Speech Enhancement and Dereverberation With Diffusion-Based Generative Models

Julius Richter ⑩, *Student Member, IEEE*, Simon Welker ⑩, *Student Member, IEEE*,
Jean-Marie Lemercier ⑩, *Student Member, IEEE*, Bunlong Lay ⑩, and Timo Gerkmann ⑩, *Senior Member, IEEE*

Abstract—In this work, we build upon our previous publication and use diffusion-based generative models for speech enhancement. We present a detailed overview of the diffusion process that is based on a stochastic differential equation and delve into an extensive theoretical examination of its implications. Opposed to usual conditional generation tasks, we do not start the reverse process from pure Gaussian noise but from a mixture of noisy speech and Gaussian noise. This matches our forward process which moves from clean speech to noisy speech by including a drift term. We show that this procedure enables using only 30 diffusion steps to generate high-quality clean speech estimates. By adapting the network architecture, we are able to significantly improve the speech enhancement performance, indicating that the network, rather than the formalism, was the main limitation of our original approach. In an extensive cross-dataset evaluation, we show that the improved method can compete with recent discriminative models and achieves better generalization when evaluating on a different corpus than used for training. We complement the results with an instrumental evaluation using real-world noisy recordings and a listening experiment, in which our proposed method is rated best. Examining different sampler configurations for solving the reverse process allows us to balance the performance and computational speed of the proposed method. Moreover, we show that the proposed method is also suitable for dereverberation and thus not limited to additive background noise removal.

Index Terms—Speech enhancement, dereverberation, diffusion models, score-based generative models, score matching.

I. INTRODUCTION

SPEECH enhancement aims to recover clean speech signals from audio recordings that are impacted by acoustic noise or reverberation [1]. To this end, computational approaches often exploit the different statistical properties of the target and interference signals [2]. Machine learning algorithms can be used to extract these statistical properties by learning useful representations from large datasets. A wide class of methods employed for speech enhancement are discriminative models that learn to directly map noisy speech to the corresponding clean speech target [3]. Common approaches include time-frequency (T-F) masking [4], complex spectral mapping [5], or operating directly in the time domain [6]. These supervised methods are trained with a variety of clean/noisy speech pairs containing multiple speakers, different noise types, and a large range of signal-to-noise ratios (SNRs). However, it is nearly impossible to cover all possible acoustic conditions in the training data to guarantee generalization. Furthermore, some discriminative approaches have been shown to result in unpleasant speech distortions that outweigh the benefits of noise reduction [7].

The use of generative models for speech enhancement, on the other hand, follows a different paradigm, namely to learn a prior distribution over clean speech data. Thus, they aim at learning the inherent properties of speech, such as its spectral and temporal structure. This prior knowledge can be used to make inferences about clean speech given noisy or reverberant input signals that are assumed to lie outside the learned distribution. Several approaches follow this idea and utilized deep generative models for speech enhancement [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. Among them are methods that employ likelihood-based models for explicit density estimation such as the variational autoencoder (VAE) [18], or leverage generative adversarial networks (GANs) [19] for implicit density estimation. Bando et al. propose a statistical framework using a VAE trained in an unsupervised fashion to learn a prior distribution over clean speech [9]. At test time they combine the speech model with a low-rank noise model to infer the signal variances of speech and noise to build a Wiener filter for denoising. However, since the VAE is trained with clean speech only, the inference model (i.e. the encoder) that predicts the latent variable remains sensitive to noise. This has been shown to cause the generative speech enhancement method to produce speech-like sounds although only noise is present [9]. To mitigate this, it has been proposed to make the inference

model robust to noisy speech by training on labeled data in a supervised manner [14], [15], or by disentangling the latent variable from high-level information such as speech activity which can be estimated by supervised classifiers [12], [13]. Nevertheless, VAE-based speech enhancement methods remain limited due to the dimensionality reduction in the latent layer and the combined use of a linear noise model based on non-negative matrix factorization [9], [10], [11], [12], [13], [14], [15].

More recently, a new class of generative models called diffusion-based generative models, has been introduced to the task of speech enhancement [20], [21], [22], [23]. Diffusion-based generative models, or simply *diffusion models*, are inspired by non-equilibrium thermodynamics and exist in several variants [24], [25], [26]. All of them share the idea of gradually turning data into noise, and training a neural network that learns to invert this process for different noise scales. More specifically, the inference model is a fixed Markov chain, that slowly transforms the data into a tractable prior, such as the standard normal distribution. The generative model is another Markov chain that is trained to revert this process iteratively [25]. Therefore, diffusion models can be considered as deep latent variable models and have similar properties to VAEs, with the crucial difference that the inference model is not trained and that the latent variables have the same dimensionality as the input. This has the advantage of not relying on surrogate objectives to approximate maximum likelihood training such as the evidence lower bound and enforces no strong restrictions on the model architecture. Recently, diffusion models have been connected with score matching [27] by looking at the stochastic differential equation (SDE) associated with the discrete-time Markov chain [28]. The forward process can be inverted, resulting in a corresponding reverse SDE which depends only on the score function of the perturbed data [29]. Using this continuous-time SDE formalism creates the opportunity to design novel diffusion processes that support the underlying generation task. In contrast to discrete Markov chains, it also allows the use of general-purpose SDE solvers to numerically integrate the reverse process for sampling.

Concerning the application of diffusion models for speech enhancement, there exist currently two approaches that differ conceptually in how the diffusion process is used. One approach is based on speech regeneration, i.e. a diffusion-based vocoder network is used to synthesize clean speech by sampling from an unconditional prior, while a conditioner network takes noisy speech as input and performs the core part of denoising by providing enhanced speech representations to the vocoder network [23], [30]. An auxiliary loss is introduced for the conditioner network to facilitate its ability to estimate clean speech representations [23]. The second approach, on the other hand, does not require any auxiliary loss and is not using two separate models for generation and denoising. Instead, it models the corruption of clean speech by environmental background noise or reverberation directly within the forward diffusion process, so that reversing this process would consequently result in generating clean speech. This has been proposed as a discrete diffusion process for time-domain speech signals [21], and as a continuous SDE-based diffusion process in the complex

spectrogram domain [22]. Interestingly, the original denoising score matching objective [31], which is to estimate the white Gaussian noise in the perturbed data, is essentially reminiscent of the goal of speech enhancement, which is to remove interfering noise or reverberation from speech signals. However, under realistic conditions, the environmental noise or reverberation may not match the assumption of stationary white Gaussian noise. Therefore, it was proposed to include real noise recordings in the diffusion process, either by linearly interpolating between clean and noisy speech along the process [21], or by defining such a transformation within the drift term of an SDE [22]. The choice of linear interpolation in [21], however, implies that the trained deep neural network (DNN) must explicitly estimate a portion of environmental noise at each step in the reverse process. This can be seen in the resulting objective function [21, Eq. (21)] which exhibits characteristics of a discriminative learning task. In contrast, an SDE-based formulation results in a pure generative objective function [22, Eq. (9)] and avoids any prior assumptions on the noise distribution.

Nonetheless, note that diffusion-based speech enhancement methods, unlike the VAE-based method described above, are not counted as unsupervised methods, since labeled data (i.e. clean and noisy speech pairs) are used for training. However, the learning objective remains generative in nature which is to learn a prior for clean speech per se rather than a direct mapping from noisy to clean speech. In fact, supervision is only exploited to learn the conditional generation of clean speech when noisy speech is given. Thus, current diffusion-based models for speech enhancement, such as [21], [22], [23], as can be considered as conditional generative models trained in a supervised manner.

In this work, we build upon our previous publication which defines the diffusion process in the complex short-time Fourier transform (STFT) domain [22]. We present a comprehensive theoretical review of the underlying score-based generative model and include an expanded discussion on the conditional generation process which is based on the continuous-time SDE formalism. By using a network architecture developed in the image processing community [28], in the work at hand we significantly improve performance in comparison to our previous model [22]. This indicates that the network, rather than the formalism, was the main limitation of our original approach. In an extensive cross-dataset evaluation, we show that the improved method can compete with recent discriminative models and achieves better generalization when evaluating on a different corpus than used for training. To confirm the effectiveness of the proposed method on non-simulated data, we perform an instrumental evaluation with real-world noisy recordings using non-intrusive metrics. We complement the results with a listening experiment, in which our proposed method is rated best. Interestingly, using the improved network, we show that the proposed method is also suitable for dereverberation when an individual model is trained on simulated reverberant data. Thus, the method is not limited to the removal of additive background noise and can also be applied to non-additive corruptions such as reverberation or, as shown in [32], for bandwidth extension. Furthermore, we investigate different sampler configurations for solving the reverse process which reveals a trade-off between

the performance and computational speed of the proposed method.

We summarize our major contributions as follows. Regarding the novelty with respect to Song et al. [28], we introduce a drift term to the SDE to achieve the required task adaptation for reconstruction problems and furthermore apply the diffusion process and score matching objective to a complex data representation. Also note that the approach in [28] is not explicitly trained on reconstruction tasks and the application is different from ours. Regarding the novelty with respect to our previous publication [22], we use an improved network architecture and increase the performance significantly. Moreover, we include an extended theoretical discussion and investigate different sampler configurations. Finally, we expand the evaluation by means of a cross-dataset evaluation, an instrumental evaluation with real-world noisy recordings, and a listening experiment.

II. METHOD: SCORE-BASED GENERATIVE MODEL FOR SPEECH ENHANCEMENT (SGMSE)

In this section, we motivate and describe in detail the approach of using score-based generative models for speech enhancement, as proposed in our previous publication [22].

A. Data Representation

We represent our data in the complex-valued STFT domain, as it has been observed that both real and imaginary parts of clean speech spectrograms exhibit clear structure and are therefore amenable to deep learning models [4]. Following the approach of complex spectral mapping [5], we use our conditional generative model to estimate the clean real and imaginary spectrograms from the noisy ones.

The use of complex coefficients as data representation allows the definition of the diffusion process in the complex spectral domain, in which additive Gaussian noise corresponds to the signal model used for the denoising task. This relates to traditional STFT-based methods, where spectral coefficients are usually assumed to be complex Gaussian distributed and mutually independent [1], [2]. Statistical approaches often consider an additive signal model assuming that the speech process and the noise process are realizations of stochastic processes that are statistically independent. Observing that the overall noise process is a sum of several independent sources, the central limit theorem ensures that the observed noise process tends to be Gaussian [1].

Although it would be theoretically possible to define the diffusion process in the magnitude domain, additive Gaussian noise would not relate to the signal model anymore. This becomes evident considering that in the magnitude domain, additive Gaussian noise could result in negative amplitudes which are physically not defined.

Thus, we operate on complex spectrograms that are elements of $\mathbb{C}^{K \times F}$, where K denotes the number of time frames dependent on the audio length, and F represents the number of frequency bins. To compensate for the typically heavy-tailed distribution of STFT speech amplitudes [33], we apply an amplitude transformation

$$\tilde{c} = \beta |c|^{\alpha} e^{i \angle(c)} \tag{1}$$

to all complex STFT coefficients c, where $\angle(\cdot)$ represents the angle of a complex number, $\alpha \in (0, 1]$ is a compression exponent which brings out frequency components with lower energy (e.g. fricative sounds of unvoiced speech) [34], and $\beta \in \mathbb{R}_+$ is a simple scaling factor to normalize amplitudes roughly to within [0,1]. Such a compression has been argued to be perceptually more meaningful in speech enhancement [35], [36], and the transformation ensures that the neural network operates on consistently scaled inputs with respect to the Gaussian diffusion noise [25].

B. Stochastic Process

The tasks at hand, speech enhancement and dereverberation, can be considered as conditional generation tasks: Given the corrupted noisy/reverberant speech, generate clean speech by using a conditional generative model. Most previously published diffusion-based generative models are adapted to such conditional tasks either through explicit conditioning channels added to the DNN [37], [38], or through combining an unconditionally trained score model with a separate model (such as a classifier) that provides conditioning in the form of a gradient [28], [39]. With our method, we explore a third possibility, which is to incorporate the particular task directly into the forward and reverse processes of a diffusion-based generative model.

1) Forward Process: Following Song et al. [28], we design a stochastic diffusion process $\{\mathbf{x}_t\}_{t=0}^{T}$ that is modeled as the solution to a linear SDE of the general form,

$$d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, \mathbf{y})dt + g(t)d\mathbf{w}, \tag{2}$$

where $\mathbf{x}_t$ is the current process state, $t \in [0, T]$ a continuous time-step variable describing the progress of the process (not to be confused with the time index of any signal in the time or T-F domain), $\mathbf{y}$ the noisy or reverberant speech, and $\mathbf{w}$ denotes a standard Wiener process. The vector-valued function $\mathbf{f}(\mathbf{x}_t, \mathbf{y})$ is referred to as the *drift coefficient*, while $g(t)$ is called the *diffusion coefficient* and controls the amount of Gaussian white noise injected at each time-step. Note that different to Song et al. [28], our drift term is now a function of $\mathbf{y}$, by which we tailor the proposed SDE to reconstruction tasks. The process is defined for each T-F bin independently. Thus, the variables in bold are assumed to be vectors in $\mathbb{C}^d$ with $d = KF$ containing the coefficients of a flattened complex spectrogram.

The forward process in (2) turns a clean speech sample $\mathbf{x}_0$ into a corrupted sample $\mathbf{x}_T$ by gradually adding noise from the Wiener process, as illustrated in Fig. 1. To account for the intended task adaptation of speech enhancement or dereverberation, we propose a drift term that ensures the mean of the process moving from clean speech $\mathbf{x}_0$ to noisy/reverberant speech $\mathbf{y}$. In particular, we define the drift coefficient $\mathbf{f}$ and the diffusion coefficient g as

$$\mathbf{f}(\mathbf{x}_t, \mathbf{y}) := \gamma(\mathbf{y} - \mathbf{x}_t), \tag{3}$$
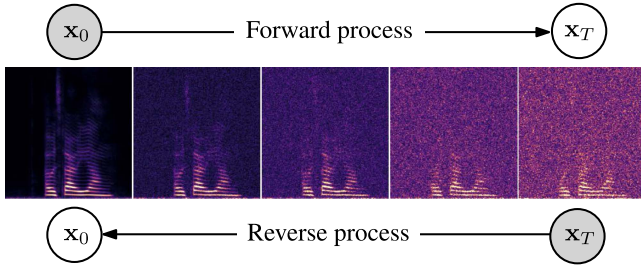
Fig. 1.   Diffusion process on a spectrogram: In the forward process noise is gradually added to the clean speech spectrogram $\mathbf{x}_0$, while the reverse process learns to generate clean speech in an iterative fashion starting from the corrupted signal $\mathbf{x}_T$.

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

where γ is a constant called *stiffness* controlling the transition from $\mathbf{x}_0$ to $\mathbf{y}$, and $\sigma_{\min}$ and $\sigma_{\max}$ are parameters defining the noise schedule of the Wiener process. Note that we choose the diffusion coefficient identical to that of the so-called *Variance Exploding SDE* from Song et al. [28]. Our novel contribution lies in the modified drift term, by which the intended task adaptation is achieved.

2) Reverse Process: Following Anderson [29] and Song et al. [28], the SDE in (2) has an associated *reverse SDE*,

$$d\mathbf{x}_t = \left[-\mathbf{f}(\mathbf{x}_t, \mathbf{y}) + g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y}) \right] dt + g(t) d\bar{\mathbf{w}}, \qquad (5)$$

where the *score* $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})$ is the term to be approximated by a DNN which is therefore called a *score model*. We denote the score model as $\mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t)$, which is parameterized by a set of parameters θ and receives the current process state $\mathbf{x}_t$, the noisy speech $\mathbf{y}$, and the current time-step t as an input. Finally, by substituting the score model into the reverse SDE in (5), we obtain the so-called *plug-in reverse SDE* [40],

$$d\mathbf{x}_t = \left[-\mathbf{f}(\mathbf{x}_t, \mathbf{y}) + g(t)^2 \mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t) \right] dt + g(t) d\bar{\mathbf{w}}, \qquad (6)$$

which can be solved by various solver procedures, to be discussed in detail in Section III.

For inference, we assume that a trained score model $\mathbf{s}_\theta$ is given, which approximates the true score for all $t \in [0, T]$. We can then generate clean speech $\mathbf{x}_0$ conditioned on the noisy or reverberant speech $\mathbf{y}$ by solving the plug-in reverse SDE in (6). To determine the initial condition of the reverse process at $t = T$, we sample

$$\mathbf{x}_T \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}_T; \mathbf{y}, \sigma(T)^2 \mathbf{I}), \qquad (7)$$

which is a strongly corrupted version of the noisy speech $\mathbf{y}$, as illustrated in Fig. 1. The denoising process which solves the task of speech enhancement or dereverberation is then based on iterating through the reverse process starting at $t = T$ and ending at $t = 0$.

C. Training Objective

Next, we derive the objective function used for training the score model $\mathbf{s}_\theta$. Since the SDE in (2) describes a Gaussian process, the mean and variance of the process state $\mathbf{x}_t$ can be derived when its initial conditions are known [41]. This allows for direct sampling of $\mathbf{x}_t$ at an arbitrary time step t given $\mathbf{x}_0$ and $\mathbf{y}$ by using the so-called *perturbation kernel*,

$$p_{0t}(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y}) = \mathcal{N}_{\mathbb{C}} \left(\mathbf{x}_t; \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t), \sigma(t)^2 \mathbf{I} \right), \qquad (8)$$

where $\mathcal{N}_{\mathbb{C}}$ denotes the circularly-symmetric complex normal distribution and $\mathbf{I}$ denotes the identity matrix. We utilize Eqs. (5.50, 5.53) in Särkkä & Solin [41] to determine closed-form solutions for the mean

$$\boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t) = \mathrm{e}^{-\gamma t} \mathbf{x}_0 + (1 - \mathrm{e}^{-\gamma t}) \mathbf{y}, \qquad (9)$$

and the variance

$$\sigma(t)^2 = \frac{\sigma_{\min}^2 \left((\sigma_{\max}/\sigma_{\min})^{2t} - \mathrm{e}^{-2\gamma t} \right) \log(\sigma_{\max}/\sigma_{\min})}{\gamma + \log(\sigma_{\max}/\sigma_{\min})}. \qquad (10)$$

Vincent [31] shows that fitting the score model $\mathbf{s}_\theta$ to the score of the perturbation kernel $\nabla_{\mathbf{x}_t} \log p_{0t}(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y})$ is equivalent to implicit and explicit score matching [27] under some regularity conditions. This technique is called *denoising score matching* and essentially results in estimating

$$\nabla_{\mathbf{x}_t} \log p_{0t}(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y}) = \nabla_{\mathbf{x}_t} \log \left[|2\pi\sigma\mathbf{I}|^{-\frac{1}{2}} \mathrm{e}^{-\frac{\|\mathbf{x}_t - \boldsymbol{\mu}\|_2^2}{2\sigma^2}} \right] \qquad (11)$$

$$= \nabla_{\mathbf{x}_t} \log |2\pi\sigma(t)\mathbf{I}|^{-\frac{1}{2}} - \nabla_{\mathbf{x}_t} \frac{\|\mathbf{x}_t - \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t)\|_2^2}{2\sigma(t)^2} \qquad (12)$$

$$= -\frac{\mathbf{x}_t - \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t)}{\sigma(t)^2}, \qquad (13)$$

where for simplicity we derived the score for the real and imaginary part of the complex normal distribution in (8), assuming they are independently distributed and each follows a real-valued multivariate normal distribution. Note that (13) involves division by $\sigma(t)^2$, which has very small numerical values (including 0) around $t = 0$. To avoid undefined values and numerical instabilities, we thus introduce a small minimum process time t_ε, as done previously in the literature [28].

At each training step, the procedure can then be described as follows: 1) sample a random $t \sim \mathcal{U}[t_\varepsilon, T]$, 2) sample $(\mathbf{x}_0, \mathbf{y})$ from the dataset, 3) sample $\mathbf{z} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{z}; 0, \mathbf{I})$, and 4) sample $\mathbf{x}_t$ from (8) by effectively computing

$$\mathbf{x}_t = \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t) + \sigma(t)\mathbf{z}. \qquad (14)$$

After passing $(\mathbf{x}_t, \mathbf{y}, t)$ to the score model, the final loss is an unweighted L_2 loss between the model output and the score of the perturbation kernel. By substituting (14) into (13), the overall training objective becomes,

$$\arg\min_\theta \mathbb{E}_{t, (\mathbf{x}_0, \mathbf{y}), \mathbf{z}, \mathbf{x}_t|(\mathbf{x}_0, \mathbf{y})} \left[\left\| \mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t) + \frac{\mathbf{z}}{\sigma(t)} \right\|_2^2 \right], \qquad (15)$$

where the expectation is approximated by sampling all random variables at each training step as described above. Note that
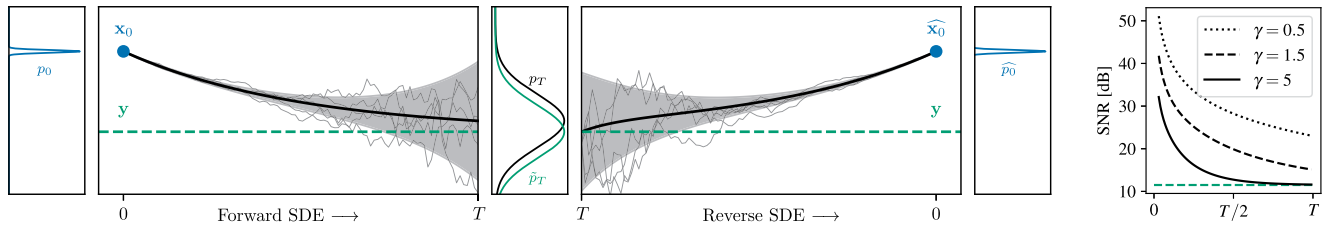
Fig. 2. (Left) The forward and reverse process illustrated with a single scalar variable. The mean $\boldsymbol{\mu}$ (thick black line) of the forward process exponentially decays from clean speech $\mathbf{x}_0$ (blue) towards noisy speech $\mathbf{y}$ (green), and the standard deviation (shaded gray region) increases exponentially. The reverse process moves back to $\mathbf{x}_0$, starting from a slightly mismatched distribution $\widetilde{p}_T$ which is centered around $\mathbf{y}$ rather than $\mathbf{x}_T$. Sample paths from both processes are shown as thin black lines. (Right) Time evolution of the SNR of the mean $\boldsymbol{\mu}$ (black) with respect to the SNR of $\mathbf{y}$ (green) for three different values of γ.

due to the cancellation of $\boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t)$, the loss function does not explicitly involve $\mathbf{y}$, only as an input to the score model. This means that the score model is not tasked with estimating any portion of the environmental noise directly. Finally, the minimization is achieved by optimizing the parameters θ using stochastic gradient descent.

D. Interpretation and Limitations

Let p_t be the distribution of the perturbed data $\mathbf{x}_t$ from the diffusion process for a given dataset. Then its time evolution can be thought of as a continuum of distributions $\{p_t\}_{t \in [0, T]}$ which is determined by the drift and the diffusion coefficient of the forward SDE. For fixed $\mathbf{x}_0$ and $\mathbf{y}$, this time evolution can be described in close form using (8), which we illustrate in Fig. 2 for a one-dimensional case. In the reverse process, the DNN has the task of learning this continuous family of distributions starting from $\tilde{p}_T$ as defined in (7). Due to the exponential increase of the diffusion coefficient in the forward process with initial condition $\sigma(0)^2 = 0$ the distribution p_0 essentially corresponds to the clean speech distribution, whereas the terminating distribution p_T is a strongly corrupted version of the noisy speech distribution. The particular characteristics in each noisy speech sample are strongly masked by the Gaussian white noise at $t = T$. Therefore, by learning the reverse process, the generative model learns a strong prior p_0 on clean speech, whereas the forward process terminates in a strongly corrupted distribution of the noisy speech, used as a weakly informative prior for generation. In Fig. 2, we simulate five sample paths from the diffusion process. All sample paths of the forward process start exactly at $\mathbf{x}_0$ but exhibit starkly different trajectories at large t. The reverse process should then turn a high-variance sample $\mathbf{x}_T$ back into a low-variance estimate of $\mathbf{x}_0$.

Looking at (9), we see that the mean $\boldsymbol{\mu}$ of the forward process exponentially decays from $\mathbf{x}_0$ to $\mathbf{y}$, which can also be seen in Fig. 2 (thick black line). However, for finite t it does not fully reach the corrupted speech $\mathbf{y}$ (dashed green line), particularly we have $\boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, T) \neq \mathbf{y}$. Thus, the final distribution of the forward process p_T exhibits a slight mismatch to the initial distribution of the reverse process $\tilde{p}_T$. We can make this mismatch arbitrarily small by either choosing a high stiffness parameter γ or by increasing $\sigma_{\max}$ to further smooth the density functions of both distributions. However, increasing γ would bring the mean close to $\mathbf{y}$ within a short time of the forward process, which may lead

to an unstable reverse process because only the last steps are concerned with removing environmental noise. This effect can be seen in Fig. 2 (right plot), where we plot the SNR of the process mean $\boldsymbol{\mu}$ averaged over 256 randomly selected files from the dataset for three values of γ. Note that we calculated the SNR in the time domain as the ratio of the power of clean speech to the power of environmental noise after inverting the non-linear amplitude transformation of (1). We see that while for $\gamma = 5$ the mismatch at $t = T$ becomes virtually zero, the change in SNR occurs mainly in the first half of the process. For $\gamma = 0.5$, on the other hand, the mismatch is already more than $10 \, \text{dB}$. However, the slope in the SNR is still apparent at the end of the process. Therefore, there is a trade-off to consider when choosing γ which depends on the dataset to be used. Increasing $\sigma_{\max}$ would come at the cost of more reverse iterations since the more white Gaussian noise is added, the less high-level information about the structure of the speech is preserved to serve as a guide in the reverse process. In the experiments, we choose a set of parameters based on empirical hyperparameter optimization.

III. NUMERICAL SDE SOLVERS

There exist several computational methods to find numerical solutions for SDEs, which are based on an approximation to discrete time steps. To this end, the interval $[0, T]$ is partitioned into N equal subintervals of width $\Delta t = T/N$, which approximates the continuous formulation into the discrete reverse process $\{\mathbf{x}_T, \mathbf{x}_{T-\Delta t}, \ldots, \mathbf{x}_0\}$. A common single-step method for solving this discretization is the Euler-Maruyama method. In each iteration step, the method refers to a previous state of the process and utilizes the drift and the Brownian motion to determine the current state.

In this work, we employ so-called predictor-corrector (PC) samplers proposed by Song et al. [28], which combine single-step methods for solving the reverse SDE with numerical optimization approaches such as annealed Langevin Dynamics [26]. PC samplers consist of two parts, a predictor and a corrector. The predictor can be any single-step method that aims to solve the reverse process by iterating through the reverse SDE. After each iteration step of the predictor, the current state of the process is refined by the corrector. The correction is based on Markov chain Monte Carlo sampling and can be understood as a stochastic gradient ascent optimizer that adds at each iteration step a small
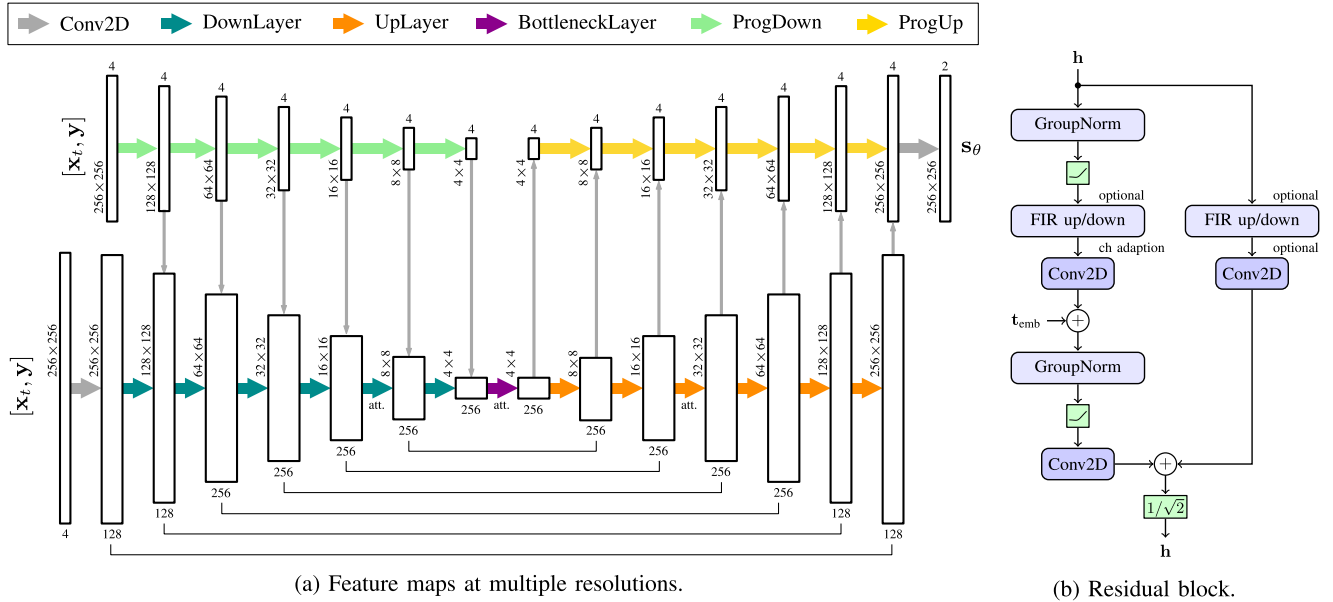
(a) Feature maps at multiple resolutions.

(b) Residual block.

Fig. 3. NCSN++ network architecture used as a score model $\mathbf{s}_\theta$: The architecture is based on a multi-resolution U-Net structure containing skip connections and an additional progressive growing path as shown in (a). Each up- and downsampling layer and the bottleneck layer consist of multiple residual blocks in series which are illustrated in (b).

amount of noise after taking a step in the direction of the estimated score. One possible intuition about the use of stochastic correctors is that they allow the process state to escape local minima by the use of stochasticity. However, Karras et al. [42] have recently argued that in the reverse process, stochasticity is only necessary to correct for numerical truncation errors of the predictor, a need that could be effectively circumvented by further improving the quality of the score model and predictor.

Another numerical way of approximating the reverse process is by solving the corresponding *probability flow ordinary differential equation (ODE)*,

$$\mathrm{d}\mathbf{x}_t = \left[-\mathbf{f}(\mathbf{x}_t, \mathbf{y}) + g(t)^2 \mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t)\right] \mathrm{d}t, \qquad (16)$$

which is the associated *deterministic process* of the stochastic reverse SDE in (6). It can be shown that for each diffusion process, there exists an ODE that describes the same marginal probability density $p_t(\mathbf{x}_t)$ [28]. Enhancing the noisy or reverberant mixture is then based on solving this ODE. In Section V-E, we also evaluate and compare this class of solvers for our task, specifically employing the Runge-Kutta method of fourth order with an error estimator of fifth order [43].

IV. NETWORK ARCHITECTURE

We utilize the *Noise Conditional Score Network* (NCSN++) architecture [28] for the score model $\mathbf{s}_\theta$ and adapt it for the use of complex spectrograms. For this purpose, we consider the real and imaginary parts of the complex input as separate channels, since the original network only works with real-valued numbers. Estimating both the real and imaginary parts of the score allows to generate complex spectrograms of clean speech.

The network is based on a multi-resolution U-Net structure, which has been experimentally shown to be powerful for tasks

such as generation and segmentation [44]. In Fig. 3(a), we illustrate the architecture by showing the feature maps at each resolution, indicating their spatial dimension and the corresponding number of channels. The transformations between the feature maps are represented by arrows, where the color of the arrow specifies the type of transformation (see the legend on top). We use Conv2D layers with a 3x3 kernel and stride 1 as input and output layers, and 1x1 Conv2D layers to aggregate information from the progressive growing path that we describe later. Up- and downsampling layers are based on residual network blocks which are taken from the BigGAN architecture [45], shown in Fig. 3(b). A residual block consists of Conv2D layers with the same configuration as above, group normalization [46], up- or downsampling with finite impulse response (FIR) filters [47], and the Swish activation function [48]. Each upsampling layer consists of three residual blocks and each downsampling layer of two residual blocks in series with the last block performing the up- or downsampling. Global attention mechanisms [49] are added at a resolution of 16×16 and in the bottleneck layer to better learn global dependencies within the feature maps.

To make the model time-dependent, information about the current progression of the diffusion process is fed into the network architecture. A common practice is to use Fourier-embeddings [49], i.e., a learned projection that maps the scalar time coordinate t to an M-dimensional vector $\mathbf{t}_{\mathrm{emb}}$ that is integrated into every residual block as can be seen in Fig. 3(b).

In addition to the main feature extraction path of the multi-resolution U-Net structure, the network incorporates a so-called progressive growing of the input which is seen at the top of Fig. 3(a). The idea is to provide a downsampled version of the input to every feature map in the contracting path, which has been successful in stabilizing high-resolution image generation [50].

Note that the downsampling operation in the progressive growing use shared weights for each resolution. The same procedure is also used in the expansive path where a progressive growing of the output is informed by the feature maps at each resolution, resulting in the final score estimate.

V. EXPERIMENTS

In this section, we describe the experimental setup for our speech enhancement and speech dereverberation experiments using the proposed method.

A. Datasets

For the evaluation of the speech enhancement task we use two datasets, the WSJ0-CHiME3 dataset and the VB-DMD dataset, which are described below. The use of two datasets allows cross-dataset evaluation, i.e. the test is performed on the other dataset than the one used for training. This mismatched condition reveals information about how well the method generalizes to unseen data with different characteristics such as distinct noise types or different recording conditions. Moreover, to train and evaluate our proposed method on the dereverberation task, we create the WSJ0-REVERB dataset, which is also described below.

1) WSJ0-CHiME3: We create the WSJ0-CHiME3 dataset using clean speech utterances from the Wall Street Journal (WSJ0) dataset [51] and noise signals from the CHiME3 dataset [52]. The mixture signal is created by randomly selecting a noise file and adding it to a clean utterance. Each utterance is used only once, and the SNR is sampled uniformly between 0 and 20 dB for the training, validation, and test set.

2) VB-DMD: We use the publicly available VoiceBank-DEMAND dataset (VB-DMD) [53] which is often used as a benchmark for single-channel speech enhancement. The utterances are artificially contaminated with eight real-recorded noise samples from the DEMAND database [54] and two artificially generated noise samples (babble and speech shaped) at SNRs of 0, 5, 10, and 15 dB. The test utterances are mixed with different noise samples at SNR levels of 2.5, 7.5, 12.5, and 17.5 dB. We split the training data into a training and validation set using speakers "p226" and "p287" for validation.

3) WSJ0-REVERB: To create the WSJ0-REVERB dataset, we use clean speech data from the WSJ0 dataset [51] and convolve each utterance with a simulated room impulse response (RIR). We use the `pyroomacoustics` engine [55] to simulate the RIRs. The reverberant room is modeled by sampling uniformly a T_{60} between 0.4 and 1.0 seconds. A dry version of the room is generated with the same geometric parameters but a fixed absorption coefficient of 0.99, to generate the corresponding anechoic target. The resulting average direct-to-reverberant ratio (DRR) is around -9 dB.

B. Instrumental Evaluation Metrics

To evaluate the performance of the proposed method we use standard metrics which we will describe in detail below. Metrics (a)–(d) employ full reference algorithms that rate the processed signal in relation to the clean reference signal using conventional digital signal analysis. On the other hand, metrics (e)-(g) are non-intrusive metrics that can be used to evaluate real recordings when the clean reference is unavailable.

1) POLQA: The Perceptual Objective Listening Quality Analysis (POLQA) is an ITU-T standard that includes a perceptual model for predicting speech quality [56]. The POLQA score takes values from 1 (poor) to 5 (excellent) as usual for mean opinion scores (MOS).

2) PESQ: The Perceptual Evaluation of Speech Quality (PESQ) is used for objective speech quality testing and is standardized in ITU-T P.862 [57]. Although it is the predecessor of POLQA, it is still widely used in the research community. The PESQ score lies between 1 (poor) and 4.5 (excellent) and there exist two variants, namely wideband PESQ and narrowband PESQ denoted as PESQ$_{nb}$.

3) ESTOI: The Extended Short-Time Objective Intelligibility (ESTOI) is an instrumental measure for predicting the intelligibility of speech subjected to various kinds of degradation [58]. The metric is normalized and lies between 0 and 1, with higher values indicating better intelligibility.

4) SI-SDR, SI-SIR, SI-SAR: Scale-Invariant (SI-) Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR) are standard evaluation metrics for single-channel speech enhancement and speech separation [59]. They are all measured in dB, with higher values indicating better performance.

5) DNSMOS: The Deep Noise Suppression MOS (DNSMOS) is a reference-free metric to evaluate perceptual speech quality [60]. The evaluation method uses a DNN that is trained on human ratings obtained by using an online framework for listening experiments [61] based on ITU-T P.808 [62].

6) SIG, BAK, OVRL: The non-intrusive speech quality assessment model DNSMOS P.835 [63] is based on a listening experiment according to ITU-T P.835 [64] and provides three MOS scores: speech quality (SIG), background noise quality (BAK), and the overall quality (OVRL) of the audio.

7) WVMOS: Wav-to-Vec MOS (WVMOS) [65] is a MOS prediction method for speech quality evaluation using a fine-tuned wav2vec2.0 model [66].

C. Listening Experiment

Instrumental evaluation metrics do not always correlate to human perception because there are many aspects of perception that are very difficult to capture by computational means. Therefore, we conduct a MUSHRA listening experiment [67] with ten participants using the webMUSHRA framework [68]. The participants were asked to rate the overall quality of twelve randomly sampled examples from the WSJ0-CHiME3 test set as reconstructed by the compared algorithms. The results are reported on a quality scale from 0 to 100.

D. Hyperparameters and Training Configuration

1) Input representation: We convert each audio input with sampling rate 16 kHz into a complex-valued STFT representation using a window size of 510, resulting in $F = 256$, a hop length of 128 (i.e. approximately 75% overlap), and a periodic

TABLE I
RESULTS FOR DIFFERENT SAMPLER CONFIGURATIONS TESTED ON VB-DMD
WITH THE AVERAGE NUMBER OF FUNCTION EVALUATIONS (NFE) AND THE
RESPECTIVE AVERAGE REAL-TIME FACTOR (RTF)[1]

| Type	Sampler settings	NFE	RTF	PESQ	SI-SDR [dB]
PC	0 corrector steps	30	0.89	2.80	15.38
PC	1 corrector steps	60	1.77	2.93	17.35
PC	2 corrector steps	90	2.65	2.92	17.52
ODE	atol=10^{-1}, rtol=10^{-1}	14	0.46	2.78	12.83
ODE	atol=10^{-6}, rtol=10^{-3}	49	1.55	2.71	12.76

(a) Varying N, fix $r = 0.33$ (b) Varying r, fix $N = 30$

Fig. 4. Model performance in PESQ and SI-SDR as a function of (a) the number of reverse steps N and (b) the step size parameter r in the annealed Langevin corrector.

Hann window. To process multiple examples for batch training, the length of each spectrogram is trimmed to $K = 256$ STFT time frames, with start and end times selected randomly at each training step. For the spectrogram transformation in (1), we have chosen $\alpha = 0.5$ and $\beta = 0.15$ empirically.

2) Stochastic process: The SDE in (2) is parameterized with $\sigma_{\min} = 0.05$, $\sigma_{\max} = 0.5$, and $\gamma = 1.5$ based on hyperparameter optimization with grid search.

3) Training configuration: We train the DNN on four Quadro RTX 6000 (24 GB memory each) for 160 epochs using the distributed data-parallel (DDP) approach in PyTorch Lightning [69], which takes about one day. We use the Adam optimizer [70] with a learning rate of 10^{-4} and an effective batch size of $4 \times 8 = 32$. We track an exponential moving average of the DNN weights with a decay of 0.999, to be used for sampling [71]. We log the average PESQ value of 20 randomly chosen examples from the validation set during training and select the best-performing model for evaluation.

E. Sampler Settings

To find optimal sampler settings for the reverse process, we run a hyperparameter search using the VB-DMD dataset.

1) Sampler type: We investigate which choice of sampler yields the best speech enhancement performance, comparing the PC sampler with different numbers of corrector steps and an ordinary ODE sampler as described in Section III. In Table I we can see that use of one correction step in the PC sampler seems to be advantageous, but the use of two steps does not lead to a further increase in performance. Thus, we decide to use the PC sampler with one corrector step for the evaluation. However, it should be noted that the use of one correction step doubles the number of function evaluations (NFE) of the sampler. The *function* being the expensive score model, this results in an average real-time factor (RTF) of 1.77, i.e., 1 sec of audio requires 1.77 sec of processing.[1] Comparing the PC sampler with the ODE sampler, we find that the PC sampler performs better in both metrics. However, with suitable settings, the ODE sampler requires only 14 NFE on average which results in an improved RTF of only 0.46.

2) Number of reverse steps N: The number of reverse steps N can be used to set a balance between the computational effort and the performance of the model. In Fig. 4(a), we show the
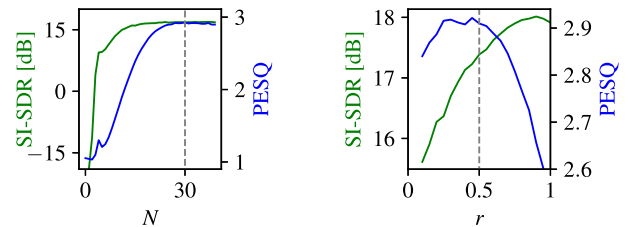
speech enhancement performance as a function of N. It can be seen that SI-SDR starts to stagnate earlier than PESQ. We opt for a value of $N = 30$, at which both metrics show no further increase in performance.

3) Step size in corrector: In Fig. 4(b), we vary the step size r of the annealed Langevin dynamics in the corrector. Interestingly, this parameter represents a compromise between PESQ and SI-SDR. We choose $r = 0.5$ to achieve a maximum PESQ value while still obtaining a good value for SI-SDR.

F. Baselines

We compare the performance of our proposed method with four generative and four discriminative baselines which we describe in more detail below. All methods are re-trained by us, except for DVAE, MetricGAN+, and CDiffuSE on VB-DMD, for which we obtained the pre-trained model from the authors who used the exact same training data.

1) STCN [11]: A generative VAE-based speech enhancement method which uses a stochastic temporal convolutional network (STCN) [72] that allows the latent variables to have both hierarchical and temporal dependencies. The parameters of the noise model and the latent variables are estimated using a Monte Carlo expectation maximization (MCEM) algorithm.

2) DVAE [17]: Generative speech enhancement method based on an unsupervised dynamical VAE (DVAE) [73] which models temporal dependencies between successive observable and latent variables. Parameters are updated at test time using a variational expectation maximization (VEM) method where the encoder is fine-tuned using stochastic gradient ascent.

3) CDiffuSE [21]: Most related to our proposed method is CDiffuSE, a generative speech enhancement method based on a conditional diffusion process defined in the time domain.

4) SGMSE [22]: Score-based Generative Model for Speech Enhancement (SGMSE) is our previous publication on which the proposed method is based. The main difference is that it uses a deep complex U-Net [74] instead of the NCSN++ architecture as the score model.

5) MetricGAN+ [75]: A discriminative speech enhancement method that uses a generator network for mask-based prediction of clean speech and introduces a discriminator network trained to approximate the PESQ score.

6) Conv-TasNet [76]: An end-to-end neural network that estimates a mask that is used for filtering a learned representation

---

[1] Average processing time for 10 audio files on an NVIDIA GeForce RTX 2080 Ti GPU, in a machine with an Intel Core i7-7800X CPU @ 3.50 GHz.

TABLE II
SPEECH ENHANCEMENT RESULTS OBTAINED FOR WSJ0-CHiME3 UNDER MATCHED AND MISMATCHED TRAINING CONDITIONS

Method	Type	Training set	POLQA	PESQ	ESTOI	SI-SDR [dB]	SI-SIR [dB]	SI-SAR[dB]	DNSMOS
Mixture	-	-	2.64 ± 0.68	2.01 ± 0.55	0.81 ± 0.12	13.5 ± 4.7	18.7 ± 5.5	15.4 ± 4.7	3.34 ± 0.37
STCN [11]	G	WSJ0	2.64 ± 0.68	2.01 ± 0.55	0.81 ± 0.12	13.5 ± 4.7	18.7 ± 5.5	15.4 ± 4.7	3.34 ± 0.37
RVAE [17]	G	WSJ0	2.97 ± 0.63	2.31 ± 0.55	0.85 ± 0.11	15.8 ± 5.0	21.6 ± 6.1	17.6 ± 4.9	3.61 ± 0.29
CDiffuse [21]	G	WSJ0-C3	3.08 ± 0.58	2.27 ± 0.51	0.83 ± 0.09	9.2 ± 2.3	19.8 ± 5.9	10.0 ± 2.3	3.43 ± 0.32
SGMSE [22]	G	WSJ0-C3	2.98 ± 0.60	2.28 ± 0.57	0.86 ± 0.09	14.8 ± 4.3	25.4 ± 5.6	15.3 ± 4.2	3.70 ± 0.27
SGMSE+	G	WSJ0-C3	$\mathbf{3.73 \pm 0.53}$	2.96 ± 0.55	0.92 ± 0.06	18.3 ± 4.4	$\mathbf{31.1 \pm 4.6}$	18.6 ± 4.5	$\mathbf{3.99 \pm 0.19}$
MetricGAN+ [75]	D	WSJ0-C3	3.52 ± 0.61	$\mathbf{3.03 \pm 0.45}$	0.88 ± 0.08	10.5 ± 4.5	24.5 ± 5.1	10.7 ± 4.6	3.67 ± 0.30
Conv-TasNet [76]	D	WSJ0-C3	3.65 ± 0.54	2.99 ± 0.58	$\mathbf{0.93 \pm 0.05}$	19.9 ± 4.3	29.2 ± 4.6	$\mathbf{20.6 \pm 4.5}$	3.79 ± 0.27
STCN [11]	G	VB	2.53 ± 0.66	1.80 ± 0.45	0.79 ± 0.12	11.9 ± 4.5	17.3 ± 4.9	13.8 ± 4.6	3.40 ± 0.34
RVAE [17]	G	VB	2.84 ± 0.61	2.08 ± 0.49	0.82 ± 0.11	13.9 ± 4.8	19.5 ± 5.9	15.8 ± 4.7	3.52 ± 0.31
CDiffuse [21]	G	VB-DMD	2.15 ± 0.57	1.79 ± 0.42	0.71 ± 0.11	3.2 ± 3.2	21.8 ± 7.0	3.4 ± 3.2	3.17 ± 0.29
SGMSE [22]	G	VB-DMD	2.66 ± 0.58	1.94 ± 0.47	0.81 ± 0.11	13.3 ± 4.3	23.5 ± 6.0	13.8 ± 4.2	3.76 ± 0.25
SGMSE+	G	VB-DMD	$\mathbf{3.43 \pm 0.61}$	$\mathbf{2.48 \pm 0.58}$	$\mathbf{0.90 \pm 0.07}$	$\mathbf{16.2 \pm 4.1}$	28.9 ± 4.6	$\mathbf{16.4 \pm 4.1}$	$\mathbf{4.00 \pm 0.19}$
MetricGAN+ [75]	D	VB-DMD	2.47 ± 0.67	2.13 ± 0.53	0.76 ± 0.12	6.8 ± 3.1	22.9 ± 4.9	7.0 ± 3.1	3.51 ± 0.29
Conv-TasNet [76]	D	VB-DMD	3.13 ± 0.60	2.40 ± 0.53	0.88 ± 0.08	15.2 ± 3.9	26.5 ± 4.6	15.6 ± 4.0	3.68 ± 0.30

Values indicate mean and standard deviation. Methods are sorted by the algorithm type, Generative (G) or Discriminative (D).

of the noisy mixture. The filtered representation is transformed back to the time domain by a learned decoder.

7) GaGNet [77]: This neural network is trained on a hybrid complex-domain and magnitude-domain regression objective for single-channel dereverberation. It uses so-called "glance" and "gaze" (GaG) modules, which respectively perform a coarse estimation of the magnitude and refine it with phase estimation in the complex domain.

8) TCN+SA+S [78]: This single-channel dereverberation approach uses a self-attention module to extract features from the input magnitude. This representation is then used by a temporal convolutional network followed by a single-layer convolutional smoother that outputs a magnitude estimate, which is used as the training objective. Griffin-Lim iterations are used to reconstruct the phase.

VI. RESULTS

A. Speech Enhancement

In Table II, we report the speech enhancement results on the WSJ0-CHiME3 test set for the matched and mismatched condition, i.e. when the training set was also WSJ0-CHiME3 or when the training set was VB-DMD. We compare our proposed method, which we call *SGMSE+*, with selected baseline methods and sort the results by the type of algorithm, which is either generative or discriminative. Considering the matched condition in the upper half of Table II, we see that SGMSE+ outperforms all other generative methods in all metrics. Note that STCN and RVAE are both unsupervised speech enhancement methods, i.e. they are trained on clean speech only (WSJ0 or VB). RVAE shows competitive results for SI-SAR, however, its VEM optimization algorithm is very time-consuming due to the fine-tuning of the encoder at test time, resulting in a RTF of >10000. This is significant in contrast to STCN with a RTF of 0.64 and SGMSE+ with a RTF of 1.77[1]. Although both VAE-based methods model temporal dependencies, they are limited in their ability to produce high-quality speech, likely due to the dimensionality reduction of the latent variable and

the encoder's sensitivity to noisy input, which causes the latent variable to be incorrectly initialized [15].

Comparing SGMSE+ to our previous model SGMSE, we find a significant improvement, especially for the perceptual metrics. We report improvements of 0.75 for POLQA and 0.68 for PESQ. This shows that the proposed generative diffusion process benefits significantly from the adapted network architecture. In our previous paper [22], we have already shown improvements over CDiffuSE for SGMSE in SI-SDR and SI-SAR, which we now back up with also reporting an improvement in ESTOI and DNSMOS and on par results in PESQ and POLQA. With SGMSE+, these improvements become even more significant, e.g. with 0.65 improvement in POLQA and 9.1 dB in SI-SDR compared to CDiffuSE. In qualitative analysis, we found that SGMSE+ is more accurate than CDiffuSE in preserving the high frequencies of fricatives after the completion of the reverse process. To compensate for that, CDiffuSE combines the enhanced files with the original noisy speech signal at a ratio of 0.2 for the final prediction [21]. This results in a trade-off between noise removal and the conservation of the signal. In our proposed approach, on the other hand, we found no significant suppression of high frequencies after completing the reverse process. Therefore, it is not necessary to mix back the noisy mixture to improve the signal quality, resulting in a significantly higher SI-SIR.

The comparison with Conv-TasNet and MetricGAN+ shows that SGMSE+ can keep up with the performance of discriminative methods and even surpasses them in terms of POLQA, SI-SIR, and DNSMOS. Discriminative methods are based on regression problems that optimize certain point-wise loss functions between the corrupted speech and a clean speech reference. For Conv-TasNet and MetricGAN+ these loss functions correspond to established intrusive metrics, namely SI-SDR for Conv-TasNet and PESQ for MetricGAN+. Note that both these discriminative methods shine in particular on the respective metric they used as a loss function. In contrast, generative methods like SGMSE+ are usually not trained to achieve the exact reconstruction of the reference clean speech but rather aim
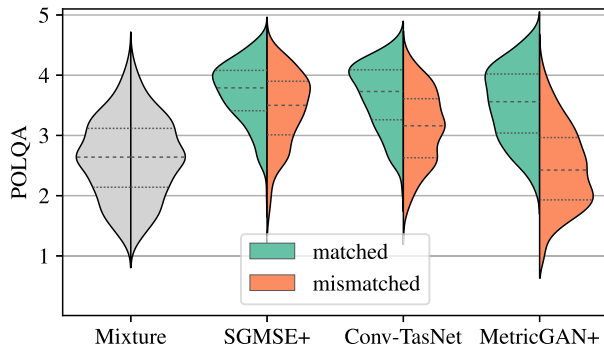
Fig. 5. Violin plots showing POLQA results for the matched and the mismatched condition with dashed and dotted lines representing median and quartiles, respectively.

Fig. 6. Boxplot showing the results of the MUSHRA listening experiment with ten participants on twelve randomly selected examples.

at generating a realization of speech that is on the manifold of clean speech. Thus, we suggest the use of non-intrusive metrics as a complementary measure since they allow an estimation of speech quality without relying on the exact reconstruction of a reference signal. In fact, for the non-intrusive metric DNSMOS our proposed method yields a significantly higher value than the discriminative baselines, indicating the strong ability of our generative model to generate high-quality clean speech.

Looking at the results for the mismatched condition in the bottom half of Table II, a general trend of decreasing metrics can be seen for all methods when compared to the corresponding values of the matched condition. This was to be expected since particular properties of the mismatched test set, such as distinct noise types or different recording characteristics of the clean speech have not been seen during training. However, generative methods generally show less degradation in the mismatched condition than discriminative methods. CDiffuSE is an exception, as this method shows significant degradation in the mismatched case. Informal listening reveals a problem with gain control, which is evident in strong volume fluctuations in the enhanced files. Furthermore, we see that SGMSE+ outperforms all other methods in all metrics under this condition, which shows the ability of our proposed method to generalize well.

Complementary to the average results above, we present in Fig. 5 violin plots of the full distribution of the POLQA scores obtained for SGMSE+, Conv-TasNet, MetricGAN+, and the noisy mixture for reference. For each method, the distributions are plotted side by side for the matched and mismatched conditions, so that the ability to generalize can be inferred from the horizontal alignment between both distributions. It can be seen that both distributions for SGMSE+ are relatively similar, whereas they are skewed for Conv-TasNet and especially for MetricGAN+.

In Fig. 6, we report the results of the MUSHRA listening experiment in a boxplot. On average, the ten participants rated the overall quality of our proposed approach with the highest score. In addition, our method remains fairly robust when the model was trained on a different training set, while discriminative methods show much stronger degradation for the mismatched condition. This also corresponds with the results of the non-intrusive metric DNSMOS in Table II and thus supports the use of non-intrusive methods for instrumental evaluation.
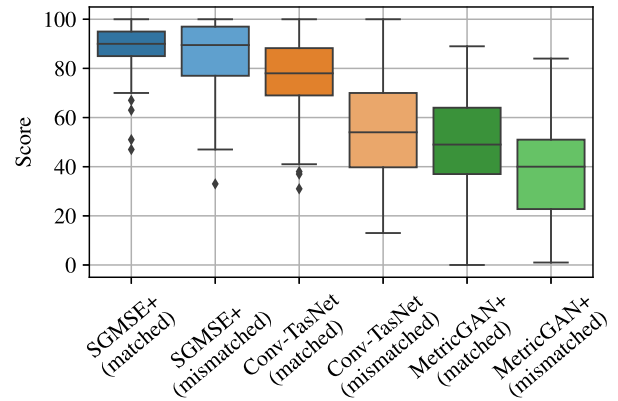
TABLE III
SPEECH ENHANCEMENT RESULTS OBTAINED FOR VB-DMD

Method	PESQ	PESQ$_{nb}$	ESTOI	SI-SDR	DNSMOS
Mixture	1.97	2.88	0.79	8.4	3.09
SEGAN* [8]	2.16	-	-	-	-
RVAE [17]	2.43	3.11	0.81	16.4	3.30
MetricGAN-U* [82]	2.45	-	0.77	8.2	-
CDiffuse [21]	2.52	3.31	0.79	12.4	3.09
SGMSE [22]	2.28	3.22	0.80	16.2	3.46
SGMSE+	2.93	**3.66**	**0.87**	17.3	**3.56**
UMX* [83]	2.35	-	0.83	14.0	-
Conv-TasNet [76]	2.63	3.42	0.85	**19.1**	3.37
MetricGAN+ [75]	**3.13**	3.63	0.83	8.5	3.37

Models marked with an asterisk (*) are additional baselines with values taken from the corresponding papers.

Interestingly, MetricGAN+ was only rated with a median score less than 50 for the matched condition, although the method performed best among all baselines for PESQ (see Table II). This reveals the discrepancy between the use of instrumental metrics for evaluation and people's actual perceptions. We suspect that MetricGAN+ has simply learned to utilize the internal operations of the PESQ algorithm to obtain a high value in this metric, neglecting the naturalness of the clean speech estimate. In fact, listening to the enhanced files, it can be recognized that the energy of the speech signal estimated by MetricGAN+ is mainly concentrated in the low- and mid-frequency area of the spectrogram, while high frequencies are strongly attenuated.

Listening to the enhanced files of our method, we notice that at very low input SNRs, some "vocalizing" artifacts with very poor articulation and no linguistic meaning are occasionally produced. In other examples, we find that breathing sounds or speech-like sounds were generated in noisy regions where no speech was originally present. These artifacts may also explain the outliers of our method in the listening experiment (see Fig. 6). For the matched condition, for example, the two lowest outliers come from the same utterance with clearly noticeable vocalizing artifacts. We hypothesize that these artifacts can be linked to the generative nature of the proposed approach. Indeed, for very noisy inputs, the score model may erroneously identify noise energy in some T-F areas as corrupted speech. The reverse diffusion process then produces speech where it did not originally exist. We argue that this behavior could be mitigated if

TABLE IV
SINGLE-CHANNEL DEREVERBERATION RESULTS OBTAINED FOR WSJ0-REVERB TEST SET

Method	Type	POLQA	PESQ	ESTOI	SI-SDR	SI-SIR	SI-SAR
Mixture	-	1.76 ± 0.29	1.36 ± 0.19	0.46 ± 0.12	-7.3 ± 5.5	-7.5 ± 5.4	-
SGMSE [22]	G	1.79 ± 0.28	1.35 ± 0.15	0.57 ± 0.07	-7.4 ± 5.8	-1.1 ± 7.0	-6.2 ± 5.5
SGMSE+	G	$\mathbf{3.24 \pm 0.46}$	$\mathbf{2.66 \pm 0.48}$	$\mathbf{0.84 \pm 0.07}$	$\mathbf{1.6 \pm 7.8}$	9.4 ± 10.2	$\mathbf{2.3 \pm 7.2}$
Conv-TasNet [76]	D	2.41 ± 0.52	1.84 ± 0.42	0.73 ± 0.10	$\mathbf{1.6 \pm 5.3}$	12.1 ± 5.1	1.9 ± 5.4
TCN+SA+S [78]	D	2.92 ± 0.33	2.29 ± 0.36	0.79 ± 0.05	-4.4 ± 5.3	-2.3 ± 5.2	-0.6 ± 5.1
GaGNet [77]	D	2.62 ± 0.47	1.98 ± 0.46	0.73 ± 0.08	-0.6 ± 4.9	6.1 ± 3.9	0.4 ± 5.1

Values indicate mean and standard deviation. Methods are sorted by the algorithm type which is either Generative (G) or Discriminative (D).

some conditioning with respect to speech activity and phoneme identity would be added to the score model.

Finally, Table III lists the results for the standardized VB-DMD dataset. This has the advantage that one can take values from other methods and copy them from the corresponding papers for a quick algorithmic comparison. It can be seen that SGMSE+ outperforms all other generative baselines, further narrowing the performance gap with discriminative methods that currently lead the benchmark based on PESQ,[2] including recent approaches such as [79] and [80]. It should however be noted that PESQ formally requires a minimum file length of 3.2 sec according to P.862.3 [81], which is not the case for most files in VB-DMD [53].

Investigating whether phase estimation has actually been improved with the modeling of the complex coefficients, we use the noisy phase in place of the estimated phase which does not show a significant performance difference. This is also in line with a recent study on the role of phase enhancement, where it has been shown that the impact of phase enhancement is rather small for ~ 32 ms spectral analysis frames but increasingly large with shorter frame lengths [84].

B. Dereverberation

We report in Table IV the performance of our approach when trained and tested on a single-channel dereverberation task. We compare with SGMSE [22] and three discriminative baselines, namely Conv-TasNet [76], GaGNet [77] and TCN+SA+S [78].

Our proposed SGMSE+ approach performs particularly well in terms of instrumental metrics compared to all other baseline models. The low average input DRR of -9 dB constitutes a real challenge for discriminative approaches, which do not manage to separate the reverberation from the target without distorting the target signal, resulting in low-quality scores. On the other hand, our approach benefits from generative modeling and is able to reconstruct speech with very high quality in most cases. When comparing our previous SGMSE model [22] with SGMSE+, we see that for speech dereverberation, the method benefits greatly from the improved network architecture. This effect is even more significant than for additive background noise removal in the speech enhancement task.

In particular, using the proposed approach SGMSE+ on a single-channel dereverberation task does not produce any of the vocalized artifacts observed in the speech enhancement

[2][Online]. Available: https://paperswithcode.com/sota/speech-enhancement-on-demand

TABLE V
SPEECH ENHANCEMENT RESULTS OBTAINED FOR REAL-WORLD NOISY RECORDINGS FROM THE DNS CHALLENGE 2020 TEST SET

Method	DNSMOS	SIG	BAK	OVRL	WVMOS
Mixture	3.05	3.05	2.51	2.26	1.12
RVAE [17]	3.29	3.16	2.91	2.44	1.87
CDiffuse [21]	3.14	3.15	3.19	2.55	1.86
SGMSE [22]	3.38	3.22	3.02	2.52	1.80
SGMSE+	**3.64**	**3.42**	**3.82**	**3.04**	**2.54**
Conv-TasNet [76]	3.07	2.87	3.59	2.52	2.07
MetricGAN+ [75]	3.26	2.88	3.39	2.45	1.52

experiments for low input SNRs. Although the reverberant signal is formally decorrelated in the time domain from the target by the randomness of reflections across the room, it still originates from the dry speech source. Therefore, we conjecture that the score model effectively detects whether the energy in a particular time-frequency area is associated with the clean speech nearby that needs to be reconstructed.

C. Evaluation on Real Data

Complementing the experiments using simulated data, we evaluate the speech enhancement performance on real-world noisy recordings. For real-world noisy recordings, there exists no clean speech reference. Thus, we can only non-intrusive metrics to evaluate the perceptual speech quality which we describe in Section V-B (e)–(g). For the evaluation, we use 300 files from the test set of the Deep Noise Suppression (DNS) Challenge 2020 [85]. In Table V, we report the results for models that were trained on VB-DMD. It turns out that our proposed method performs better than all other methods in all non-intrusive metrics, demonstrating its robustness to real-world noisy examples. Interestingly, a trend of degradation in speech quality (SIG) can be observed for the discriminative methods, whereas all generative models improve this metric with respect to the mixture. For the background noise quality (BAK) metric, on the other hand, discriminative models seem to perform well, yet our proposed method performs superior. It is important to note that non-intrusive metrics do not require a corresponding clean reference signal and only assess speech quality based on the method's estimate. We hypothesize that our generative model works well on these metrics, as it was trained to generate clean speech. However, "vocalizing" artifacts as mentioned above or phonetic confusions may not be captured with these metrics.

We provide on our project page[3] some listening examples for all evaluated tasks. Furthermore, we include real reverberant examples from the MC-WSJ-AV dataset [86].

VII. CONCLUSION

In this work, we built upon our existing work [22] that uses a novel stochastic diffusion process to design a generative model for speech enhancement in the complex STFT domain. We presented an extended theoretical analysis of the underlying score-based generative model and derived in detail the objective function used for training. In further explorations, we considered the time evolution of the conditional diffusion process which revealed a slight mismatch between the forward and reverse process, which can be adjusted with a careful parameterization of the forward SDE.

By using an adopted network architecture, we were able to significantly improve the performance compared to our previous model. In addition, we trained and evaluated the proposed method on the task of speech dereverberation and show significantly superior performance compared to discriminative baseline methods. Hence, we showed that with our proposed method, a single framework can be used to train individual models for different distortion types. For the task of speech enhancement, we evaluated performance under matched and mismatched conditions, i.e. when the training and test data were taken from the same or different corpora. For the matched condition, the proposed generative speech enhancement method performs on par with competetive discriminative methods. For the mismatched condition, our method shows strong generalization capabilities and outperforms all baselines in all metrics, as confirmed by a listening experiment. In very adverse conditions, however, we observe that the proposed method sometimes introduces vocalizing and breathing artifacts. We argue that these could be mitigated in future work if some conditioning concerning speech activity and phoneme information would be added to the score model.

In addition, we explored different sampling strategies to solve the reverse process at test time which allows us to balance the performance and computational speed of the proposed method. Future work could include other sampling techniques to further reduce the number of diffusion steps [87] and thus the computational complexity.

ACKNOWLEDGMENT

We would like to thank J. Berger and Rohde&Schwarz SwissQual AG for their support with POLQA.

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*. San Rafael, CA, USA: Morgan & Claypool, 2013.

[2] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech*, E. Enhancement, T. V. Virtanen, and S. Gannot, Eds., Hoboken, NJ, USA: Wiley, 2018.

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

[4] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[5] S.-W. Fu, T.-Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.

[6] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. IEEE Asia-Pacific Signal Inf. Proc. Assoc. Annu. Summit Conf.*, 2017, pp. 006–012.

[7] P. Wang, K. Tan, and D. L. Wang, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 39–48, 2020.

[8] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Proc. ISCA Interspeech*, 2017, pp. 3642–3646.

[9] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, pp. 716–720, 2018.

[10] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *Proc. IEEE 28th Int. Workshop Mach. Learn. Signal Process.*, 2018, pp. 1–6.

[11] J. Richter, G. Carbajal, and T. Gerkmann, "Speech enhancement with stochastic temporal convolutional networks," in *Proc. ISCA Interspeech*, 2020, pp. 4516–4520.

[12] G. Carbajal, J. Richter, and T. Gerkmann, "Guided variational autoencoder for speech enhancement with a supervised classifier," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 681–685.

[13] G. Carbajal, J. Richter, and T. Gerkmann, "Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement," in *Proc. IEEE Workshop Appl. Signal Proccess. Audio Acoust.*, 2021, pp. 126–130.

[14] Y. Bando, K. Sekiguchi, and K. Yoshii, "Adaptive neural speech enhancement with a denoising variational autoencoder," in *Proc. ISCA Interspeech*, 2020, pp. 2437–2441.

[15] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 676–680.

[16] A. A. Nugraha, K. Sekiguchi, and K. Yoshii, "A flow-based deep latent variable model for speech spectrogram modeling and enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1104–1117, 2020.

[17] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2993–3007, 2022.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Representations*, 2014.

[19] I. Goodfellow et al., "Generative adversarial networks," *Adv. Neural Inf. Process. Syst.*, vol. 27, pp. 2672–2680, 2014.

[20] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2021, pp. 659–666.

[21] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7402–7406.

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

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

[24] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[25] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 6840–6851, 2020.

[26] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Adv. Neural Inf. Process. Syst.*, pp. 11918–11930, 2014, 2019.

[27] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching," *J. Mach. Learn. Res.*, vol. 6, no. 4, pp. 695–709, 2005.

[28] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Representations*, 2021.

[29] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Process. Their Appl.*, vol. 12, no. 3, pp. 313–326, 1982.

[30] Y. Koizumi, H. Zen, K. Yatabe, N. Chen, and M. Bacchiani, "SpecGrad: Diffusion probabilistic model based neural vocoder with adaptive noise spectral shaping," in *Proc. ISCA Interspeech*, 2022, pp. 803–807.

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

[32] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[33] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2010, pp. 1–4.

[34] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *Proc. IEEE Int. Conf. Telecom. Signal Process.*, 2021, pp. 72–76.

[35] C. H. You, S. N. Koh, and S. Rahardja, "/spl beta/-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 13, no. 4, pp. 475–486, Jul. 2005.

[36] C. Breithaupt and R. Martin, "Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 277–289, Feb. 2011.

[37] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," in *Proc. Int. Conf. Learn. Representations*, 2021.

[38] G. Batzolis, J. Stanczuk, C.-B. Schönlieb, and C. Etmann, "Conditional image generation with score-based diffusion models," 2021, *arXiv:2111.13606*.

[39] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *Adv. Neural Inf. Process. Syst.*, 2021, pp. 8780–8794.

[40] C.-W. Huang, J. H. Lim, and A. C. Courville, "A variational perspective on diffusion-based generative models and score matching," *Adv. Neural Inf. Process. Syst.*, 2021, pp. 22863–22876.

[41] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations*. Cambridge, U.K: Cambridge Univ. Press, 2019.

[42] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *Adv. Neural Inf. Process. Syst.*, 2022.

[43] J. R. Dormand and P. J. Prince, "A family of embedded Runge-Kutta formulae," *J. Comput. Appl. Math.*, vol. 6, pp. 19–26, 1980.

[44] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2015, pp. 234–241.

[45] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Representations*, 2018.

[46] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[47] R. Zhang, "Making convolutional networks shift-invariant again," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7324–7334.

[48] P. Ramachandran, B. Zoph, and Q. V. Le, "Swish: A self-gated activation function," 2017, *arXiv:1710.05941*.

[49] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, pp. 5998–6008, 2017.

[50] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8110–8119.

[51] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete," 1993. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S6A

[52] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 504–511.

[53] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proc. ISCA Speech Synth. Workshop*, 2016, pp. 146–152.

[54] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multichannel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. 3591–3591, 2013.

[55] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 351–355.

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

[57] 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," *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2001, vol. 2, pp. 749–752.

[58] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

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

[60] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6493–6497.

[61] B. Naderi and R. Cutler, "An open source implementation of ITU-T recommendation P.808 with validation," in *Proc. ISCA Interspeech*, 2020, pp. 2862–2866.

[62] ITU-T Rec. P.808, "Subjective evaluation of speech quality with a crowdsourcing approach," in *Proc. Int. Telecommun. Union*, 2021. [Online]. Available: https://www.itu.int/rec/T-REC-P.808-202106-I/en

[63] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 886–890.

[64] ITU-T Rec. P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," in *Proc. Int. Telecommun. Union*, 2003. [Online]. Available: https://www.itu.int/rec/T-REC-P.835-200311-I/en

[65] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi : A unified framework for bandwidth extension and speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, 2022, pp. 1–5.

[66] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Adv. Neural Inf. Process. Syst.*, 2020, pp. 12449–12460.

[67] *Method for the subjective assessment of intermediate quality level of audio systems*, Rec. BS.1534-3, International Telecommunication Union, Geneva, Switzerland, 2014. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1534

[68] M. Schoeffler et al., "Webmushra–A comprehensive framework for web-based listening tests," *J. Open Res. Softw.*, vol. 6, no. 1, 2018.

[69] W. Falcon et al., "Pytorch lightning," *GitHub*, vol. 3, 2019. https://github.com/PyTorchLightning/pytorch-lightning

[70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[71] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," *Adv. Neural Inf. Process. Syst.*, 2020, pp. 12438–12448.

[72] E. Aksan and O. Hilliges, "Stcn: Stochastic temporal convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2018.

[73] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda, "Dynamical variational autoencoders: A comprehensive review," *Found. Trends Mach. Learn.*, vol. 15, no. 1/2, pp. 1–175, 2021.

[74] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," 2019, *arXiv:1903.03107*.

[75] S.-W. Fu et al., "MetricGAN : An improved version of MetricGAN for speech enhancement," in *Proc. Interspeech*, 2021, pp. 201–205, doi: 0.21437/Interspeech.2021-599.

[76] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[77] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Appl. Acoust.*, vol. 187, 2022, Art. no. 108499.

[78] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1598–1607, 2020.

[79] G. Yu, A. Li, H. Wang, Y. Wang, Y. Ke, and C. Zheng, "DBT-Net: Dual-branch federative magnitude and phase estimation with attention-in-attention transformer for monaural speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2629–2644, 2022.

[80] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based metric GAN for speech enhancement," in *Proc. ISCA Interspeech*, 2022, pp. 936–940.

[81] ITU-T Rec. P.862.3, *Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2*, Rec. P.862.3, International Telecommunication Union, Geneva, Switzerland, 2007. [Online]. Available: https://www.itu.int/rec/T-REC-P.862.3/en

[82] S.-W. Fu, C. Yu, K.-H. Hung, M. Ravanelli, and Y. Tsao, "MetricGAN-U: Unsupervised speech enhancement/dereverberation based only on noisy/reverberated speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7412–7416.

[83] S. Uhlich and Y. Mitsufuji, "Open-unmix for speech enhancement (UMX SE)," 2020. [Online]. Available: https://github.com/sigsep/open-unmix-pytorch

[84] T. Peer and T. Gerkmann, "Phase-aware deep speech enhancement: It's all about the frame length," *JASA Exp. Lett.*, vol. 2, no. 10, 2022, Art. no. 104802.

[85] C. K. Reddy et al., "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. ISCA Interspeech*, 2020, pp. 2492–2496.

[86] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2005, pp. 357–362.

[87] D. Watson, W. Chan, J. Ho, and M. Norouzi, "Learning fast samplers for diffusion models by differentiating through sample quality," in *Proc. Int. Conf. Learn. Representations*, 2021.

Jean-Marie Lemercier (Student Member, IEEE) received the M.Eng. degree in electrical engineering in 2019 from Ecole Polytechnique, Paris, France. In 2020, he received the M.Sc. degree in communications and signal processing from Imperial College London, London, U.K. He is currently with the Signal Processing group, Universität Hamburg, Hamburg, Germany, under the supervision of Prof. Dr.-Ing. Timo Gerkmann. His research interests include machine learning-based speech enhancement and dereverberation for hearing devices applications. His recent works also include the design and analysis of diffusion-based generative models for various speech restoration tasks.

Julius Richter (Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering in 2017 and 2019, respectively, from the Technical University of Berlin, Berlin, Germany. He is currently with the Signal Processing Group, Universität Hamburg, Hamburg, Germany, under the supervision of Prof. Dr.-Ing. Timo Gerkmann. His research interests include deep generative models and multi-modal learning with applications to audio-visual speech processing.

Simon Welker (Student Member, IEEE) received the B.Sc. degree in computing in science and M.Sc. degree in bioinformatics degrees in 2019 and 2021, respectively, from Universität Hamburg, Hamburg, Germany. He is currently a second-year Ph.D. student in signal processing with the labs of Prof. Timo Gerkmann, Universität Hamburg and Prof. Henry N. Chapman Center for Free-Electron Laser Science, DESY, Hamburg, researching machine learning techniques for solving inverse problems that arise in speech processing and X-ray imaging.

Bunlong Lay received the B.Sc. and M.Sc. degrees in mathematics in 2015 and 2017 respectively, from the University of Bonn, Bonn, Germany. From 2018 to 2021, he was with Research Institute Fraunhofer FKIE, Wachtberg, Germany, where he focused on research in the field of electronic warfare. In 2021 he started working toward the Ph.D. degree with the University of Hamburg, Hamburg, Germany. He is currently researching on diffusion-based models for Speech Enhancement.

Timo Gerkmann (Senior Member, IEEE) is a Professor for Signal Processing with the Universität Hamburg, Hamburg, Germany. He has held positions with Technicolor Research & Innovation, University of Oldenburg, Oldenburg, Germany, KTH Royal Institute of Technology, Stockholm, Sweden, Ruhr-Universität Bochum, Bochum, Germany, and Siemens Corporate Research, Princeton, NJ, USA. His research interests include statistical signal processing and machine learning for speech and audio applied to communication devices, hearing instruments, audio-visual media, and human-machine interfaces. He was the recipient of the VDE ITG award 2022. Timo Gerkmann is a Member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and the Senior Area Editor of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING.