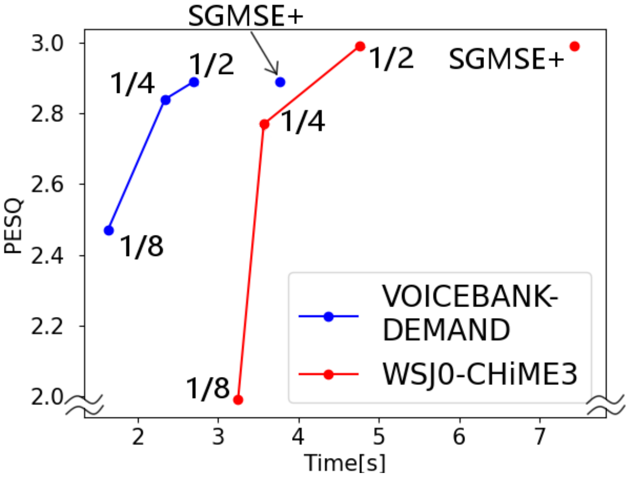

Fig. 4. Line graph for computational time and PESQ. ”Time” indicates the average computational time required for speech enhancement per utterance, including the processing time of EncDec. Fractions represent the value of $\xi$.

We also measure the time that the GPU uses for generation. We use torch.cuda.Event for recording time. The GPU we used for generation is one RTX 3090. The CPU is Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz. The RAM has 32GB.

### C. Results

TABLE I
Results on VB-DMD. ”Time” in the table indicates the average time required for speech enhancement per utterance, including the processing time of EncDec. $\xi = 1/2, 1/4, 1/8$ are proposed method, LDMSE. Values after ± mean standard deviation. MGAN+ means MetricGAN+. The results of metrics and generation time in MGAN+ and CDiffuSE are from our preliminary experiment because [12] does not show the standard deviation of results.

| | PESQ↑ | ESTOI↑ | SI-SDR↑ | Time[s] |
|---|---|---|---|---|
| SGMSE+ | 2.89±0.60 | 0.86±0.10 | 16.8±3.4 | 3.77 |
| $\xi = 1/2$ | 2.89±0.63 | 0.86±0.10 | 17.2±3.0 | 2.70 |
| $\xi = 1/4$ | **2.84±0.65** | **0.86±0.10** | **17.3±2.9** | **2.34** |
| $\xi = 1/8$ | 2.47±0.62 | 0.81±0.12 | 14.8±3.3 | 1.63 |
| CDiffuSE | 2.53±0.57 | 0.77±0.11 | 12.4±2.7 | 0.033 |
| MGAN+ | 3.15±0.57 | 0.82±0.12 | 8.0±3.2 | 0.0012 |

TABLE II
Results on WSJ-C3. The meanings of each value and word are the same as in VB-DMD. The results of metrics in MGAN+(=MetricGAN+) and CDiffuSE are reported value in [12].

| | PESQ↑ | ESTOI↑ | SI-SDR↑ | Time[s] |
|---|---|---|---|---|
| SGMSE+ | 2.99±0.55 | 0.91±0.07 | 18.2±4.3 | 7.43 |
| $\xi = 1/2$ | **2.99±0.55** | **0.91±0.06** | **17.6±4.0** | **4.76** |
| $\xi = 1/4$ | 2.77±0.57 | 0.90±0.07 | 17.2±3.5 | 3.57 |
| $\xi = 1/8$ | 1.99±0.41 | 0.84±0.09 | 12.4±3.6 | 3.25 |
| CDiffuSE | 2.27±0.51 | 0.83±0.09 | 9.2±2.3 | 0.099 |
| MGAN+ | 3.03±0.45 | 0.88±0.08 | 10.5±4.5 | 0.0024 |

Fig. 4 shows the comparison of our proposed methods, LDMSE, with different $\xi$ values. For $\xi = 1/2$, our method succeeded in reducing the time without losing quality in VB-DMD, and in $\xi = 1/4$ with WSJ-C3 it did not lose quality in the $\xi = 1/2$ case, nor did it lose much. In these $\xi$, the effectiveness of the proposed method was confirmed. For $\xi = 1/8$, the quality has been noticeably reduced for both data sets. This may be because the degree of compression was too large, and thus, the necessary dimension in the real space was not obtained in the latent space.

Table I and II show the detailed results of the quality evaluation of speech enhancement for VB-DMD and WSJ-C3, respectively. The PESQ and Time values in Table I and II are the same as the values in Fig. 4. Results of MetricGAN+ and CDiffuSE are shown for comparison. With SI-SDR and ESTOI, up to $\xi = 1/4$, the speedup was achieved without losing quality. The bold values in Table I, II are those considered to have been particularly successful. Compared to the other methods, LDMSE is slower than CDiffuSE in generation time but better in quality. Compared to CDiffuSE, LDMSE outperforms in quality in SI-SDR and ESTOI, while MetricGAN+ outperforms in quality in PESQ. This may be because the objective of MetricGAN+ during training is to maximize PESQ. In all quality metrics, there is no significant difference in t-test at a 5% significance level.

## V. Conclusion

We have proposed a method called LDMSE to reduce the computational cost of diffusion model based speech enhancement, it is shown that the quality of the diffusion model with SGMSE+ can be maintained while reducing the computational cost by using the latent space. In particular, if it is $\xi = 1/4$ in VB-DMD and $\xi = 1/2$ in WSJ-C3 we can achieve a speedup of more than 35% without quality loss.

In the future, we would like to aim to use VAE for EncDec [20], [21]. Additionally, we would like to investigate the degree of data dispersion in latent space. Finally, we want to examine how we can apply the proposed method to other diffusion model-based denoising methods, such as CDiffuSE, MGAN+, and complex number-based frameworks.

## References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979. DOI: 10.1109/TASSP.1979.1163209.

[2] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989. DOI: 10.1109/5.18626.

[3] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016. DOI: 10.1109/TASLP.2015.2512042.

[4] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," Sep. 2017, pp. 1–6. DOI: 10.1109/MLSP.2017.8168119.

[5] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. arXiv: http://arxiv.org/abs/1312.6114v10 [stat.ML].

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14, Montreal, Canada: MIT Press, 2014, pp. 2672–2680.

[7] S.-W. Fu, C. Yu, T.-A. Hsieh, *et al.*, "Metricgan+: An improved version of metricgan for speech enhancement," Aug. 2021, pp. 201–205. DOI: 10 . 21437 / Interspeech.2021-599.

[8] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20, Vancouver, BC, Canada: Curran Associates Inc., 2020, ISBN: 9781713829546.

[9] 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*, 2021. [Online]. Available: https://openreview.net/forum?id=PxTIG12RRHS.

[10] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982, ISSN: 0304-4149. DOI: https://doi.org/10.1016/0304-4149(82)90051-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0304414982900515.

[11] P. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations* (Stochastic Modelling and Applied Probability). Springer Berlin Heidelberg, 2011, ISBN: 9783540540625. [Online]. Available: https://books.google.co.jp/books?id=BCvtssom1CMC.

[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*, vol. 31, pp. 2351–2364, 2023. DOI: 10.1109/TASLP.2023.3285241.

[13] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=St1giarCHLP.

[14] D. Watson, W. Chan, J. Ho, and M. Norouzi, "Learning fast samplers for diffusion models by differentiating through sample quality," in *International Conference*

*on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=VFBjuF8HEp.

[15] B. Kim and J. C. Ye, "Denoising MCMC for accelerating diffusion-based generative models," in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., ser. Proceedings of Machine Learning Research, vol. 202, PMLR, 23–29 Jul 2023, pp. 16 955–16 977. [Online]. Available: https://proceedings.mlr.press/v202/kim23z.html.

[16] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," in *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 2016, pp. 146–152. DOI: 10.21437/SSW.2016-24.

[17] 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)*, 2022, pp. 7402–7406. DOI: 10.1109/ICASSP43922.2022.9746901.

[18] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Proc. Interspeech 2022*, 2022, pp. 2928–2932. DOI: 10.21437/Interspeech.2022-10653.

[19] D. Bank, N. Koenigstein, and R. Giryes, *Autoencoders*, Mar. 2020.

[20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 10 684–10 695.

[21] H. Liu, Z. Chen, Y. Yuan, *et al.*, "AudioLDM: Text-to-audio generation with latent diffusion models," in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., ser. Proceedings of Machine Learning Research, vol. 202, PMLR, 23–29 Jul 2023, pp. 21 450–21 474. [Online]. Available: https://proceedings.mlr.press/v202/liu23f.html.

[22] Z. Liu, Y. Guo, and K. Yu, "Diffvoice: Text-to-speech with latent diffusion," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095100.

[23] A. Vahdat, K. Kreis, and J. Kautz, "Score-based generative modeling in latent space," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=P9TYG0j-wtG.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241, ISBN: 978-3-319-24574-4.

[25] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013, pp. 1–4. DOI: 10.1109/ICSDA.2013.6709856.

[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. 035 081, Jun. 2013, ISSN: 1939-800X. DOI: 10.1121/1.4799597. eprint: https://pubs.aip.org/asa/poma/article-pdf/doi/10.1121/1.4799597/14782232/035081_1_online.pdf. [Online]. Available: https://doi.org/10.1121/1.4799597.

[27] e. a. P. L. D. C. Garofolo John S., *Csr-i (wsj0) complete ldc93s6a. web download*, Sep. 1993. DOI: 10.1109/MLSP.2018.8516711.

[28] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511. DOI: 10.1109/ASRU.2015.7404837.

[29] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, 749–752 vol.2. DOI: 10.1109/ICASSP.2001.941023.

[30] 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. DOI: 10.1109/TASLP.2016.2585878.

[31] J. L. 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)*, 2019, pp. 626–630. DOI: 10.1109/ICASSP.2019.8683855.