Diffusion-based speech enhancement in matched and mismatched conditions using a Heun-based sampler

P. Gonzalez[1] Z.-H. Tan[2] J. Østergaard[2] J. Jensen[2] T. S. Alstrøm[3] T. May[1]

[1]Department of Health Technology
Technical University of Denmark

[2]Department of Electronic Systems
Aalborg University

[3]Department of Applied Mathematics and Computer Science
Technical University of Denmark

April 16th, 2024

Introduction – Diffusion models

- Diffusion models are a new class of generative models that have recently been applied to speech enhancement[1,2]
- They iteratively add noise to the training data and learn to undo this process

[1] Y.-J. Lu *et al.*, "Conditional diffusion probabilistic model for speech enhancement," *ICASSP*, 2022
[2] J. Richter *et al.*, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, 2023

Introduction – Diffusion models

- Diffusion models are a new class of generative models that have recently been applied to speech enhancement[1,2]
- They iteratively add noise to the training data and learn to undo this process

Forward process

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{f}(t, \boldsymbol{x}_t)\,\mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{\omega}_t \longrightarrow$$

[1] Y.-J. Lu *et al.*, "Conditional diffusion probabilistic model for speech enhancement," *ICASSP*, 2022

[2] J. Richter *et al.*, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, 2023

Introduction – Diffusion models

- Diffusion models are a new class of generative models that have recently been applied to speech enhancement[1][2]
- They iteratively add noise to the training data and learn to undo this process

Forward process

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{f}(t, \boldsymbol{x}_t)\,\mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{\omega}_t$$

- **Noise schedule**: defined by the hyperparameters $f(t, \boldsymbol{x}_t)$ and $g(t)$

[1] Y.-J. Lu *et al.*, "Conditional diffusion probabilistic model for speech enhancement," *ICASSP*, 2022
[2] J. Richter *et al.*, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, 2023

Introduction – Diffusion models

- Diffusion models are a new class of generative models that have recently been applied to speech enhancement[1][2]
- They iteratively add noise to the training data and learn to undo this process

Forward process

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{f}(t, \boldsymbol{x}_t)\,\mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{\omega}_t$$

$$\mathrm{d}\boldsymbol{x}_t = \left[\boldsymbol{f}(t, \boldsymbol{x}_t) - g(t)^2 \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)\right] \mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{\omega}_t$$

Reverse process

- **Noise schedule**: defined by the hyperparameters $f(t, \boldsymbol{x}_t)$ and $g(t)$

[1] Y.-J. Lu *et al.*, "Conditional diffusion probabilistic model for speech enhancement," *ICASSP*, 2022

[2] J. Richter *et al.*, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, 2023

Introduction – Diffusion models

- Diffusion models are a new class of generative models that have recently been applied to speech enhancement[1][2]
- They iteratively add noise to the training data and learn to undo this process

Forward process

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{f}(t, \boldsymbol{x}_t)\,\mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{\omega}_t$$

$$\mathrm{d}\boldsymbol{x}_t = \left[\boldsymbol{f}(t, \boldsymbol{x}_t) - g(t)^2 \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)\right]\mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{\omega}_t$$

Reverse process

- **Noise schedule**: defined by the hyperparameters $f(t, \boldsymbol{x}_t)$ and $g(t)$
- **Sampler**: numerical method used to integrate the reverse process

[1] Y.-J. Lu *et al.*, “Conditional diffusion probabilistic model for speech enhancement,” *ICASSP*, 2022
[2] J. Richter *et al.*, “Speech enhancement and dereverberation with diffusion-based generative models,” *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, 2023

Introduction – Adaptation to speech enhancement

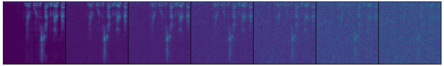
- Output image $\longleftrightarrow$ Clean spectrogram
  Text prompt $\longleftrightarrow$ Noisy spectrogram

Introduction – Adaptation to speech enhancement

- Output image $\longleftrightarrow$ Clean spectrogram
  Text prompt $\longleftrightarrow$ Noisy spectrogram

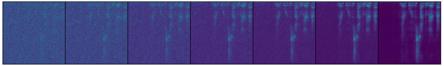
Forward process

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{f}(t, \boldsymbol{x}_t)\,\mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{\omega}_t \longrightarrow$$



Reverse process

$$\mathrm{d}\boldsymbol{x}_t = \left[-\boldsymbol{f}(t, \boldsymbol{x}_t) - g(t)^2 \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)\right] \mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{\omega}_t \longrightarrow$$



Introduction – Adaptation to speech enhancement

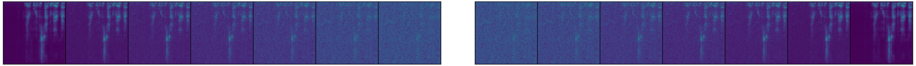
- Output image $\longleftrightarrow$ Clean spectrogram
  Text prompt $\longleftrightarrow$ Noisy spectrogram
- Provoke a drift towards the noisy spectrogram[1 2]

Forward process

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{f}(t, \boldsymbol{x}_t)\,\mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{\omega}_t \longrightarrow$$

Reverse process

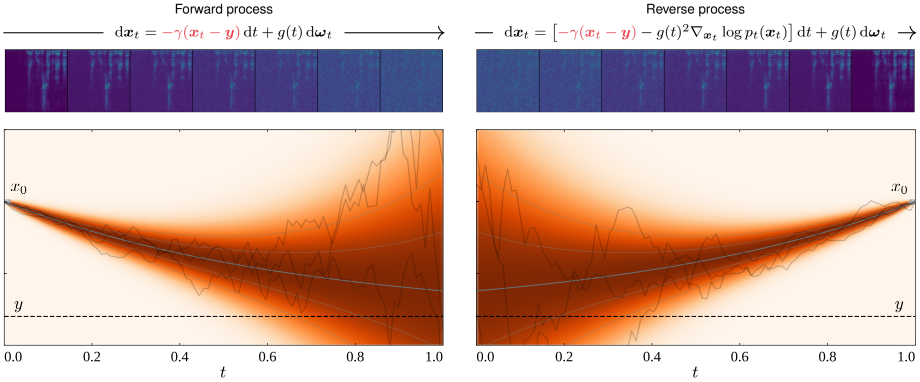
$$\mathrm{d}\boldsymbol{x}_t = \left[-\boldsymbol{f}(t, \boldsymbol{x}_t) - g(t)^2 \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)\right] \mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{\omega}_t \longrightarrow$$



[1] S. Welker *et al.*, "Speech enhancement with score-based generative models in the complex STFT domain," *INTERSPEECH*, 2022
[2] J. Richter *et al.*, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, 2023

Introduction – Adaptation to speech enhancement

- Output image ⟷ Clean spectrogram
  Text prompt ⟷ Noisy spectrogram
- Provoke a drift towards the noisy spectrogram[1 2]



[1] S. Welker *et al.*, "Speech enhancement with score-based generative models in the complex STFT domain," *INTERSPEECH*, 2022
[2] J. Richter *et al.*, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, 2023

Motivation

- **Goal**: apply findings from image generation literature[1] to improve speech enhancement performance
  - Neural network preconditioning based on first principles
  - Second-order Heun-based sampler

[1] T. Karras *et al.*, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, 2022

Motivation

- **Goal**: apply findings from image generation literature[1] to improve speech enhancement performance
  - Neural network preconditioning based on first principles
  - Second-order Heun-based sampler
- **Problem**: the drift coefficient $-\gamma(\boldsymbol{x}_t - \boldsymbol{y})$ cannot be written in the form $f(t)\boldsymbol{x}_t$

[1] T. Karras *et al.*, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, 2022

Motivation

- **Goal**: apply findings from image generation literature[1] to improve speech enhancement performance
  - Neural network preconditioning based on first principles
  - Second-order Heun-based sampler
- **Problem**: the drift coefficient $-\gamma(\boldsymbol{x}_t - \boldsymbol{y})$ cannot be written in the form $f(t)\boldsymbol{x}_t$
- **Solution**: change of variable $\boldsymbol{n}_t = \boldsymbol{x}_t - \boldsymbol{y}$

$$\mathrm{d}\boldsymbol{x}_t = -\gamma(\boldsymbol{x}_t - \boldsymbol{y})\,\mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{\omega}_t$$

$$\Updownarrow$$

$$\mathrm{d}\boldsymbol{n}_t = -\gamma\boldsymbol{n}_t\,\mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{\omega}_t$$

Or more generally,

$$\mathrm{d}\boldsymbol{n}_t = f(t)\boldsymbol{n}_t\,\mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{\omega}_t$$

[1] T. Karras *et al.*, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, 2022

Motivation

- **Goal**: apply findings from image generation literature[1] to improve speech enhancement performance
  - Neural network preconditioning based on first principles
  - Second-order Heun-based sampler
- **Problem**: the drift coefficient $-\gamma(\boldsymbol{x}_t - \boldsymbol{y})$ cannot be written in the form $f(t)\boldsymbol{x}_t$
- **Solution**: change of variable $\boldsymbol{n}_t = \boldsymbol{x}_t - \boldsymbol{y}$

$$\mathrm{d}\boldsymbol{x}_t = -\gamma(\boldsymbol{x}_t - \boldsymbol{y})\,\mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{\omega}_t$$

$$\Updownarrow$$

$$\mathrm{d}\boldsymbol{n}_t = -\gamma\boldsymbol{n}_t\,\mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{\omega}_t$$

Or more generally,

$$\mathrm{d}\boldsymbol{n}_t = f(t)\boldsymbol{n}_t\,\mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{\omega}_t$$

- This allows to write

$$p(\boldsymbol{n}_t|\boldsymbol{n}_0) = \mathcal{N}(\boldsymbol{n}_t; s(t)\boldsymbol{n}_0, s(t)^2\sigma(t)^2\mathbf{I})$$

where

$$s(t) = \exp\int_0^t f(\xi)\,\mathrm{d}\xi \quad \text{and} \quad \sigma(t)^2 = \int_0^t \frac{g(\xi)^2}{s(\xi)^2}\,\mathrm{d}\xi$$

[1] T. Karras *et al.*, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, 2022

Experimental setup

- Modifications to SGMSE+M[1]:
  - Shifted-cosine noise schedule[2]
  - Neural network preconditioning[3]
  - Heun-based sampler[3] instead of PC sampler[4]

[1] J.-M. Lercercier *et al.*, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," *ICASSP*, 2023
[2] E. Hoogeboom *et al.*, "simple diffusion: End-to-end diffusion for high resolution images," *ICML*, 2023
[3] T. Karras *et al.*, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, 2022
[4] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2021

Experimental setup

- Modifications to SGMSE+M[1]:
  - Shifted-cosine noise schedule[2]
  - Neural network preconditioning[3]
  - Heun-based sampler[3] instead of PC sampler[4]
- Evaluation in matched and mismatched conditions:
  - Noisy speech is generated using 5 speech corpora, 5 noise databases and 5 room impulse response databases

Speech:	TIMIT	LibriSpeech	WSJ	Clarity	VCTK
Noise:	TAU	NOISEX	ICRA	DEMAND	ARTE
Room:	Surrey	ASH	BRAS	CATT	AVIL

[1] J.-M. Lercercier *et al.*, “Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration,” *ICASSP*, 2023
[2] E. Hoogeboom *et al.*, “simple diffusion: End-to-end diffusion for high resolution images,” *ICML*, 2023
[3] T. Karras *et al.*, “Elucidating the design space of diffusion-based generative models,” *NeurIPS*, 2022
[4] Y. Song *et al.*, “Score-based generative modeling through stochastic differential equations,” *ICLR*, 2021

Experimental setup

- Modifications to SGMSE+M[1]:
  - Shifted-cosine noise schedule[2]
  - Neural network preconditioning[3]
  - Heun-based sampler[3] instead of PC sampler[4]

- Evaluation in matched and mismatched conditions:
  - Noisy speech is generated using 5 speech corpora, 5 noise databases and 5 room impulse response databases
  - $N = 1$ or $N = 4$ databases are used for the training condition and the remaining databases are used for the mismatched condition
  - Results are averaged across 5 different combinations

Speech:	TIMIT	LibriSpeech	WSJ	Clarity	VCTK
Noise:	TAU	NOISEX	ICRA	DEMAND	ARTE
Room:	Surrey	ASH	BRAS	CATT	AVIL

[1] J.-M. Lemercier *et al.*, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," *ICASSP*, 2023
[2] E. Hoogeboom *et al.*, "simple diffusion: End-to-end diffusion for high resolution images," *ICML*, 2023
[3] T. Karras *et al.*, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, 2022
[4] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2021

Experimental setup

- Modifications to SGMSE+M[1]:
  - Shifted-cosine noise schedule[2]
  - Neural network preconditioning[3]
  - Heun-based sampler[3] instead of PC sampler[4]
- Evaluation in matched and mismatched conditions:
  - Noisy speech is generated using 5 speech corpora, 5 noise databases and 5 room impulse response databases
  - $N = 1$ or $N = 4$ databases are used for the training condition and the remaining databases are used for the mismatched condition
  - Results are averaged across 5 different combinations

	Training	Testing			
Speech:	TIMIT	LibriSpeech	WSJ	Clarity	VCTK
Noise:	TAU	NOISEX	ICRA	DEMAND	ARTE
Room:	Surrey	ASH	BRAS	CATT	AVIL

[1] J.-M. Lemercier *et al.*, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," *ICASSP*, 2023
[2] E. Hoogeboom *et al.*, "simple diffusion: End-to-end diffusion for high resolution images," *ICML*, 2023
[3] T. Karras *et al.*, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, 2022
[4] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2021

Experimental setup

- Modifications to SGMSE+M[1]:
  - Shifted-cosine noise schedule[2]
  - Neural network preconditioning[3]
  - Heun-based sampler[3] instead of PC sampler[4]
- Evaluation in matched and mismatched conditions:
  - Noisy speech is generated using 5 speech corpora, 5 noise databases and 5 room impulse response databases
  - $N = 1$ or $N = 4$ databases are used for the training condition and the remaining databases are used for the mismatched condition
  - Results are averaged across 5 different combinations

Speech:	TIMIT	LibriSpeech	WSJ	Clarity	VCTK
Noise:	TAU	NOISEX	ICRA	DEMAND	ARTE
Room:	Surrey	ASH	BRAS	CATT	AVIL

Training
Testing

[1] J.-M. Lemercier *et al.*, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," *ICASSP*, 2023
[2] E. Hoogeboom *et al.*, "simple diffusion: End-to-end diffusion for high resolution images," *ICML*, 2023
[3] T. Karras *et al.*, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, 2022
[4] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2021

Experimental setup

- Modifications to SGMSE+M[1]:
  - Shifted-cosine noise schedule[2]
  - Neural network preconditioning[3]
  - Heun-based sampler[3] instead of PC sampler[4]
- Evaluation in matched and mismatched conditions:
  - Noisy speech is generated using 5 speech corpora, 5 noise databases and 5 room impulse response databases
  - $N = 1$ or $N = 4$ databases are used for the training condition and the remaining databases are used for the mismatched condition
  - Results are averaged across 5 different combinations

	Testing		Training	Testing	
Speech:	TIMIT	LibriSpeech	WSJ	Clarity	VCTK
Noise:	TAU	NOISEX	ICRA	DEMAND	ARTE
Room:	Surrey	ASH	BRAS	CATT	AVIL

Training
Testing

[1] J.-M. Lemercier *et al.*, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," *ICASSP*, 2023
[2] E. Hoogeboom *et al.*, "simple diffusion: End-to-end diffusion for high resolution images," *ICML*, 2023
[3] T. Karras *et al.*, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, 2022
[4] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2021

Experimental setup

- Modifications to SGMSE+M[1]:
  - Shifted-cosine noise schedule[2]
  - Neural network preconditioning[3]
  - Heun-based sampler[3] instead of PC sampler[4]

- Evaluation in matched and mismatched conditions:
  - Noisy speech is generated using 5 speech corpora, 5 noise databases and 5 room impulse response databases
  - $N = 1$ or $N = 4$ databases are used for the training condition and the remaining databases are used for the mismatched condition
  - Results are averaged across 5 different combinations

Speech:	TIMIT	LibriSpeech	WSJ	Clarity	VCTK
Noise:	TAU	NOISEX	ICRA	DEMAND	ARTE
Room:	Surrey	ASH	BRAS	CATT	AVIL

Training

Testing

[1] J.-M. Lemercier *et al.*, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," *ICASSP*, 2023
[2] E. Hoogeboom *et al.*, "simple diffusion: End-to-end diffusion for high resolution images," *ICML*, 2023
[3] T. Karras *et al.*, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, 2022
[4] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2021

Experimental setup

- Modifications to SGMSE+M[1]:
  - Shifted-cosine noise schedule[2]
  - Neural network preconditioning[3]
  - Heun-based sampler[3] instead of PC sampler[4]

- Evaluation in matched and mismatched conditions:
  - Noisy speech is generated using 5 speech corpora, 5 noise databases and 5 room impulse response databases
  - $N = 1$ or $N = 4$ databases are used for the training condition and the remaining databases are used for the mismatched condition
  - Results are averaged across 5 different combinations

Speech:	TIMIT	LibriSpeech	WSJ	Clarity	VCTK
Noise:	TAU	NOISEX	ICRA	DEMAND	ARTE
Room:	Surrey	ASH	BRAS	CATT	AVIL

Training
Testing

[1] J.-M. Lemercier *et al.*, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," *ICASSP*, 2023
[2] E. Hoogeboom *et al.*, "simple diffusion: End-to-end diffusion for high resolution images," *ICML*, 2023
[3] T. Karras *et al.*, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, 2022
[4] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2021

Experimental setup

- Modifications to SGMSE+M[1]:
  - Shifted-cosine noise schedule[2]
  - Neural network preconditioning[3]
  - Heun-based sampler[3] instead of PC sampler[4]

- Evaluation in matched and mismatched conditions:
  - Noisy speech is generated using 5 speech corpora, 5 noise databases and 5 room impulse response databases
  - $N = 1$ or $N = 4$ databases are used for the training condition and the remaining databases are used for the mismatched condition
  - Results are averaged across 5 different combinations

Speech:	TIMIT	LibriSpeech	WSJ	Clarity	VCTK
Noise:	TAU	NOISEX	ICRA	DEMAND	ARTE
Room:	Surrey	ASH	BRAS	CATT	AVIL

Training

Testing

[1] J.-M. Lemercier *et al.*, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," *ICASSP*, 2023
[2] E. Hoogeboom *et al.*, "simple diffusion: End-to-end diffusion for high resolution images," *ICML*, 2023
[3] T. Karras *et al.*, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, 2022
[4] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2021

Experimental setup

- Modifications to SGMSE+M[1]:
  - Shifted-cosine noise schedule[2]
  - Neural network preconditioning[3]
  - Heun-based sampler[3] instead of PC sampler[4]

- Evaluation in matched and mismatched conditions:
  - Noisy speech is generated using 5 speech corpora, 5 noise databases and 5 room impulse response databases
  - $N = 1$ or $N = 4$ databases are used for the training condition and the remaining databases are used for the mismatched condition
  - Results are averaged across 5 different combinations

	Training				Testing
Speech:	TIMIT	LibriSpeech	WSJ	Clarity	VCTK
Noise:	TAU	NOISEX	ICRA	DEMAND	ARTE
Room:	Surrey	ASH	BRAS	CATT	AVIL

Training
Testing

[1] J.-M. Lemercier *et al.*, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," *ICASSP*, 2023
[2] E. Hoogeboom *et al.*, "simple diffusion: End-to-end diffusion for high resolution images," *ICML*, 2023
[3] T. Karras *et al.*, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, 2022
[4] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2021

Experimental setup

- Modifications to SGMSE+M[1]:
  - Shifted-cosine noise schedule[2]
  - Neural network preconditioning[3]
  - Heun-based sampler[3] instead of PC sampler[4]

- Evaluation in matched and mismatched conditions:
  - Noisy speech is generated using 5 speech corpora, 5 noise databases and 5 room impulse response databases
  - $N = 1$ or $N = 4$ databases are used for the training condition and the remaining databases are used for the mismatched condition
  - Results are averaged across 5 different combinations

Speech:	TIMIT	LibriSpeech	WSJ	Clarity	VCTK
Noise:	TAU	NOISEX	ICRA	DEMAND	ARTE
Room:	Surrey	ASH	BRAS	CATT	AVIL

Training
Testing

[1] J.-M. Lemercier *et al.*, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," *ICASSP*, 2023
[2] E. Hoogeboom *et al.*, "simple diffusion: End-to-end diffusion for high resolution images," *ICML*, 2023
[3] T. Karras *et al.*, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, 2022
[4] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2021

Experimental setup

- Modifications to SGMSE+M[1]:
  - Shifted-cosine noise schedule[2]
  - Neural network preconditioning[3]
  - Heun-based sampler[3] instead of PC sampler[4]

- Evaluation in matched and mismatched conditions:
  - Noisy speech is generated using 5 speech corpora, 5 noise databases and 5 room impulse response databases
  - $N = 1$ or $N = 4$ databases are used for the training condition and the remaining databases are used for the mismatched condition
  - Results are averaged across 5 different combinations

	Training		Testing	Training	
Speech:	TIMIT	LibriSpeech	WSJ	Clarity	VCTK
Noise:	TAU	NOISEX	ICRA	DEMAND	ARTE
Room:	Surrey	ASH	BRAS	CATT	AVIL

Training
Testing

[1] J.-M. Lemercier *et al.*, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," *ICASSP*, 2023
[2] E. Hoogeboom *et al.*, "simple diffusion: End-to-end diffusion for high resolution images," *ICML*, 2023
[3] T. Karras *et al.*, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, 2022
[4] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2021

Experimental setup

- Modifications to SGMSE+M[1]:
  - Shifted-cosine noise schedule[2]
  - Neural network preconditioning[3]
  - Heun-based sampler[3] instead of PC sampler[4]

- Evaluation in matched and mismatched conditions:
  - Noisy speech is generated using 5 speech corpora, 5 noise databases and 5 room impulse response databases
  - $N = 1$ or $N = 4$ databases are used for the training condition and the remaining databases are used for the mismatched condition
  - Results are averaged across 5 different combinations

Speech:	TIMIT	LibriSpeech	WSJ	Clarity	VCTK
Noise:	TAU	NOISEX	ICRA	DEMAND	ARTE
Room:	Surrey	ASH	BRAS	CATT	AVIL

Training
Testing

[1] J.-M. Lemercier *et al.*, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," *ICASSP*, 2023
[2] E. Hoogeboom *et al.*, "simple diffusion: End-to-end diffusion for high resolution images," *ICML*, 2023
[3] T. Karras *et al.*, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, 2022
[4] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2021

Experimental setup

- Modifications to SGMSE+M[1]:
  - Shifted-cosine noise schedule[2]
  - Neural network preconditioning[3]
  - Heun-based sampler[3] instead of PC sampler[4]

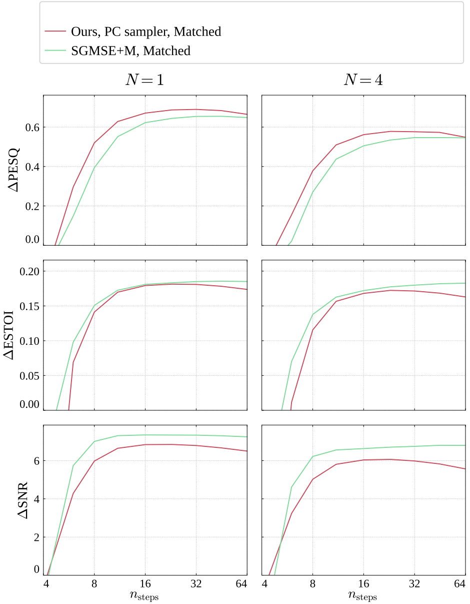
- Evaluation in matched and mismatched conditions:
  - Noisy speech is generated using 5 speech corpora, 5 noise databases and 5 room impulse response databases
  - $N = 1$ or $N = 4$ databases are used for the training condition and the remaining databases are used for the mismatched condition
  - Results are averaged across 5 different combinations

	Testing	Training			
Speech:	TIMIT	LibriSpeech	WSJ	Clarity	VCTK
Noise:	TAU	NOISEX	ICRA	DEMAND	ARTE
Room:	Surrey	ASH	BRAS	CATT	AVIL

Training
Testing

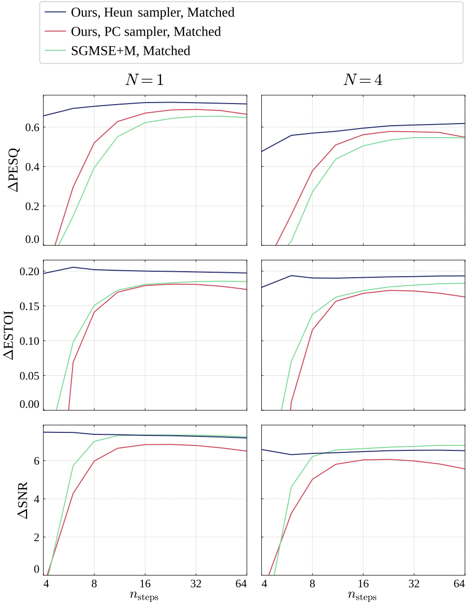
[1] J.-M. Lemercier *et al.*, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," *ICASSP*, 2023
[2] E. Hoogeboom *et al.*, "simple diffusion: End-to-end diffusion for high resolution images," *ICML*, 2023
[3] T. Karras *et al.*, "Elucidating the design space of diffusion-based generative models," *NeurIPS*, 2022
[4] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2021

Results



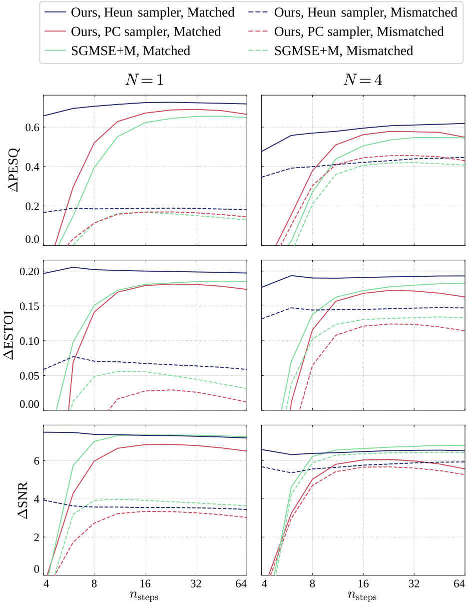
Ours, PC sampler, Matched
SGMSE+M, Matched
$N=1$
$N=4$
ΔPESQ
0.6
0.4
0.2
0.0
ΔESTOI
0.20
0.15
0.10
0.05
0.00
ΔSNR
6
4
2
0
4
8
16
32
64
n_{steps}

Results



Ours, Heun sampler, Matched
Ours, PC sampler, Matched
SGMSE+M, Matched
$N = 1$
$N = 4$
ΔPESQ
0.0
0.2
0.4
0.6
ΔESTOI
0.00
0.05
0.10
0.15
0.20
ΔSNR
0
2
4
6
4
8
16
32
64
n_{steps}

Results



Ours, Heun sampler, Matched
Ours, PC sampler, Matched
SGMSE+M, Matched
Ours, Heun sampler, Mismatched
Ours, PC sampler, Mismatched
SGMSE+M, Mismatched
$N=1$
$N=4$
ΔPESQ
ΔESTOI
ΔSNR
n_{steps}
4 8 16 32 64

Results

		ΔPESQ	ΔESTOI	ΔSNR
Matched	Conv-TasNet	0.63	0.19	**8.58**
	DCCRN	0.41	0.12	7.11
	MANNER	0.68	0.17	7.24
	SGMSE+M	0.65	0.18	7.33
	Ours	**0.72**	**0.20**	7.27
Mismatched	Conv-TasNet	0.12	0.01	3.24
	DCCRN	0.11	0.02	3.25
	MANNER	0.17	0.04	3.40
	SGMSE+M	0.15	0.04	**3.79**
	Ours	**0.19**	**0.06**	3.52

(a) $N = 1$

		ΔPESQ	ΔESTOI	ΔSNR
Matched	Conv-TasNet	0.44	0.16	**7.38**
	DCCRN	0.34	0.12	6.52
	MANNER	0.52	0.15	6.23
	SGMSE+M	0.55	0.18	6.74
	Ours	**0.61**	**0.19**	6.54
Mismatched	Conv-TasNet	0.29	0.09	5.77
	DCCRN	0.23	0.07	5.42
	MANNER	0.38	0.10	5.61
	SGMSE+M	0.42	0.13	**6.42**
	Ours	**0.44**	**0.15**	5.88

(b) $N = 4$

Table: Average ΔPESQ, ΔESTOI and ΔSNR scores in matched and mismatched conditions when training with $N = 1$ (a) or $N = 4$ (b) speech corpora, noise databases and BRIR databases. SGMSE+M and Ours use $n_{\text{steps}} = 32$.

Conclusion

- The drift towards the noisy speech previously proposed for diffusion-based speech enhancement makes it difficult to apply recent advances from image generation literature
- To overcome this, the diffusion process is reformulated using a change of variable
- A different neural network preconditioning and noise schedule only had a small effect on performance
- The Heun-based sampler substantially improved the performance at few sampling steps
- All systems substantially benefited from training with multiple corpora in mismatched conditions

Thank you!