Diffusion-based Speech Enhancement: Demonstration of Performance and Generalization

Julius Richter*
Signal Processing Group
University of Hamburg
Germany
julius.richter@uni-hamburg.de

Timo Gerkmann
Signal Processing Group
University of Hamburg
Germany
timo.gerkmann@uni-hamburg.de

Abstract

This demo presents advanced techniques in speech enhancement using deep generative models. It highlights the generalization capabilities of score-based generative models for speech enhancement and compares directly with Schrödinger bridge approaches. The presented methods focus on generating high-quality super-wideband speech at a sampling rate of 48 kHz. Participants will record speech using a single microphone in a noisy environment, such as a conference venue. These recordings will then be enhanced and played back through headphones, demonstrating the model's effectiveness in improving speech quality and intelligibility.

1 Introduction

Generative speech enhancement has recently shown promising advancements in improving speech quality in noisy environments [1]. Several diffusion-based frameworks exist, such as score-based generative models for speech enhancement (SGMSE) [2] and the Schrödinger bridge [3, 4].

This demo showcases the generalization capabilities of diffusion-based models for speech enhancement. A comparative evaluation is presented for SGMSE and the Schrödinger bridge approach. All presented methods generate high-quality super-wideband speech enhancement at a sampling rate of 48 kHz. Participants are asked to record their speech using a single microphone in a noisy environment, such as a conference venue. Once the speech is recorded, it undergoes enhancement and is subsequently played back through headphones.

2 Methods

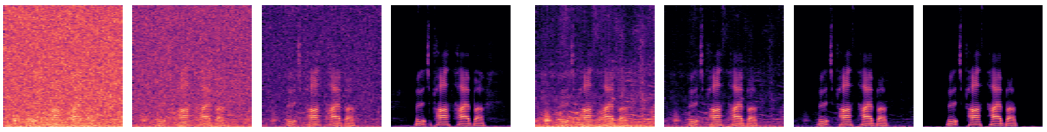
We represent audio signals in the time-frequency domain using the short-time Fourier transform (STFT). Thus, we have complex spectrograms $\bar{\mathbf{x}} \in \mathbb{C}^{T \times F}$ where T is the number of time frames, and F is the number of frequency bins. We apply an amplitude transformation $\widetilde{c} = \beta |c|^{\alpha} \mathrm{e}^{i \angle(c)}$ to all STFT coefficients c, where $\angle(\cdot)$ represents the angle of a complex number, $\alpha \in (0, 1]$, and $\beta \in \mathbb{R}$ [2]. Finally, we flatten $\bar{\mathbf{x}}$ into a vector of dimension $d = TF$, resulting in the input representation $\mathbf{x} \in \mathbb{C}^d$.

2.1 Score-based generative models for speech enhancement (SGMSE)

In SGMSE [2], we use a task-adapted diffusion process for the conditional generation of clean speech $\mathbf{x}_0 \in \mathbb{C}^d$ given a noisy input $\mathbf{y} \in \mathbb{C}^d$. Using the continuous-time formulation for diffusion models [5],

*We acknowledge the support by the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



(a) SDE solver for SGMSE. (b) ODE solver for the Schrödinger bridge.

Figure 1: Comparison between SGMSE and Schrödinger bridge.

	POLQA	**SI-SDR** [dB]	**PESQ**	**ESTOI**	**SIGMOS**	**DNSMOS**
Noisy	1.71 ± 0.56	5.98 ± 6.10	1.24 ± 0.22	0.49 ± 0.15	1.95 ± 0.39	2.74 ± 0.29
SGMSE+ [2]	3.40 ± 0.73	16.78 ± 4.47	2.50 ± 0.62	$\mathbf{0.73 \pm 0.13}$	3.41 ± 0.41	$\mathbf{3.88 \pm 0.26}$
SB [3]	3.46 ± 0.84	$\mathbf{17.85 \pm 4.37}$	2.33 ± 0.67	$\mathbf{0.73 \pm 0.13}$	$\mathbf{3.44 \pm 0.42}$	3.83 ± 0.27
SB-PESQ [4]	$\mathbf{3.71 \pm 0.80}$	16.29 ± 4.16	$\mathbf{3.09 \pm 0.63}$	$\mathbf{0.73 \pm 0.13}$	3.18 ± 0.44	3.72 ± 0.28

Table 1: Speech enhancement performance on EARS-WHAM. Mean and standard deviation.

the forward process is modeled as the solution to the stochastic differential equation (SDE)

$$\mathrm{d}\mathbf{x}_t = \gamma(\mathbf{y} - \mathbf{x}_t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}\,, \tag{1}$$

where $\mathbf{x}_t \in \mathbb{C}^d$ denotes the process state at time $t \in [0, 1]$, $\gamma \in \mathbb{R}$ controls the transition from $\mathbf{x}_0$ to $\mathbf{y}$, and $g(t) \in \mathbb{R}$ is the diffusion coefficient that controls the amount of Gaussian noise induced by a standard Wiener process $\mathbf{w}$.

The forward process can be time-inverted [5], resulting in a corresponding reverse process

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

where $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})$ is the conditional score, and $\bar{\mathbf{w}}$ is the Wiener process backward in time.

The score function is typically intractable and approximated by a score model $\mathbf{s}_\theta$ with parameters θ. Once the score model is trained, the reverse process in Eq. (2) can be solved by replacing the score function with its approximation $\mathbf{s}_\theta$ and using a numerical SDE solver. Fig. 1a illustrates an example of reverse sampling utilizing an SDE solver. We use the same hyperparameters as in [6].

2.2 Schrödinger bridge for speech enhancement

The Schrödinger bridge (SB) for speech enhancement involves a pair of symmetric forward and reverse SDEs [3]. The general SB formulation is typically intractable, but closed-form solutions exist for certain cases, such as Gaussian boundary conditions [7].

Training the SB involves directly predicting the clean data $\mathbf{x}_0$, which differs from SGMSE, where Gaussian noise is predicted instead. This allows the incorporation of perceptual loss terms [4]. The reverse SDE is solved during inference using an ODE sampler [7]. The SB approach benefits from its optimal transport characteristics, interpolating efficiently between clean and noisy signals, as shown in Fig. 1b. We use the same hyperparameters as in [4], the same STFT as in [6], and $\alpha_{\mathrm{P}} = 10^{-4}$.

3 Experimental Setup and Results

We train all models using the EARS-WHAM dataset [6], which comprises 86.8 hours of audio. We report results on the EARS-WHAM test set in Table 1. In the demo, speech is recorded using a microphone connected to a sound card, which is in turn connected to a laptop. The laptop is equipped with a GPU to facilitate efficient processing. After the audio enhancement process is complete, we play back the enhanced audio through headphones. An example of enhanced audio from a noisy YouTube video is available at the following link: `https://www.youtube.com/watch?v=H5FiO0JxPK4`.

4 Conclusion

This demo illustrates the advanced capabilities of generative speech enhancement, showcasing its ability to improve speech quality and intelligibility in real-world, noisy environments. We highlight the generation of high-quality super-wideband speech, thereby emphasizing the potential of these techniques for practical applications.

References

[1] Jean-Marie Lemercier, Julius Richter, Simon Welker, Eloi Moliner, Vesa Välimäki, and Timo Gerkmann. Diffusion models for audio restoration. *Signal Processing Magazin, accepted paper*, 2025.

[2] Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2351–2364, 2023.

[3] Ante Jukić, Roman Korostik, Jagadeesh Balam, and Boris Ginsburg. Schrödinger bridge for generative speech enhancement. In *ISCA Interspeech*, pages 1175–1179, 2024.

[4] Julius Richter, Danilo de Oliveira, and Timo Gerkmann. Investigating training objectives for generative speech enhancement. *arXiv preprint arXiv:2409.10753*, 2024.

[5] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021.

[6] Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinjii Watanabe, Alexander Richard, and Timo Gerkmann. EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. In *ISCA Interspeech*, pages 4873–4877, 2024.

[7] Zehua Chen, Guande He, Kaiwen Zheng, Xu Tan, and Jun Zhu. Schrödinger bridges beat diffusion models on text-to-speech synthesis. *arXiv preprint arXiv:2312.03491*, 2023.