Interspeech 2024
1-5 September 2024, Kos, Greece

SWiBE: A Parameterized Stochastic Diffusion Process for Noise-Robust Bandwidth Expansion

Yin-Tse Lin[1], *Shreya G. Upadhyay*[2], *Bo-Hao Su*[2], *Chi-Chun Lee*[1,2]

[1]Institute of Communication Engineering, National Tsing Hua University, Taiwan
[2]Department of Electrical Engineering, National Tsing Hua University, Taiwan

alexanderlin890625@gapp.nthu.edu.tw, shreya@gapp.nthu.edu.tw, borrissu@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw

Abstract

Speech recordings frequently encounter a variety of distortions, making the task of eliminating them essential yet challenging. In this study, leveraging the current success of score-based generative modeling (SGM), we propose a novel noise-robust bandwidth expansion (BWE) framework based on an innovative parameterized stochastic diffusion process, achieved through stepwise bandwidth expansion in the spectrogram. Our proposed Step-Wised Bandwidth Expansion (SWiBE) method outperforms baseline approaches over considered metrics, including the current state-of-the-art noise-robust BWE model and various diffusion and GAN-based models. Moreover, we analyze the interaction between the hyperparameters and performance across different aspects including perceptual quality and spectral reconstruction. Our findings reveal that the score-based model manifests distinct characteristics under varying parameterizations.

Index Terms: bandwidth expansion, speech enhancement, score-based generative modeling

1. Introduction

Bandwidth expansion (BWE), also referred to as audio super-resolution, is a task that aims at bridging the gap between audio signals at a low sampling rate and a high sampling rate. By restoring the lost high-frequency components, BWE reconstructs essential information, improving the intelligibility and quality of speech signals. Moreover, this restoration not only enhances the listening experience but also benefits downstream tasks such as automated speech recognition (ASR) and speech synthesis, where clear and well-formed speech signals are essential for accurate processing. Moreover, as the real-world speech recordings inevitably contend with a multitude of disturbances stemming from either interference or device artifacts, including background noise, reverberation, clipping, etc, ensuring the robustness of in-the-wild BWE performance has also become an important research direction [1].

When focusing on BWE, previous studies can be broadly classified into regression-based approaches (i.e., direct mapping methods) and generative methods. Regression-based methods have received attention due to their utilization of neural network architectures, aiming at address the temporal patterns of speech signals through various modules, either in the time or frequency domain [2, 3, 4, 1]. Conversely, generative methods do not directly establish a mapping between narrow-band input and wide-band output. Instead, they model the underlying patterns or distributions from a provided database and generate speech samples by emulating the probability distribution. Notable among generative methods are *generative adversarial networks* (GANs) [5, 6] and diffusion probabilistic models [7, 8, 9, 10, 11], which have shown leading-edge performance and attracted considerable interest in the field.

Specifically, diffusion probabilistic models [12, 13, 14] are nowadays particularly popular for their exceptional performance in addressing a wide range of distortions within the speech enhancement domain [15, 16, 17]. For instance, Serr‘a et al. [18] introduced a universal speech enhancer employing score-based diffusion to address 55 different distortions concurrently, while Richter et al. [17] proposed a model for both speech enhancement and speech dereverberation, offering empirical insights into the diffusion process along with a theoretical exploration. These studies have tested various forms of distortion on diffusion models and has become the current SOTA methods. In these methods, distinctive types of distortions are treated equally through similar diffusion processes. However, we argue that different types of distortion may involve distinct underlying diffusion processes. In other words, to align with the unique characteristics observed in each type of distortion, tailored diffusion processes should be devised. For instance, in the context of BWE, band limitation poses a distinctive challenge by truncating specific information within the speech signal, in contrast to noise addition. Directly applying the commonly used diffusion process for noise removal might not yield the optimal solution. This is because the diffusion process designed for general purpose enhancement would lead models to fill the lost high-frequency parts in the first few steps and then refine the high-frequency components gradually in the subsequent steps. Instead, a more intuitive strategy would involve progressively filling the frequency band from low-to-high in a step-by-step manner.

In this work, based on *score-based generative modeling* (SGM), we propose SWiBE (Step-Wised Bandwidth Expansion), a parameterized stochastic diffusion process specifically designed for noise-robust BWE. SGM emerges as the SOTA diffusion model, involving a process to maximize the log-likelihood of data points $\mathbf{x}$ within an underlying data distribution $p(\mathbf{x})$. Our approach under the SGM framework delineates a process from narrow-band to wide-band, making the diffusion process within the data distribution $p(\mathbf{x})$ become a shift across different sampling rates. We evaluate our method on the noise-robust BWE task (8k to 16k) and our results demonstrate superior performance across all considered metrics compared to both the generative model baselines and the current SOTA noise-robust BWE model [1]. Additionally, we parameterize the diffusion process in order to control the overall expansion trajectory. With the parameterization, we conduct an experimental analysis to explore the characteristic differences as a function of parameter choices. Our experimental result reveals clear trends in metrics of perceptual quality and spectral reconstruction under different parameterizations.

10.21437/Interspeech.2024-463

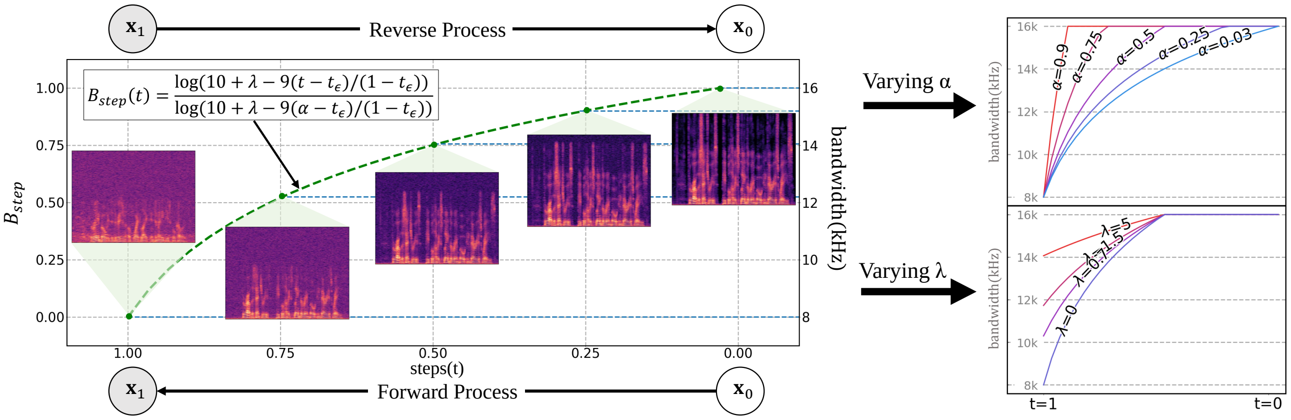


Figure 1: *A visualization of the proposed B_{step} in the diffusion process of SWiBE, where $\mathbf{x}_1$ and $\mathbf{x}_0$ denotes to noise perturbed 8k signal and clean 16k siganl, respectively.*

2. Approach

In this section, we describe our proposed SWiBE by elucidating the design of the stochastic diffusion process, with a focus on the adaptation made for BWE.

2.1. Modeling backbone

In this work, we choose SGMSE [17, 19], an SGM-based speech enhancement and dereverberation framework, as the backbone. SGM [20, 14], served as the fundamental structure behind SGMSE, is a category of diffusion models designed to model the gradient of log probability density function $\bigtriangledown_{\mathbf{x}} \log p(\mathbf{x})$, i.e., the score function, from the data distribution $p(\mathbf{x})$. The overall generation process involves calculating the score function while simultaneously perturbing the training data with noise, resulting in a gradual shift on $p(\mathbf{x})$ to maximize the log-likelihood. Consequently, the score-based model $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t) \approx \bigtriangledown_{\mathbf{x}} \log p_t(\mathbf{x})$ is defined as the score function predictor, where $p_t(\mathbf{x})$ is the marginal distribution under different levels of noise perturbation on $p(\mathbf{x})$.

Song et al. in [14] models the noise perturbation procedure in SGM as a continuous time stochastic process and formulated it into a *stochastic differential equation* (SDE). In SGMSE, the SDE is rewritten with the introduction of the noisy speech signal y and can be expressed as:

$$d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, \mathbf{y})dt + g(t)d\mathbf{w} \tag{1}$$

where $\mathbf{x}_t$ is the current state on the distribution $p_t(\mathbf{x})$, indexed by the continuous time variable $t \in [0, T]$, $\mathrm{f}(\mathbf{x}_t, \mathbf{y})$ is a vector-valued function called the *drift* coefficient, $g(\mathbf{t})$ is a real-valued function called the *diffusion* coefficient, and $\mathbf{w}$ denotes a standard Brownian motion. In accordance with the SDE, a corresponding reverse SDE exists, which is used for sample generation. SGMSE defined the equation as follows:

$$d\mathbf{x}_t = [\mathbf{f}(\mathbf{x}_t, \mathbf{y}) - g^2(t) \bigtriangledown_{\mathbf{x}} \log p_t(\mathbf{x}_t|\mathbf{y})]dt + g(t)d\bar{\mathbf{w}} \tag{2}$$

where dt represents a negative infinitesimal time step from $t = T$ to $t = 0$. Utilizing the reverse SDE, the score-based model can generate samples by transforming the prior distribution $p_T(\mathbf{x})$ back into the data distribution $p_0(\mathbf{x})$.

2.2. Stochastic diffusion process for BWE

In SWiBE, we follow the formulation of SDE and reverse SDE in equation 1 and equation 2. The difference is that, in our work, the SDE delineates a function that gradually distorts $\mathbf{x}_t$ from clean 16k to noisy 8k with $\mathbf{x}_0$ and $\mathbf{y}$ representing the clean 16k signal and the noisy 8k signal, respectively. To achieve this objective, the drift coefficient $\mathbf{f}$ is defined as:

$$\mathbf{f}(\mathbf{x}_t, \mathbf{y}) := \gamma(F_b(\mathbf{y}) - F_b(\mathbf{x}_t)) \tag{3}$$

where F_b denotes an ideal low-pass filter with a cutoff frequency b, and γ is a constant called *stiffness*, determining the transition from $\mathbf{x}_0$ to $\mathbf{y}$. The diffusion coefficient g follows the definition of SGMSE which can be expressed as:

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

where $\sigma_{\min}$ and $\sigma_{\max}$ are parameters used to control the noise schedule. Consequently, once the score $\bigtriangledown_{\mathbf{x}} \log p_t(\mathbf{x}_t)$ of each marginal distribution $p_t(\mathbf{x}_t)$ is determined, we can derive the reverse diffusion process from the reverse SDE and simulate it to sample the clean 16k signal $\mathbf{x}_0$ from p_0.

It is noteworthy in equation 3 that as t approaches 0, b increases accordingly, making x_{t_1} an audio with higher-sampling-rate than x_{t_2} for $0 \leqq t_1 < t_2 \leqq 1$. This adjustment transforms the forward process into a noise perturbation and downsampling procedure while the reverse process becomes a denoising upsampling procedure as illustrated in the left-hand side of Figure 1. This may help the score-based model uncover the data distribution $p(x)$ consisting of audio samples with different sampling rates. In particular, b can be derived from the function $B_{step}(t)$ called *bandstep*, which is depicted as the green curve in Figure 1 and can be written as follows:

$$B(t) = \log(10 + \lambda - \frac{9(t - t_\epsilon)}{1 - t_\epsilon}) \tag{5}$$

$$B_{step}(t) = B(t)/B(\alpha) \tag{6}$$

$$b = \begin{cases} f_{tgt} * \frac{B_{step}(t)+1}{f_{tgt}/f_{src}} & \text{if} \quad B_{step}(t) <= 1 \\ f_{tgt} & \text{if} \quad B_{step}(t) > 1 \end{cases} \tag{7}$$

where $t \in [0, 1]$, t_ϵ is the minimum time in practical implementation, $\alpha \in [t_\epsilon, 1)$, $\lambda \in [0, \infty)$, and f_{tgt} and f_{src} denotes the target sampling rate and the source sampling rate for BWE, expressed in Hz. In our case, f_{tgt} and f_{src} are 16k and 8k, respectively, but they can be generalized to any chosen combination of source and target sampling rate. The variations of B_{step} over time steps according to α and λ are shown in the right-hand side of Figure 1. As illustrated, α and λ are used to regulate the overall expansion trajectory. Specifically, α governs the saturation speed, in which, as α increases, the score-based model accelerates its progression toward complete expansion, which makes the input audio expand rapidly from 8k to 16k and spend the remaining time on refining and denoising. On the other hand, λ controls the starting point of expansion; a higher value of λ results in a starting point with a higher frequency, thereby smoothing the expansion curve.

2.3. Score-based model training

The score-based model training is done by estimating the score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)$ with the use of denoising score matching [21, 22], involving minimizing the term

$$\mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_t|(\mathbf{x}_0,y)} w(t)[\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t,\mathbf{y},t) - \nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t|\mathbf{x}_0,\mathbf{y})\|_2^2] \quad (8)$$

over uniformly sampled $t \in [0,1]$, $\mathbf{x}_0 \sim p_0(\mathbf{x})$, and $\mathbf{x}_t \sim p_{0t}(\mathbf{x}_t|\mathbf{x}_0,\mathbf{y})$, where $w(t)$ is a positive weighting function. In the equation 8, $p_0(\mathbf{x})$ is already known, while the transition kernel $p_{0t}(\mathbf{x}_t|\mathbf{x}_0,\mathbf{y})$ can be solved by computing the closed-form solution of mean and variance of $\mathbf{x}_t$ with the equations mentioned in [23, Chapter 5-5]. Accordingly, we can sample $\mathbf{x}_t$ with the equations below:

$$\mathbf{x}_t = \boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},t) + \sigma(t)\mathbf{z} \quad (9)$$

$$\boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},t) = e^{-\gamma t} F_b(\mathbf{x}_0) + (1 - e^{-\gamma t}) F_b(\mathbf{y}) \quad (10)$$

$$\sigma(t)^2 = \frac{\sigma_{\min}^2((\sigma_{\max}/\sigma_{\min})^{2t} - e^{-2\gamma t}) log(\sigma_{\max}/\sigma_{\min})}{\gamma + log(\sigma_{\max}/\sigma_{\min})} \quad (11)$$

where $t \sim \mathcal{U}[t_\epsilon, 1]$ and $\mathbf{z} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{z}; 0, \mathbf{I})$.

Finally, the loss term of the score-based model, as shown in [17], can be rewritten from equation 8 to:

$$\arg\min_{\theta} E_{t,\mathbf{x}_0,\mathbf{x}_t|(\mathbf{x}_0,\mathbf{y})} w(t)[\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t,\mathbf{y},t) + \mathbf{z}/\sigma(t)\|_2^2] \quad (12)$$

where we select $w(t)$ to be $g(t)^2$ as suggested by Song et al. in [24]. In practical implementation, score-based model $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t,\mathbf{y},t)$ takes $\mathbf{x}_t$, t, and $\mathbf{y}$ as input, outputs the score in each marginal distribution, implying the relationship between different sampling rates, and finally updates the model using equation 12. Here, $\mathbf{y}$ serves as a conditioner, and to provide additional information for the score-based model, we append the high-frequency part with the low-frequency component for x_t when it has not fully expanded yet.

The network utilized in this study is NCSN++ [14], which is adapted for complex spectrogram input as described in [17]. It is a U-net-structured conditional network composed of several downsampling and upsampling residual blocks. The continuous time variable t is fed into the network with random Fourier feature embeddings [25].

3. Experiment

3.1. Database

In this study, we use the VoiceBank-DEMAND dataset [26] as the primary dataset. It is an English corpus comprises 28/2 speakers for the training/testing set. The dataset provides both the clean version and noisy version where the noisy one is mixed with noise sourced from DEMAND [27]. The training set has 40 noisy conditions, and the testing set has another 20 different noisy conditions. We downsample the clean utterances to 16k from the original 48k as our reconstruction target $\mathbf{x}_0$. Meanwhile, the noisy 8k signal $\mathbf{y}$ is generated by first downsampling the noisy utterance to 8k and then upsampling to 16k using sinc interpolation. For validation, we split speakers p226 and p227 from the training set. In our setup, we have an 8-hour 48-minute training set, a 34-minute testing set, and a 38-minute validation set, containing 108030, 824, and 742 pieces of speech, respectively.

3.2. Experimental setup

In our experiments, the training speeches are transformed into complex-valued spectrograms with an FFT length of 510 and a hop length of 128, which is then cropped into samples with 256 STFT frames. The model is optimized by Adam optimizer with a learning rate of 1e-4 and a batch size of 8 for 150 epochs. Exponential moving average (EMA) is applied to parameters when sampling [28], with an EMA rate of 0.999. γ in equation 3 is set to 1.5, while $\sigma_{\min}$ and $\sigma_{\max}$ in equation 4 are set to 0.05 and 0.5, respectively. A grid search is applied to B_{step} in equation 6 by varying α and λ from {0.03, 0.15, 0.25, 0.5} and {0.0, 0.7, 1.5, 5.0}, respectively. Our model in table 1 is selected with the best performance across all metrics. During inference, we utilize the Predictor-Corrector samplers [14] as the SDE solver, with a combination of reverse diffusion predictor and annealed Langevin dynamics corrector. The number of reverse steps is set to 30. Our framework is trained on NVIDIA A100 80GB, and model selection is based on validation performance. The time and memory costs for training and inference are approximately {24, 6} GBs and {30, 1.5} hours, respectively. Our source code can be found in the Github link[1].

3.3. Evaluation metrics

We employ several metrics for model evaluation. For reconstruction performance, SNR (Signal-to-Noise Ratio) assesses the waveform reconstruction, while LSD (Log Spectral Distance) evaluates the spectrogram reconstruction. PESQ [30] (Perceptual Evaluation of Speech Quality) gives the perceptual quality, and ESTOI [31] (Extended Short-Time Objective Intelligibility) represents speech intelligibility. CSIG, CBAK, COVL [32] are metrics often used for speech enhancement evaluation, which represents the mean opinion score (MOS) of speech distortion, intrusiveness of background noise, and overall processed speech quality, respectively. Within these metrics, the lower value of LSD is better, otherwise the higher is better.

3.4. Baseline models

To compare our proposed method SWiBE, we consider five baselines approaches including a regression-based network, a GAN-based network, and three diffusion model-based networks. In this study, all models are implemented with the publicly available source codes.

The first baseline model, *EP-WUN* [1], is a noise-robust BWE model trained using a modified triplet loss on the model embedding domain to adapt the entire representation toward a clean space. The second model, *CMGAN* [29], is a Conformer

[1] https://github.com/alexlinander/SWiBE

Table 1: *Metric result on VoiceBank-DEMAND test set, as* $\mathbf{R}$ *and* $\mathbf{G}$ *denote regression-based and generative methods, respectively. The 95% confidence interval is computed for SWiBE, showing as follows:* [15.75, 16.18] *for SNR,* [2.79, 2.87] *for PESQ,* [0.846, 0.859] *for ESTOI,* [1.04, 1.05] *for LSD,* [3.81, 3.88] *for CSIG,* [3.29, 3.35] *for CBAK, and* [3.30, 3.38] *for COVL.*

Model	type	#Params	SNR↑	PESQ↑	ESTOI↑	LSD↓	CSIG↑	CBAK↑	COVL↑
Unprocessed	*	*	8.78	1.96	0.771	2.98	1.00	2.41	1.06
EP-WUN [1]	**R**	4.58M	14.67	2.25	0.810	1.06	3.50	2.94	2.86
CMGAN [29]	**G**	1.83M	2.36	2.21	0.784	1.70	2.40	2.60	2.30
CDiffuSE [15]	**G**	4.28M	11.67	2.22	0.738	1.40	3.10	2.72	2.64
NU-Wave2 [8]	**G**	1.7M	12.01	1.86	0.761	1.13	3.11	2.67	2.46
SGMSE+ [17]	**G**	65.6M	15.85	2.82	0.851	1.13	3.77	3.30	3.30
SWiBE	**G**	65.6M	**15.97**	**2.83**	**0.852**	**1.04**	**3.84**	**3.32**	**3.34**

(convolution-augmented transformer)-based Metric GAN that utilizes two-stage conformer blocks in a generator and a metric discriminator, achieving great performance on speech enhancement. The remaining are diffusion model-based methods, including: *CDiffuSE* [15], a conditional diffusion probabilistic model for speech enhancement which incorporates characteristics of the observed noisy speech signal into the diffusion and reverse processes; *NU-Wave2* [8], a diffusion model for neural audio super-resolution that generates 48k Hz audio signals from inputs of various sampling rates with a single model; *SGMSE+* [17], a score-based generative model in the complex STFT domain for both speech enhancement and speech dereverberation.

4. Results and Analyses

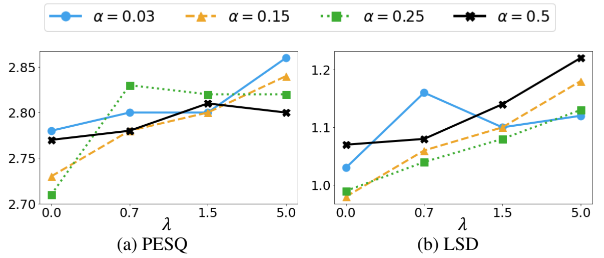
4.1. Baseline comparison

Table 1 shows the comparison between SWiBE and other baselines for reconstructing clean 16k speech signals from noisy 8k speech signals, evaluated on the VoiceBank-DEMAND test set, with the demonstration of the corresponding method type and the model size. SWiBE in Table 1 is obtained with hyperparameters set as $\alpha = 0.25$ and $\lambda = 0.7$. The 95% confidence interval of SWiBE is also provided.

Overall, our proposed SWiBE demonstrates superior performance across all metrics. In perceptual quality measurements such as PESQ and STOI, while SGMSE+ has already shown outstanding performance, SWiBE achieves further improvement. Additionally, compared with SGMSE+, SWiBE obtains better CSIG and CBAK scores, indicating that SWiBE at the same time achieves better speech signal preservation and improved noise suppression. When considering spectral reconstruction, generative methods commonly show poorer LSD scores although achieving competitive performance on other metrics. This is in contrast to the great LSD score obtained by the regression-based EP-WUN. Nevertheless, SWiBE shows significant improvement with a 9% and 2% reduction in LSD score compared to SGMSE+ and EP-WUN, respectively. The result implies that the stochastic diffusion process in SWiBE operates more proficiently, especially regarding expansion in the frequency domain, which may be facilitated by the stepwise expansion strategy.

4.2. Effect on parameterized stochastic diffusion process

With the utilization of B_{step} formulated from equation 6, adjustments to α and λ allow us to delineate distinct expansion trajectories in the reverse process, as illustrated in Figure 1. To assess the impact of these expansion paths on model characteristics and the quality of reconstructed speech, we conduct an experimental grid search on varying α and λ. The results are presented in Figure 2, which depicts the variations of PESQ scores in 2a and LSD scores in 2b, along different saturation speeds. Each line in the graph represents a different starting bandwidth, delineated by different colors. As shown, the PESQ curves demonstrate an overall trend that as the starting bandwidth increases, the model produces reconstructed speeches with higher perceptual quality. In contrast, the LSD curves suggest that a lower starting bandwidth leads to better spectral reconstruction. Among these trends, a model with $\lambda = 0.7$ appears to achieve the best balance, as there is a positive correlation between higher perceptual quality and poorer spectral reconstruction. Regarding the impact on different saturation speeds, we observe from the LSD curves that $\alpha = 0.15$ or 0.25 may be the most appropriate since higher or lower speeds tend to result in poorer spectral reconstruction.



(a) PESQ (b) LSD

Figure 2: *Speech quality metrics variation along to* α*, i.e., different saturation speed, under different* λ*, i.e., different starting bandwidths.*

5. Conclusion

In this paper, we propose SWiBE, an SGM-based noise-robust BWE framework featuring a BWE-focused stochastic diffusion process. Our proposed SWiBE exhibits superior performance across all considered metrics, notably showing significant improvement in LSD score compared to other generative methods. This underscores the effectiveness of the score-based model derived from our diffusion process in understanding frequency domain expansion. Additionally, we parameterize our diffusion process to enable expansion along distinct trajectories, revealing diverse model characteristics influenced by varying hyperparameters. However, the relationship between these variations and downstream tasks like ASR remains unexplored. Moreover, our study is limited by its focus on specific metrics, hence, our future research will expand the analysis to assess how different diffusion processes affect downstream tasks, alongside conducting comprehensive investigations into various metrics.

6. References

[1] Y.-T. Lin, B.-H. Su, C.-H. Lin, S.-C. Kuo, J.-S. R. Jang, and C.-C. Lee, "Noise-Robust Bandwidth Expansion for 8K Speech Recordings," in *Proc. INTERSPEECH 2023*, 2023, pp. 5107–5111.

[2] V. Kuleshov, S. Z. Enam, and S. Ermon, "Audio super-resolution using neural networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017.

[3] S. Birnbaum, V. Kuleshov, Z. Enam, P. W. W. Koh, and S. Ermon, "Temporal film: Capturing long-range sequence dependencies with feature-wise modulations." *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[4] H. Liu, W. Choi, X. Liu, Q. Kong, Q. Tian, and D. Wang, "Neural Vocoder is All You Need for Speech Super-resolution," in *Proc. Interspeech 2022*, 2022, pp. 4227–4231.

[5] X. Li, V. Chebiyyam, K. Kirchhoff, and A. Amazon, "Speech audio super-resolution for speech recognition." in *INTERSPEECH*, 2019, pp. 3416–3420.

[6] S. Kim and V. Sathe, "Bandwidth extension on raw audio via generative adversarial networks," *arXiv preprint arXiv:1903.09027*, 2019.

[7] J. Lee and S. Han, "Nu-wave: A diffusion probabilistic model for neural audio upsampling," *Proc. Interspeech 2021*, pp. 1634–1638, 2021.

[8] S. Han and J. Lee, "NU-Wave 2: A General Neural Audio Upsampling Model for Various Sampling Rates," in *Proc. Interspeech 2022*, 2022, pp. 4401–4405.

[9] C.-Y. Yu, S.-L. Yeh, G. Fazekas, and H. Tang, "Conditioning and sampling in variational diffusion models for speech super-resolution," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[11] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Peer, and T. Gerkmann, "Causal diffusion models for generalized speech enhancement," *IEEE Open Journal of Signal Processing*, 2024.

[12] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.

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

[14] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *9th International Conference on Learning Representations, ICLR*, 2021.

[15] 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)*. IEEE, 2022, pp. 7402–7406.

[16] H. Yen, F. G. Germain, G. Wichern, and J. Le Roux, "Cold diffusion for speech enhancement," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[17] 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*, 2023.

[18] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv preprint arXiv:2206.03065*, 2022.

[19] 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.

[20] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[21] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching." *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.

[22] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[23] S. Särkkä and A. Solin, *Applied stochastic differential equations*. Cambridge University Press, 2019, vol. 10.

[24] Y. Song, C. Durkan, I. Murray, and S. Ermon, "Maximum likelihood training of score-based diffusion models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1415–1428, 2021.

[25] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7537–7547, 2020.

[26] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech." in *SSW*, 2016, pp. 146–152.

[27] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5_Supplement, pp. 3591–3591, 2013.

[28] Y. Song and S. Ermon, "Advances in neural information processing systems," *Advances in Neural Information Processing Systems*, pp. 12 438–12 448, 2020.

[29] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based Metric GAN for Speech Enhancement," in *Proc. Interspeech 2022*, 2022, pp. 936–940.

[30] 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 *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[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*, vol. 24, no. 11, pp. 2009–2022, 2016.

[32] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.