Reducing the Prior Mismatch of Stochastic Differential Equations for Diffusion-based Speech Enhancement

Bunlong Lay[1], Simon Welker[1,2], Julius Richter[1], Timo Gerkmann[1]

[1]Signal Processing (SP), University of Hamburg, Germany
[2]Center for Free-Electron Laser Science, DESY, Hamburg, Germany
{bunlong.lay, simon.welker, julius.richter, timo.gerkmann}@uni-hamburg.de

Abstract

Recently, score-based generative models have been successfully employed for the task of speech enhancement. A stochastic differential equation is used to model the iterative forward process, where at each step environmental noise and white Gaussian noise are added to the clean speech signal. While in limit the mean of the forward process ends at the noisy mixture, in practice it stops earlier and thus only at an approximation of the noisy mixture. This results in a discrepancy between the terminating distribution of the forward process and the prior used for solving the reverse process at inference. In this paper, we address this discrepancy and propose a forward process based on a Brownian bridge. We show that such a process leads to a reduction of the mismatch compared to previous diffusion processes. More importantly, we show that our approach improves in objective metrics over the baseline process with only half of the iteration steps and having one hyperparameter less to tune.

Index Terms: speech enhancement, diffusion models, stochastic differential equations, Brownian bridge.

1. Introduction

Speech enhancement aims to recover the clean speech signal from a noisy mixture that is corrupted by environmental noise [1]. Classical approaches try to exploit statistical relations of the clean speech signal and the environmental noise [2]. Numerous machine learning methods have been proposed that treat speech enhancement as a discriminative learning task [3, 4].

Different from discriminative approaches that learn a direct mapping from noisy to clean speech, generative approaches learn a prior distribution over clean speech data. Recently, so-called *score-based generative models* (or *diffusion models*) were introduced to the task of speech enhancement [5–9]. The idea is to iteratively add Gaussian noise to the data using a discrete and fixed Markov chain called *forward process*, thereby transforming data into a tractable distribution such as a normal distribution. Then, a neural network is trained to invert this diffusion process in a so-called *reverse process* [10]. When the step size between two discrete Markov chain states is taken to zero, the discrete Markov chain becomes a continuous-time stochastic differential equation (SDE) under mild constraints. Utilizing SDEs offers more flexibility and opportunities than approaches based on discrete Markov chains [11]. For example, SDEs allow to use general-purpose SDE solvers to numerically integrate the reverse process, impacting the performance and number of iteration steps. An SDE can be interpreted as a trans-formation between two given distributions, where one is called the initial distribution and the other the terminating distribution. In the case of speech enhancement, we transform between the distribution of clean speech data and the distribution of noisy mixture data. Under mild constraints, we can find for each forward SDE a reverse SDE inverting the forward SDE [12, 13]. This reverse SDE starts from a noisy mixture and ends at the clean speech. It can be therefore used for speech enhancement.

Currently, for the task of speech enhancement, there are different approaches that integrate the corruption of environmental noise in the diffusion process [6–8]. To compensate for non-Gaussian noise characteristics, these approaches use an interpolation between clean speech and noisy speech data along the forward process. In [7,8] a continuous-time SDE is used, which includes a drift term that allows the transformation between clean and noisy speech. Interestingly, the mean of the process in [7,8] evolves from clean speech perfectly to noisy speech only for an infinitely long forward diffusion process. In practice, however, the mean of the forward process ends at an approximation of the noisy speech data. Therefore, when solving the reverse SDE to perform speech enhancement, there exists a mismatch between the terminating distribution of the forward process and the initial distribution of the reverse process [8]. We call the initial distribution of the reverse process the prior distribution of the generative model and the corresponding mismatch the *prior mismatch*. Moreover, the SDEs in [7, 8, 14, 15] includes a stiffness parameter controlling the pull of the terminating distribution of the forward process and the prior distribution. Consequently, this stiffness parameter determines the degree of the resulting prior mismatch. Increasing the stiffness reduces the prior mismatch, but may also negatively affect the speech enhancement performance as the reverse process may become unstable [8, Section II D].

To overcome this limitation, we seek to reduce the prior mismatch without destabilizing the reverse process. To this end, we propose to replace the forward process in [7,8] with an SDE based on a Brownian bridge process. A Brownian bridge seems suitable for this purpose because it has fixed starting and end points and follows a Brownian motion in between. We show that the resulting diffusion process does not only drastically decrease the prior mismatch, but also eliminates the dataset-dependent and hard-to-tune stiffness parameter of the SDE in [7, 8]. In the experiments, we demonstrate that using the proposed SDE outperforms the baseline SDE while having one hyperparameter less to tune and using only half as many function evaluations [1].

[1]code online available https://github.com/sp-uhh/sgmse-bbed

2. Background

The task of speech enhancement is to estimate the clean speech signal $\mathbf{S}$ from a noisy mixture $\mathbf{Y} = \mathbf{S} + \mathbf{N}$, where $\mathbf{N}$ is environmental noise. All variables in bold are the coefficients of a complex valued short-time Fourier transform (STFT), e.g. $\mathbf{Y} \in \mathbb{C}^d$ and $d = KF$ with K number of STFT frames and F number of frequency bins.

2.1. Stochastic Differential Equations

Following the approach in [7, 8], we model the forward process of the score-based generative model with an SDE defined on $0 \leq t < T_{\max}$:

$$d\mathbf{X}_t = \mathbf{f}(\mathbf{X}_t, \mathbf{Y})dt + g(t)d\mathbf{w}, \qquad (1)$$

where $\mathbf{w}$ is the standard Wiener process [16], $\mathbf{X}_t$ is the current process state with initial condition $\mathbf{X}_0 = \mathbf{S}$, and t a continuous diffusion time-step variable describing the progress of the process ending at the last diffusion time-step $T_{\max}$. Moreover, $\mathbf{f}(\mathbf{X}_t, \mathbf{Y})dt$ can be integrated by Lebesgue integration [17], and $g(t)d\mathbf{w}$ follows Ito integration [16]. The functions $\mathbf{f}(\mathbf{X}_t, \mathbf{Y})$ and $g(t)$ are called drift and diffusion coefficient, respectively. The diffusion coefficient g regulates the amount of Gaussian noise that is added to the process, and the drift $\mathbf{f}$ affects mainly in the case of linear SDEs the mean of $\mathbf{X}_t$ (see [16, (6.10)]). The process state $\mathbf{X}_t$ follows a Gaussian distribution [18, Ch. 5], called the *perturbation kernel*:

$$p_{0t}(\mathbf{X}_t | \mathbf{X}_0, \mathbf{Y}) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{X}_t; \boldsymbol{\mu}(\mathbf{X}_0, \mathbf{Y}, t), \sigma(t)^2 \mathbf{I}\right). \qquad (2)$$

By Anderson [12], each forward SDE as in (1) can be associated to a reverse SDE:

$$d\mathbf{X}_t = \left[-\mathbf{f}(\mathbf{X}_t, \mathbf{Y}) + g(t)^2 \nabla_{\mathbf{X}_t} \log p_t(\mathbf{X}_t | \mathbf{Y})\right] dt + g(t)d\bar{\mathbf{w}}, \qquad (3)$$

where $d\bar{\mathbf{w}}$ is a Wiener process going backwards in time. In particular, the reverse process starts at $t = T$ and ends at $t = 0$. Here $T < T_{\max}$ is a parameter that needs to be set for practical reasons as the last diffusion time-step $T_{\max}$ is only reached in limit. The *score function* $\nabla_{\mathbf{X}_t} \log p_t(\mathbf{X}_t | \mathbf{Y})$ is approximated by a neural network called *score model* $s_\theta(\mathbf{X}_t, \mathbf{Y}, t)$, which is parameterized by a set of parameters θ. Assuming that s_θ is available, we can generate an estimate of the clean speech $\mathbf{X}_0$ from $\mathbf{Y}$ by solving the reverse SDE.

The prior mismatch discussed in this paper is defined by the difference of $\boldsymbol{\mu}(\mathbf{X}_0, \mathbf{Y}, T)$ to $\mathbf{Y}$. In this work and previous work [7, 8], we consider only SDEs where the mean is of the form

$$\boldsymbol{\mu}(\mathbf{X}_0, \mathbf{Y}, t) = (1 - k(t))\mathbf{X}_0 + k(t)\mathbf{Y}, \qquad (4)$$

where $0 \leq k(t) < 1$ is an increasing function. In the sequel, we will simply write $\boldsymbol{\mu}(t)$ for brevity. The mismatch of such an SDE is determined by $k(T)$ and we call $k(T)$ the *maximal interpolation factor* (MIF) for the rest of the paper. It is desired that the MIF $k(T)$ is close to 1 and we will see in the following sections to which degree this goal is met.

3. Design choices of different SDEs

3.1. Ornstein-Uhlenbeck with Variance Exploding (OUVE)

In [7, 8] an SDE is used with the drift coefficient $f(\mathbf{X}_t, \mathbf{Y})$ and diffusion coefficient $g(t)$ defined as

$$\mathbf{f}(\mathbf{X}_t, \mathbf{Y}) = \gamma(\mathbf{Y} - \mathbf{X}_t), \qquad (5)$$

$$g(t) = \sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^t \sqrt{2 \log\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)}, \qquad (6)$$

for $0 \leq t \leq T < T_{\max} = \infty$ and parameters $\gamma, \sigma_{\min}, \sigma_{\max} \in \mathbb{R}_+$. Such a drift term is typical for an Ornstein-Uhlenbeck process [16], whereas the diffusion coefficient is taken from the so-called *Variance Exploding* SDE [11]. Thus, we call the baseline SDE *Ornstein-Uhlenbeck with Variance Exploding* (OUVE). A reparameterization of (6) with $\sigma_{\max} := k\sigma_{\min}$ and $c := \sigma_{\min}^2 2 \log(\frac{\sigma_{\max}}{\sigma_{\min}})$ yields

$$g(t) = \sqrt{c}k^t, \quad \text{where } c, k > 0. \qquad (7)$$

We argue that this equivalent representation of the diffusion coefficient may increase the intuition of (6), as $\sqrt{c}$ simply scales the diffusion coefficient and k is the base of the exponential term. We will simply use the parameterization of Eq. (7) for the rest of this work.

The closed-form solution for the mean and variance of the perturbation kernel of this SDE are given by:

$$\sigma(t)^2 = \frac{c\left(k^{2t} - e^{-2\gamma t}\right)}{2(\gamma + \log(k))}, \qquad (8)$$

and

$$\boldsymbol{\mu}(t) = e^{-\gamma t}\mathbf{X}_0 + (1 - e^{-\gamma t})\mathbf{Y}. \qquad (9)$$

We see from (9) that for large $t \to \infty$, we have that $\mathbf{X}_t$ has mean $\mathbf{Y}$. However, as in practice we need to decide for a finite final diffusion time-step T, a certain difference between the mean of $\mathbf{X}_T$ and $\mathbf{Y}$ remains. If we parameterize the OUVE SDE as in [8], i.e. $\gamma = 1.5$ and $T = 1$, then we find that the MIF is $k(T) = (1 - e^{-1.5}) = 0.78$. As it is desired to have a MIF close to 1, we argue that the difference between $\boldsymbol{\mu}(T)$ and $\mathbf{Y}$ is relatively large. Note that increasing T for fixed γ to obtain a better MIF is equivalent to fixing T and increasing γ. Moreover, we have that increasing γ, yields a better MIF, but also worsens the performance of this approach, as the sampling from the reserve SDE becomes unstable [8, Section II D]. Therefore, increasing the MIF $k(T) = 1 - e^{-\gamma T}$ for this SDE is not straightforward.

3.2. Brownian Bridge with Exponential Diffusion Coefficient (BBED)

In order to reduce the prior mismatch, we propose to employ an SDE that has a linear interpolation factor $k(t) = t$, where $0 \leq t \leq T < T_{\max} = 1$. Substituting $k(t) = t$ the mean of the SDE in (4) becomes

$$\boldsymbol{\mu}(t) = (1 - t)\mathbf{X}_0 + t\mathbf{Y}. \qquad (10)$$

One can find an SDE with the following drift coefficient that has the desired mean from (10) by solving [16, (6.12)]

$$\mathbf{f}(\mathbf{X}_t, \mathbf{Y}) = \frac{\mathbf{Y} - \mathbf{X}_t}{1 - t}. \qquad (11)$$

Comparing (10) and (4), we see that the MIF is $k(T) = T$. Note, that the choice of $T < 1$ is limited due to numerical
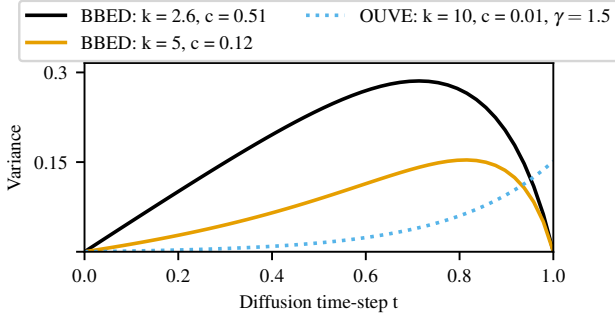
Figure 1: *Variance evolution of BBED and OUVE. Solid curves are variances for BBED with different parameterizations. The dotted curve is the variance of the OUVE with parameterization as in Section 4.4.*
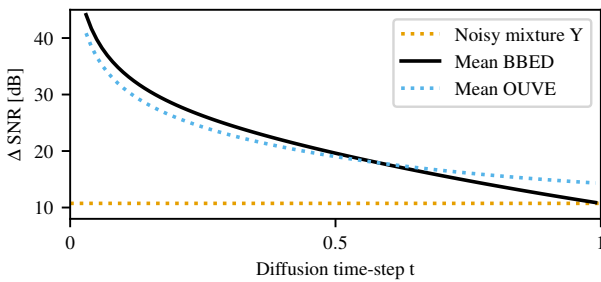

Figure 2: *Black and blue curves are the averaged $\Delta SNR(\mu(t))$ as defined in (15) of the mean evolutions of BBED and OUVE for the WSJ0-CHiME3 test set. The dotted yellow line is the SNR of $\mathbf{Y}$. The OUVE SDE is parameterized with $\gamma = 1.5$ as described in Section 4.4.*

stability as we divide by $(1 - t)$ in (11). However, it is still possible to achieve a much better MIF compared to the MIF of the OUVE SDE, as we will see in Section 5.2.

For a fair comparison to the OUVE SDE, we want to utilize the same diffusion coefficient as from the OUVE SDE in Eq. (7). The resulting variance can be computed from [16, (6.11)]:

$$\sigma(t)^2 = (1 - t)c \left[(k^{2t} - 1 + t) + \log(k^{2k^2})(1 - t)E \right], \quad (12)$$

$$E = \text{Ei} \left[2(t - 1) \log(k) \right] - \text{Ei} \left[-2 \log(k) \right], \quad (13)$$

where $\text{Ei}[\cdot]$ denotes the exponential integral function [19]. The variance trajectory exhibits one peak and vanishes for $t = 0$ and $t = 1$. The position of the peak is solely determined by k, where larger k shifts the peak closer to $t = 1$.

In the literature, SDEs that linearly transform the starting condition ($\mathbf{X}_0 = \mathbf{S}$ and zero variance for $t = 0$) to the terminal condition ($\mathbf{X}_T = \mathbf{Y}$ and zero variance for $t = 1$) with a constant diffusion coefficient of $g(t) = 1$ are called Brownian bridges [16]. As the SDE with drift coefficient (11) and diffusion coefficient (7) differs from that definition only in the diffusion coefficient, we call the SDE a *Brownian Bridge with Exponential Diffusion coefficient* (BBED).

4. Experimental setup

To allow a fair comparison between BBED SDE with OUVE SDE, we train the corresponding score models with the same configuration and follow the experimental setup from [8, Section V].

4.1. Training

For the score model $s_\theta(\mathbf{X}_t, \mathbf{Y}, t)$, we employ the Noise Conditional Score Network (NCSN++) architecture (see [8, 11] for more details). The network is optimized based on denoising score matching:

$$\arg \min_\theta \mathbb{E}_{t, (\mathbf{X}_0, \mathbf{Y}), \mathbf{Z}, \mathbf{X}_t | (\mathbf{X}_0, \mathbf{Y})} \left[\left\| \mathbf{s}_\theta(\mathbf{X}_t, \mathbf{Y}, t) + \frac{\mathbf{Z}}{\sigma(t)} \right\|_2^2 \right],$$
$$(14)$$

where $\mathbf{X}_t = \boldsymbol{\mu}(t) + \sigma(t)\mathbf{Z}$ with $\mathbf{Z} \sim \mathcal{N}_\mathbb{C}(\mathbf{0}, \mathbf{I})$. We train the network with the ADAM optimizer [20] with a learning rate of 10^{-4} and a batch size of 16. An exponential moving average of network parameters is tracked with a decay of 0.999, to be used for sampling [7, 11]. We train for 250 epochs and log the averaged PESQ value of 10 random files from the validation set during training and select the best-performing model for evaluation. Experiments are conducted on an NVIDIA A6000 and training lasts for approximately 4 days.

4.2. Dataset and input representation

We use the same WSJ0-CHiME3 dataset as in [8]. This dataset mixes clean speech utterances from the Wall Street Journal (WSJ0) dataset [21] to noise signals from the CHiME3 dataset [22] with an uniformly sampled signal-to-noise ratio (SNR) between 0 and 20 dB. The dataset is split into a train (12777 files), validation (1206 files) and test set (615 files).

Each file from the WSJ0-CHiME3 dataset is converted into a complex STFT representation with a window size of 510, resulting in 256 frequency bins, a hop size of 128 and a periodic Hann window. We randomly crop the STFT representation to a length of 256 frames at each training step. To compensate for the typically heavy-tailed distribution of STFT speech magnitudes [23], as in [8], each complex coefficient c of the STFT representation is transformed via $\beta |c|^\alpha e^{i\angle(c)}$ with $\beta = 0.15$ and $\alpha = 0.5$.

4.3. Sampling and metrics

For the baseline OUVE SDE and the proposed BBED SDE we use the same sampler settings for a fair comparison. We use a Predictor-Corrector scheme as in [8, 11], where the Predictor is the Euler-Maruyama method [18] and the Corrector is the Annealed Langevin Dynamics (ALD) method [11]. As in [8], the step size for ALD is chosen as 0.5 and the number of reverse steps is 30. Equivalently, the step size in the reverse process is $h = T/30$, where T is set for the OUVE SDE and BBED SDE individually (see Section 4.4). For the reverse process, we set the reverse starting time at $t_{\text{rs}} = T$. We also report results when experimenting with the reverse starting times $t_{\text{rs}} < T$ in Section 5.3 while keeping the step size $h = T/30$ fixed.

We evaluate the performance on perceptual metrics, wideband PESQ [24] and POLQA [25], on energy-based metrics SI-SDR, SI-SIR and SI-SAR [26] and on intelligibility metric ESTOI [27].

4.4. OUVE and BBED

As in [8], the parameters T, $\sigma_{\min}$, $\sigma_{\max}$ and θ were already tuned by a grid search. Therefore, we set as in [8] $T = 1$ and $\gamma = 1.5$, and the diffusion coefficient parameters in Eq. 6 are set to $\sigma_{\min} = 0.05$ and $\sigma_{\max} = 0.5$, or in the equivalent representation in Eq. 7, we set $k = 10$ and $c = 0.01$.

For the BBED SDE, we search for the largest T in $\{0.9, 0.99, 0.999, 0.9999\}$ so that training and inference

Table 1: *Speech enhancement results (average and standard deviation over the test set) obtained for WSJ0-CHiME3. The OUVE SDE is parameterized as described in 4.4 and the BBED SDE is parameterized with $k = 2.6, c = 0.51$. t_{rs} denotes the reverse starting time as defined in Section 4.3.*

SDE	POLQA	PESQ	ESTOI	SI-SDR [dB]	SI-SIR [dB]	SI-SAR [dB]
Mixture	2.63 ± 0.67	1.70 ± 0.49	0.78 ± 0.14	10.0 ± 5.7	10.0 ± 5.7	-
Baseline OUVE [8]	3.71 ± 0.51	2.92 ± 0.53	0.92 ± 0.05	17.78 ± 4.57	31.51 ± 4.9	18.00 ± 4.65
BBED $t_{\mathrm{rs}} = 0.5$	3.97 ± 0.48	3.05 ± 0.53	0.93 ± 0.05	18.96 ± 4.28	31.42 ± 5.19	19.28 ± 4.36
BBED $t_{\mathrm{rs}} = 0.999$	$\mathbf{4.01 \pm 0.49}$	$\mathbf{3.08 \pm 0.57}$	$\mathbf{0.94 \pm 0.05}$	$\mathbf{19.26 \pm 4.43}$	$\mathbf{31.64 \pm 5.08}$	$\mathbf{19.59 \pm 4.53}$

is numerically stable. The parameter k in (7) is determined as the empirically optimal choice of $K :=$ $\{0.02, 0.2, 0.6, 1.1, 1.5, 2.6, 5, 27\}$. The values of the grid have been chosen in such a way that the resulting variances have their peaks ranging from 0.2 to 0.9. For example, the resulting variance for $k = 5$ has its maximum at 0.8, the variance for $k = 2.6$ has its maximum at 0.7, etc. For each $k \in K$, we set the normalization factor c so that the variances admit a maximum value of either 0.15 or 0.3. This choice is based on the OUVE SDE parameterization also having a maximum value of 0.15. Exemplary, we plot two parameterizations of the variance of the BBED SDE in Fig. 1.

5. Results

First, we present the results when parameterizing the BBED SDE as described in Section 4.4. Second, we discuss if the proposed BBED SDE reduces the prior mismatch compared to the baseline OUVE SDE. Last, we discuss the performance differences in terms of objective metrics, number of iterations in the reverse process and subjective differences of the OUVE SDE and BBED SDE.

5.1. Parameterization of the BBED SDE

When training and testing the score-model with the BBED SDE with different $k \in K$, we argue that it is beneficial to have the variance maximum towards the end of the forward process, as the Gaussian noise would better mask the speech features corrupted by the environmental noise. At the same, if the variance maximum is too close to the end of the forward process, which is at $t = 1$, then the diffusion coefficient becomes numerically large and consequently the reverse process may become unstable. Empirically, we found that $k = 2.6$ with maximum variance 0.3 results in the best performance (see Fig. 1 black line). When training and testing the score-model with the BBED SDE with different $T \in \{0.9, 0.99, 0.999, 0.9999\}$, we found that $T = 0.999$ is the largest value that causes no numerical issues.

5.2. Reducing the prior mismatch

As we set for the BBED SDE $T = 0.999$, we have that the MIF is $k(T) = 0.999$. This is much closer to 1 than the MIF of 0.78 achieved by the OUVE SDE as discussed in Section 3.1. We illustrate this prior mismatch in terms for SNR in Fig. 2. To this end, let y' and s be time-domain signals, where s is the clean speech signal and y' is any clean speech signal corrupted with environmental noise. We define the $\mathrm{SNR}(\mathbf{Y}', \mathbf{S})$ to be $20 \log_{10} \frac{||s||_2}{||y'-s||_2}$, $|| \cdot ||_2$ denotes the ℓ^2 norm. In Fig. 2, we averaged

$$\Delta \mathrm{SNR}(\mu(t)) := \mathrm{SNR}(\mu(t), \mathbf{S}) - \mathrm{SNR}(\mathbf{Y}, \mathbf{S}) \qquad (15)$$

for the BBED SDE and OUVE SDE over the WSJ0-CHiME3 test set. We have that $\mathrm{SNR}(\mu(t), \mathbf{S})$ approaches $\mathrm{SNR}(\mathbf{Y}, \mathbf{S})$

if $\mu(t) = \mathbf{Y}$. This is the case for the BBED SDE as it can be observed in Fig. 2. In comparison, we find that the OUVE SDE differs to the $\mathrm{SNR}(\mathbf{Y}, \mathbf{S})$ by 3.6 dB at $t = 1$ in Fig. 2, showing that the BBED SDE indeed reduces the prior mismatch compared to the OUVE SDE.

5.3. OUVE vs. BBED

In Tab. 1 we show that BBED outperforms OUVE in all reported metrics. When listening to enhanced files generated by BBED and OUVE, we observe that the enhanced files generated by BBED contain less background noise and breathing artifacts than the enhanced files generated by OUVE. We provide some listening examples in the supplementary material[2].

Remarkably, when experimenting with t_{rs} we found that the BBED SDE largely maintains performance when changing $t_{\mathrm{rs}} = T = 0.999$ to $t_{\mathrm{rs}} = 0.5$ as it can be seen in Tab. 1. This is in contrast to OUVE SDE which loses 0.31 in PESQ when we set $t_{\mathrm{rs}} = 0.5$. Since we keep the reverse step size $h = T/30$ fixed when starting inference at $t_{\mathrm{rs}} = 0.5$, the number of iterations is halved with only negligible performance loss for the BBED SDE as compared to when starting the reverse process at $t_{\mathrm{rs}} = 0.999$. In particular BBED even outperforms OUVE when only using half as many iterations for enhancement.

The proposed BBED SDE has a different drift coefficient compared to the OUVE SD (compare Eq. (11) and Eq. (5)), which results in different mean evolutions (see Fig. 2) and different variance evolutions (see Fig. 1). Thus, there could be various reasons why the BBED SDE outperforms the OUVE SDE. We hypothesize that a much higher variance of the BBED could be mainly responsible for the improvements, as a higher variance potentially helps to generate better speech estimates. We also believe that too large values for the diffusion coefficient $g(t)$ may lead to numerical instability of the reverse process. We leave this discussion for future work.

6. Conclusions

In this paper, we aimed to minimize the prior mismatch in score-based generative modeling for speech enhancement. To this end, we constructed the BBED SDE that is inspired by Brownian bridges. The BBED SDE yields a much smaller prior mismatch compared to the baseline OUVE SDE and has one hyperparameter less to tune. As a result, we consistently improve in all reported metrics over the OUVE SDE. Moreover, the BBED SDE achieves improvements of 0.13 in PESQ and 0.26 in POLQA even when only using half as many function evaluations as the OUVE SDE.

---

[2]https://www.inf.uni-hamburg.de/en/inst/ab/sp/publications/sgmse-bbed

7. References

[1] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state-of-the-art.* Morgan & Claypool, 2013.

[2] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. John Wiley & Sons, 2018.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 26, no. 10, pp. 1702–1726, 2018.

[4] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 27, no. 8, pp. 1256–1266, 2019.

[5] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," *IEEE Asia-Pacific Signal and Inf. Proc. Assoc. Annual Summit and Conf. (APSIPA ASC)*, pp. 659–666, 2021.

[6] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2022.

[7] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," *Interspeech*, 2022.

[8] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, 2023. [Online]. Available: https://arxiv.org/abs/2208.05830

[9] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *(under review)*, 2022. [Online]. Available: https://arxiv.org/abs/2206.03065

[10] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.

[11] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *Int. Conf. on Learning Representations (ICLR)*, 2021.

[12] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[13] U. G. Haussmann and E. Pardoux, "Time reversal of diffusions," *The Annals of Probability*, pp. 1188–1205, 1986.

[14] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. S. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," in *International Conference on Learning Representations*, 2022.

[15] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech," *Int. Conf. on Machine Learning (ICML)*, 2021.

[16] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed. Springer, 1996.

[17] W. Rudin, *Real and Complex Analysis*, 3rd ed. McGraw-Hill, Inc., 1987.

[18] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations*, ser. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2019, no. 10.

[19] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers.* McGraw-Hill, 1978.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Int. Conf. on Learning Representations (ICLR)*, 2015.

[21] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete." [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S6A

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

[23] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," *Int. Workshop on Acoustic Echo and Noise Control*, 2010.

[24] 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," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, vol. 2, pp. 749–752, 2001.

[25] ITU-T Rec. P.863, "Perceptual objective listening quality prediction," *Int. Telecom. Union (ITU)*, 2018. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

[26] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR– half-baked or well done?" in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2019, pp. 626–630.

[27] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 24, no. 11, pp. 2009–2022, 2016.