Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain

Simon Welker\textsuperscript{1,2,†}, Julius Richter\textsuperscript{1,†}, Timo Gerkmann\textsuperscript{1}

\textsuperscript{1}Signal Processing (SP), Universität Hamburg, Germany
\textsuperscript{2}Center for Free-Electron Laser Science, DESY, Hamburg, Germany

\textsuperscript{†}Authors contributed equally to this work.

Abstract

Score-based generative models (SGMs) have recently shown impressive results for difficult generative tasks such as the unconditional and conditional generation of natural images and audio signals. In this work, we extend these models to the complex short-time Fourier transform (STFT) domain, proposing a novel training task for speech enhancement using a complex-valued deep neural network. We derive this training task within the formalism of stochastic differential equations (SDEs), thereby enabling the use of predictor-corrector samplers. We provide alternative formulations inspired by previous publications on using generative diffusion models for speech enhancement, avoiding the need for any prior assumptions on the noise distribution and making the training task purely generative which, as we show, results in improved enhancement performance.

Index Terms: speech enhancement, generative modeling, score-based generative models, deep learning

1. Introduction

Speech enhancement aims at estimating clean speech signals from audio recordings that are impacted by acoustic noise [1]. The task is well studied in the signal processing literature, and conventional approaches often make assumptions regarding the statistical properties of speech signals [2]. With the advent of deep learning, approaches to speech enhancement have made significant progress in the last decade. Most methods are based on a discriminative learning task that aims to minimize a certain distance between clean and noisy speech. However, since supervised methods are inevitably trained on a finite set of training data with limited model capacity for practical reasons, they may not generalize to unseen situations, e.g., different noise types, reverberation, and different signal-to-noise ratios (SNRs). In addition, some discriminative approaches have been shown to result in unpleasant speech distortions that outweigh the benefits of noise reduction [3].

Instead of learning a direct mapping from noisy to clean speech, generative models aim to learn the distribution of clean speech as a prior for speech enhancement. Several approaches have utilized deep generative models for speech enhancement using generative adversarial networks (GANs) [4, 5], variational autoencoders (VAEs) [6–11], flow-based models [12], and more recently generative diffusion models [13, 14]. The main principle of these approaches is to learn the inherent properties of clean speech, such as its spectral and temporal structure, which are then used as prior knowledge for making inferences about clean speech given a noisy input. Thus, they are trained solely to generate clean speech and are therefore considered more robust to different acoustic environments compared to their discriminative counterparts. In fact, generative approaches have shown to perform better under mismatched training and test conditions [10, 11, 14, 15]. However, they are currently less studied and still lag behind discriminative approaches, which is a strong incentive to conduct more research to realize their full potential.

Although GANs and VAEs have become a popular choice for deep generative modeling, generative diffusion models [16–18] provide a powerful alternative which has recently shown to be state-of-the-art in natural image generation [19]. The idea is to gradually turn data into noise, and to train a neural network that learns to invert this procedure for different noise scales (see Fig. 1). Other works have adopted this scheme for generating speech in the time domain using clean Mel spectrograms as a conditioner [20, 21]. Interestingly, the model in [20] investigated its zero-shot speech denoising capabilities for different noise types. Although it was only trained to remove white noise added in the diffusion process, it already shows preliminary results for performing speech enhancement. Lu et al. built upon this idea and designed a supportive reverse process using the same architecture but with noisy spectrograms as the conditioner [13]. In their follow-up paper they devised a conditional diffusion process with an adopted forward and reverse process incorporating the noisy data into the diffusion process [14]. However, the derivation of their objective function is based on the assumption that the global distribution of the additive noise follows a standard white Gaussian, which is normally not the case for environmental noise. Moreover, the neural network used in their reverse process is trained to explicitly estimate the difference between clean and noisy speech, which is usually considered a discriminative learning task.

We acknowledge the support by DASHH (Data Science in Hamburg - HELMHOLTZ Graduate School for the Structure of Matter) with the Grant No. HIDSS-0002 and the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169).
In this work, we propose an alternative diffusion process which is based on stochastic differential equations (SDEs) for score-based generative models (SGMs) [18] which is a specific class of generative diffusion models (see Section 2). Using this formulation, we are able to avoid the need for any prior assumptions on the noise distribution and make the training task purely generative. Unlike [13, 14] which work in the time domain, we perform generative speech enhancement in the complex STFT domain, i.e., working directly on the Fourier coefficients with amplitude and phase. This choice should allow deep-learning models to exploit the rich structure of speech in the time-frequency domain. In contrast to approaches that modify only the amplitudes, this choice avoids the need for phase retrieval techniques, since the phase information is enhanced directly as well. In our experiments, we demonstrate good speech enhancement performance and show that our method introduces less speech distortions compared to the baseline method.

2. Background

2.1. Score-based generative models

SGMs [16, 18] are generative diffusion models that center around the idea of corrupting training data with slowly increasing levels of noise (forward process), and training a score model to reverse this corruption (reverse process) by estimating the score \( \nabla_x \log p_{\text{data}}(x) \), i.e., the gradient of the log probability density with respect to the data. Once a score model is trained, an iterative procedure called Langevin dynamics can be used to draw samples from it [22].

2.2. SGMs through stochastic differential equations

Song et al. [18] introduced the formalism of using SDEs for score-based generative modeling. Following the Itô interpretation [23, ch. 4], the process \( \{x_t\} \) can be defined by the SDE

\[
\text{d}x_t = f(x_t, t)\text{d}t + g(t)\text{d}w,
\]

where \( f \) is the drift, \( g \) is the diffusion, \( t \) is a coordinate which denotes how far the forward process has run, \( \text{d}t \) is an infinitesimal time step, and \( w \) is the standard Wiener process. We stress that the “time” \( t \) mentioned in this context is a conceptual idea related to the progression of the process, and is not related to the time axis of an audio signal or its STFT. According to Anderson [24], every such SDE has a corresponding reverse SDE

\[
\text{d}x_t = \left[ f(x_t, t) - g(t)^2 \nabla_x \log p_t(x_t) \right] \text{d}t + g(t)\text{d}w,
\]

where \( \text{d}t \) is an infinitesimal negative time step and \( \nabla_x \log p_t(x_t) \) is with respect to the log density of the process at time \( t \) and can be approximated by a learned time-dependent score model \( s_t(x_t, t) \). The reverse SDE is then solved by means of some solver procedure (see Section 3.4), providing the basis for score-based generative modeling with SDEs.

3. Proposed method

3.1. A stochastic process for speech enhancement

Lu et al. [14] proposed the idea of linearly interpolating between clean speech \( x_0 \) (at \( t = 0 \)) and noisy speech \( y = x_0 + n \) (at \( t = 1 \)) along a discrete-time forward process, so that the reverse process should lead to clean speech at \( t = 0 \). Note, that discrete-time diffusion processes using Markov chains [17] are different from the continuous SDE formulation used in the context of SGMs [18] and in this work. Furthermore, as can be seen from their subsequent derivations, the choice of linear interpolation implies that the trained deep neural network (DNN) must explicitly estimate a portion of environmental noise \( n \) at each step in the reverse process. From an intuitive standpoint, this occurs because it is necessary to estimate the slope between every value of \( x_0 \) and \( y \) to be able to formulate the reverse process. As a consequence, the resulting evidence lower bound (ELBO) loss [14, eq. (21)] exhibits characteristics of a discriminative learning task, making it difficult to assess whether the network is learning a prior for clean speech per se or the mapping from noisy to clean speech. To avoid this, we propose the following novel diffusion process based on an SDE, which leads to a pure generative training objective without the model estimating any portion of \( n \):
and noisy speech \( y \) as elements of \( \mathbb{C}(\mathbb{F}^{-1}) \times \mathbb{T} \), where \( \mathbb{F} \) is the discrete Fourier transform (DFT) length and \( \mathbb{T} \) is the number of time frames, dependent on the audio length. Since the global distribution of STFT speech amplitudes is typically heavy-tailed [27], the information visible in untransformed spectrograms is dominated by only a small portion of bins. Furthermore, the naturally encountered amplitudes often lie far outside the interval \([0, 1]\) often used in SGMs [17, 18]. As an engineering trick, we thus apply an amplitude transformation to all complex STFT coefficients \( c \) [28], in an effort to bring out frequency components with lower energy (e.g. fricative sounds of unvoiced speech) and to normalize amplitudes roughly to within \([0, 1]\) to better fit the usual SGM assumptions. The transformation and its associated inverse transformation is defined as:

\[
\tilde{c} = \frac{|c|^\alpha}{\beta} e^{i\alpha(c)} \Leftrightarrow c = \beta |c|^{\frac{1}{\alpha}} e^{i\alpha(c)}
\]

where we have chosen \( \alpha = 0.5 \) and \( \beta = 3 \) empirically. The forward and backward Gaussian processes, as well as the input and output of the score-based DNN model, are then formulated and applied within this transformed domain.

### 3.3. Training task

We adopt the training strategy of denoising score matching [29], which trains a model to approximate the score \( \nabla_x \log p_0(x) \) by estimating the Gaussian noise added by the forward process at time \( t \). This training strategy can be efficiently realized by sampling \( t \) uniformly from \([0, 1]\) and then sampling \( x \), according to (5) for each data point in the training batch. The training task is to learn parameters \( \theta \) that minimize the following term, where \( x_0 \sim p_0(x) \), \( x_1 \sim p_0(x_1|x_0, y) \):

\[
E_{x_0, x_1, x_0 \sim p_0} \left[ \| s(x_1, t, y) - \nabla_x \log p_0(x_1|x_0, y) \|^2 \right]
\]

We note that in the baseline work, CDiffuseSE [14, eq. (21)], the loss function includes a term \((y - x_0)\), i.e., the network is in part trained to remove the environmental noise directly in each step. In contrast, our objective function in (9) does not task the DNN with estimating a portion of the environmental noise \( n \), and \( y = x_0 + n \) enters (9) only as an input to the score-based model, functioning as a conditioning signal. Our model is thus trained to only estimate the Gaussian noise that is artificially added during the forward process. We therefore argue that the CDiffuseSE training task has significant discriminative characteristics, whereas ours remains purely generative in nature.

### 3.4. Speech enhancement procedure

To proceed with speech enhancement, an initial complex spectrogram \( x_1 \) is determined by sampling from the corresponding prior distribution, with mean \( y \) and standard deviation \( \sigma(1) \):

\[
x_1 \sim \mathcal{N}(y, \sigma(1)^2 \mathbf{I})
\]

We then choose a reverse sampling method to start from \( x_1 \) and solve the reverse SDE, from \( t = 1 \) up until a small \( t_s \approx 0 \) to avoid numerical issues close to \( t = 0 \), see [18]. There are various possible choices of reverse sampling methods. Song et al. [18] show impressive performance of so-called Predictor-Corrector samplers, which we therefore adopt as well. We use the specific combination of reverse diffusion sampling as the predictor and annealed Langevin dynamics as the corrector, as in [18, Algorithm 2]. The corrector requires a so-called SNR parameter \( r \), which we empirically set to \( r = 0.33 \). We use one corrector step per iteration, at 50 total iterations. The enhanced audio is finally retrieved by applying the backward transformation (8) and a subsequent inverse STFT.

### 4. Experimental setup

#### 4.1. Neural network architecture

For our experiments in this paper, we construct a complex-valued U-Net architecture based on blocks shown schematically in Fig. 2 and parameterized as listed in Table 1, resulting in 3.56M parameters overall. We adapt the Deep Complex U-Net (DCUNet) architecture [30] in several ways for our task: (1) We insert time-embedding layers into all encoder and decoder blocks, providing the DNN with information about the time-step \( t \). We encode \( t \) via 128 random Fourier feature embeddings [31], which are passed through complex-valued affine layers and activations in each block. These embeddings are broadcasted over and added to each respective channel of the layer output, as in previous SGM-based work [17,18,20]. (2) We add multiple non-strided encoder/decoder pairs at the top of the U-Net to increase the DNN’s capacity for finer, non-downsampled features. (3) We add exponentially increasing dilations in the frequency axis to the lower layers. This enlarges the receptive field in the frequency direction, to help the DNN differentiate between spectral characteristics encountered in speech without affecting the amount of required computation. (4) We change the network to have one complex-valued output channel estimating the score, and two complex-valued input channels \((x_1, y)\), so that it has access to the information contained in the noisy speech \( y \). (5) We replace all \( 3 \times 3 \) kernels by \( 4 \times 4 \) kernels, to avoid checkerboard patterns arising from the kernel size not being divisible by the stride in the transposed convolutions [32].

#### 4.2. Dataset

We use the standardized VoiceBank-DEMAND [33, 34] dataset for training and testing, as was done in the baseline work (Diffuse, [13]). We normalize the pairs of clean and noisy audio \((x_0, y)\) by the maximum absolute value of \( x_0 \). We then convert
Table 1: Encoder and decoder parameters of our modified DCUNet architecture. \( C_i \) / \( C_o \) are input / output channels, \( K \) is the kernel size, \( S \) is the stride, and \( D \) is the dilation. Tuples indicate the axes (frequency, time).

<table>
<thead>
<tr>
<th>Depth</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_i )</td>
<td>2</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>64</td>
<td>128</td>
</tr>
<tr>
<td>( C_o )</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
</tr>
<tr>
<td>( K )</td>
<td>(4.4, 4.4)</td>
<td>(4.4, 4.4)</td>
<td>(4.4, 4.4)</td>
<td>(4.4, 4.4)</td>
<td>(4.4, 4.4)</td>
<td>(4.4, 4.4)</td>
</tr>
<tr>
<td>( S )</td>
<td>(1.1, 1.1)</td>
<td>(1.1, 1.1)</td>
<td>(2.1, 2.1)</td>
<td>(2.2, 2.2)</td>
<td>(4.1, 4.1)</td>
<td>(8.1, 8.1)</td>
</tr>
</tbody>
</table>

each input into the complex-valued one-sided STFT representation, using a DFT length of \( F = 512 \), a hop length of 128 (i.e., 75% overlap) and a periodic Hann window. We randomly crop each spectrogram to 256 STFT time frames during each epoch.

4.3. Training procedure and hyperparameters

We train our DNN for 325 epochs, using the Adam optimizer [35] with a learning rate of \( 10^{-4} \) and a batch size of 32. We parameterize our SDE \( (3) \) as follows: \( \gamma = 1.5, \sigma_{\text{min}} = 0.05, \sigma_{\text{max}} = 0.5, \nu = 0.03 \). We track an exponential moving average of the DNN weights with a decay of 0.999, to be used for sampling [36]. We also train two baseline models [13,14] to compare against, using publicly available code.

4.4. Evaluation

For evaluation, we follow the speech enhancement procedure described in Section 3.4. We report the scale-invariant signal-to-distortion ratio (SI-SDR), signal-to-interference ratio (SI-SIR) and signal-to-artifacts ratio (SI-SAR) [37], comparing against the noisy speech \( y \) and the baseline method. We avoid the use of the Perceptual Evaluation of Speech Quality (PESQ) measure, since the P.862.3 standard [38] states that “there should be a minimum of 3.2s active speech in the reference”, which does not hold for the majority of audio files in the VoiceBank-DEMAND [33] dataset used by us and (C)DiffuSE [13,14]. We provide listening examples in order to assess the perceptual quality of the compared methods\(^1\).

5. Results

In Table 2, we compare the performance of our proposed method (SGMSE) with DiffuSE [13], CDiffuSE [14] and the noisy mixture. The results from CDiffuSE, for which executable code was not available at the time of submission, were an extended journal article with additional analysis and improved performance is currently in preparation [39].

Table 2: Average performance of our method (SGMSE) and DiffuSE [13], comparing their raw output to the noisy speech mixture. Best values in each column are bold. Values for \( m = (0.8 \tilde{e} + 0.2 y) \), as used in [13,14], are listed in gray.

<table>
<thead>
<tr>
<th>Model</th>
<th>SI-SDR</th>
<th>SI-SIR</th>
<th>SI-SAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixture</td>
<td>8.4 ± 38</td>
<td>8.4 ± 38</td>
<td>∞</td>
</tr>
<tr>
<td>DiffuSE(^{[13]})</td>
<td>10.5 ± 14</td>
<td>30.0 ± 71</td>
<td>10.8 ± .11</td>
</tr>
<tr>
<td>CDiffuSE(^{[14]})</td>
<td>12.1 ± 10</td>
<td>28.2 ± 36</td>
<td>12.3 ± .09</td>
</tr>
<tr>
<td>SGMSE</td>
<td>15.1 ± 27</td>
<td>24.9 ± 42</td>
<td>15.7 ± .25</td>
</tr>
<tr>
<td>DiffuSE_m(^{[13]})</td>
<td>11.4 ± 18</td>
<td>19.1 ± .46</td>
<td>13.0 ± .11</td>
</tr>
<tr>
<td>CDiffuSE_m(^{[14]})</td>
<td>12.6 ± 15</td>
<td>18.8 ± 34</td>
<td>14.4 ± .09</td>
</tr>
<tr>
<td>SGMSE_m</td>
<td>14.5 ± .29</td>
<td>17.9 ± .36</td>
<td>17.7 ± .25</td>
</tr>
</tbody>
</table>

\( p257\_398\).wav: “I was very young.”

Figure 3: Power spectrograms for an example audio file from VoiceBank-DEMAND. As can be seen from the highlighted regions, our approach (SGMSE) exhibits fewer voice distortions than CDiffuSE [14] at some expense of noise reduction.

and noisy) and the corresponding estimates of our method and CDiffuSE. Both methods are able to remove the environmental noise, which is clearly visible in the specified dynamic range of 50 dB. Furthermore, it can be seen that our method preserves the high frequencies of the fricatives better than CDiffuSE. Interestingly, CDiffuSE removes too much energy around the formants, whereas our method maintains the natural structure.

6. Conclusion

In this work, we have designed a novel stochastic process for a score-based generative modeling approach to speech enhancement in the complex STFT domain, based on the formalism of SDEs. To our knowledge, this is the first work to apply SGMs in the complex time-frequency domain and, after CDiffuSE [14], the second theoretically principled approach to use SGMs specifically for speech enhancement. Our approach exhibits more natural sounding reconstructions with fewer artifacts than previous methods at some expense of noise reduction, which together leads to an SI-SDR improvement of about 0.6 dB over DiffuSE and 0 dB over CDiffuSE when comparing their raw output to the noisy speech mixture. Further investigation into other SDEs, data representations and DNN architectures may prove to be a fruitful research avenue for speech enhancement using generative models. An extended journal article with additional analysis and improved performance is currently in preparation [39].

\(^1\)https://uhh.de/inf-sp-sgmse
7. References


