Conditioning and Sampling in Variational Diffusion Models for Speech Super-Resolution

Chin-Yun Yu [1] Sung-Lin Yeh [2] György Fazekas [1] Hao Tang [2]

Queen Mary University of London

[1]Centre for Digital Music, Queen Mary University of London
[2]Institute for Language, Cognition and Computation, University of Edinburgh

THE UNIVERSITY of EDINBURGH

Motivation

Most diffusion-based speech restoration models treat the input audio as condition input to the neural networks. There has been no similar attempt to design a sampling method for recovering speech signals beyond additive noise. We aim to bridge this gap and explore the possibility of enhancing existing diffusion speech super-resolution (SR) models.

Variational Diffusion Models

Variational diffusion models [1] assume an audio sample $\mathbf{x}$ is generated by a chain of T latent variables $\mathbf{z}_t$.

$$p(\mathbf{x}, \mathbf{z}_{1:T}) = p(\mathbf{x}|\mathbf{z}_1)p(\mathbf{z}_T)\prod_{t=2}^{T} p(\mathbf{z}_{t-1}|\mathbf{z}_t)$$

$p(\mathbf{z}_{t-1}|\mathbf{z}_t)$ is usually parameterised as $q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x} = \hat{\mathbf{x}}_\theta(\mathbf{z}_t; t))$ where $\hat{\mathbf{x}}_\theta(\mathbf{z}_t; t)$ is modelled by neural networks and $q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})$ is a known Gaussian distribution.

System Overview

Inspired by the recent success of image inpainting with diffusion models, we cast the task of speech SR also as an inpainting problem **in the frequency-domain**, and propose an algorithm to condition the diffusion reverse step using filters.

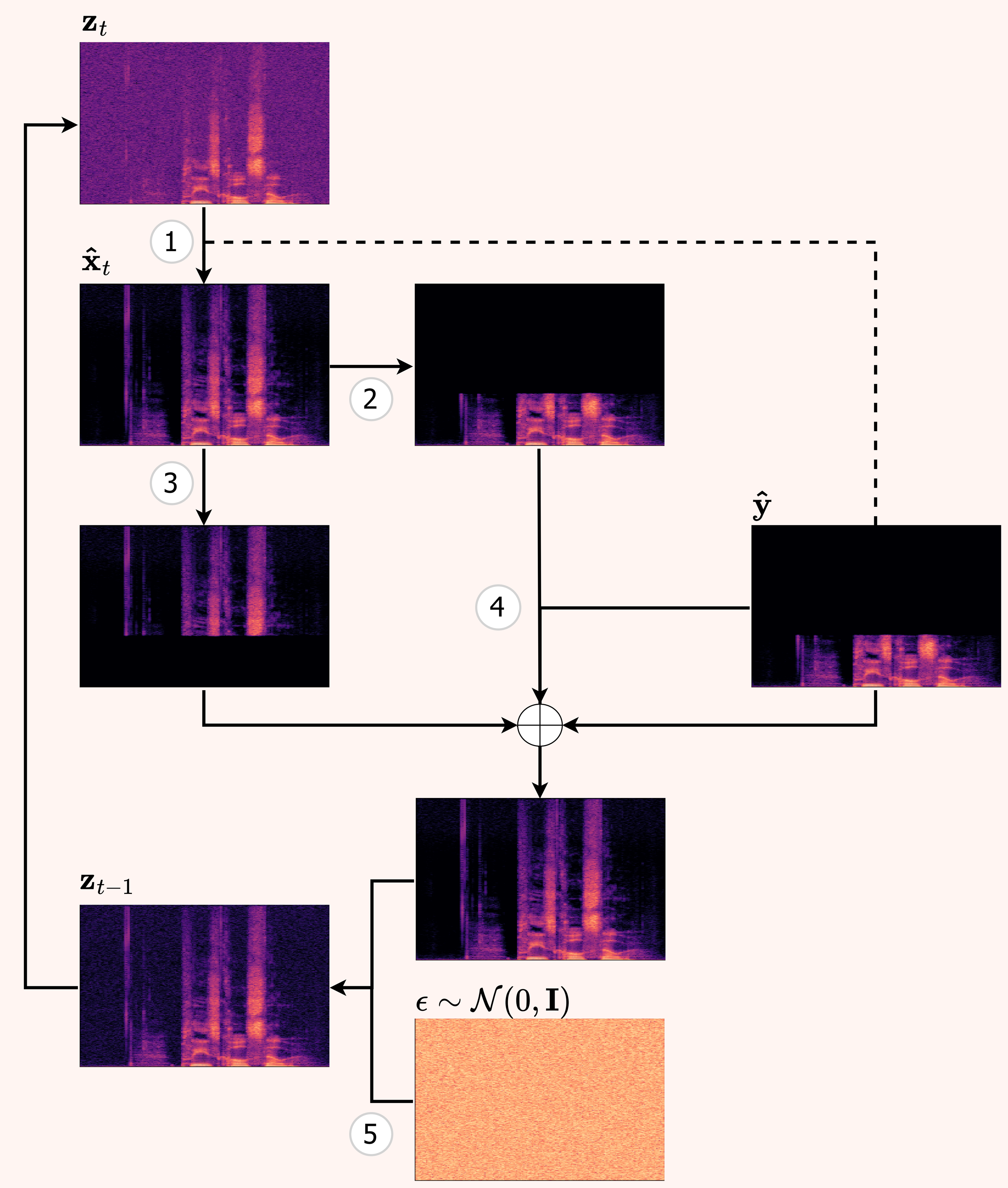

Figure 1. Visual schematic of our proposed conditional diffusion step. $\hat{\mathbf{y}}$ is the low-resolution input. 1 Predict $\hat{\mathbf{x}}_t$ from noisy $\mathbf{z}_t$; 2 Low pass filtering; 3 High pass filtering; 4 MCG [2] correction step; 5 Sample $\mathbf{z}_{t-1}$ using standard diffusion sampling.

The first author is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by UK Research and Innovation [grant number EP/S022694/1] and Queen Mary University of London.

Proposed Approach

To incorporate the available low frequencies in the low-resolution audio $\hat{\mathbf{y}}$, we parameterise the *conditional reverse diffusion* step $p(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{y})$ by replacing low-frequency region of the prediction as

$$q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x} = \hat{\mathbf{y}} + \hat{\mathbf{x}}_\theta(\mathbf{z}_t; t) - \text{lowpass}(\hat{\mathbf{x}}_\theta(\mathbf{z}_t; t))).$$

Manifold Constraint Gradients (MCG)

We subtract the high frequencies of the gradients $\frac{\partial}{\partial \mathbf{z}_t}\|\hat{\mathbf{y}} - \text{lowpass}(\hat{\mathbf{x}}_\theta(\mathbf{z}_t; t))\|_2^2$ before sampling, similar to [2].

Unbiased Diffusion Loss

We trained the model on the original variational lower bound.

$$\begin{aligned}-VLB(\mathbf{x}) = &- \mathbb{E}_{q(\mathbf{z}_1|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z}_1)] + D_{KL}(q(\mathbf{z}_T|\mathbf{x})||p(\mathbf{z}_T)) \\ &+ \frac{\delta_{max} - \delta_{min}}{2}\mathbb{E}_{\boldsymbol{\epsilon}, v}\left[\|\boldsymbol{\epsilon} - \tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_v; v)\|_2^2\right]\end{aligned}$$

Evaluation

We adopted unconditional DiffWave [3] as our denoiser.

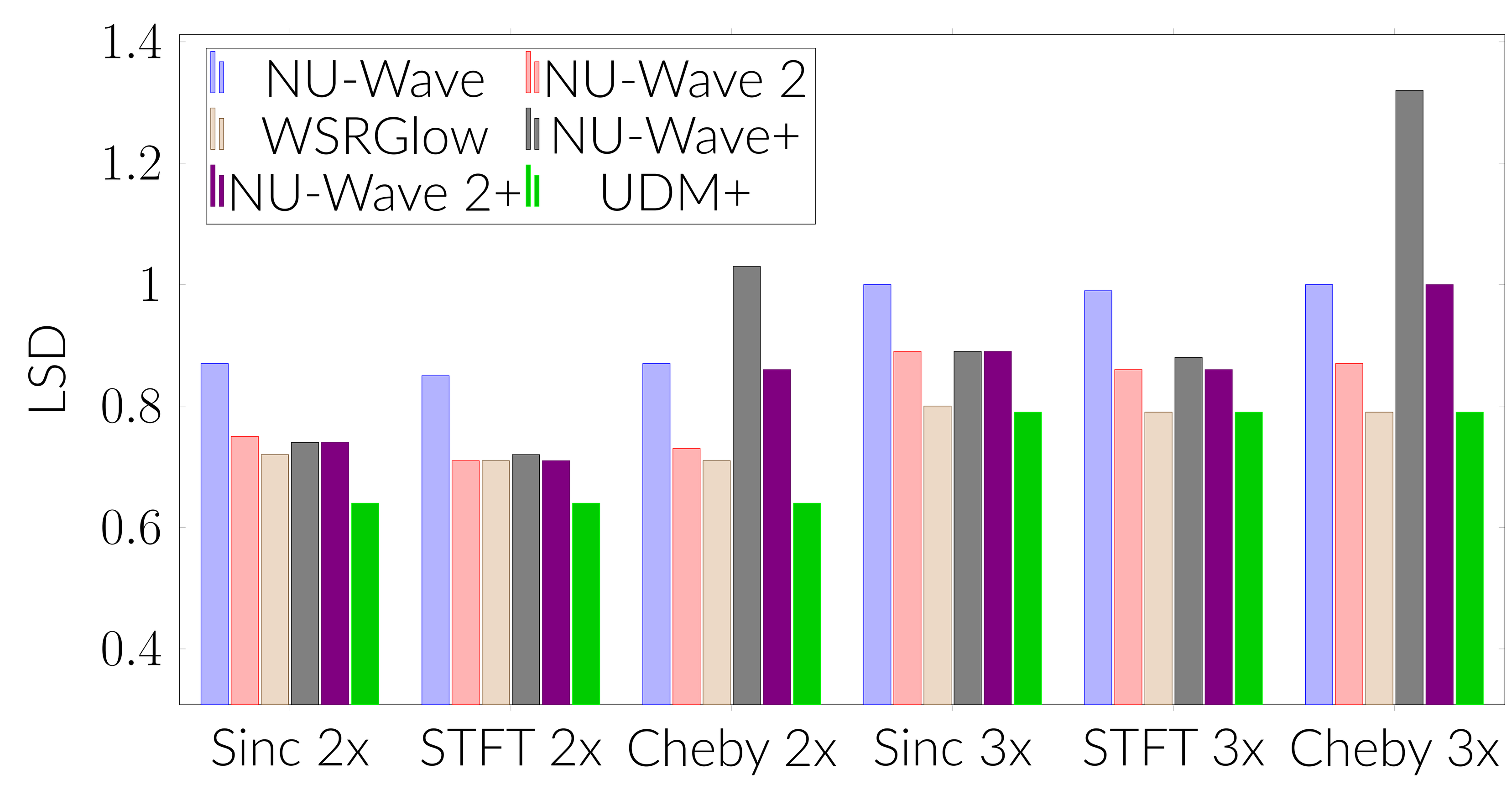

Figure 2. Evaluation results on the 48 kHz VCTK Multi-Speaker benchmark with different upscaling ratios and filter settings.

Conclusion

1. Robust to **various downsampling schemes**.
2. A **drop-in replacement** for the vanilla sampling process and can enhance the performance of the existing works.
3. Can upscale audio up to **48** kHz.

References

[1] D. P. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," in *Advances in Neural Information Processing Systems*, pp. 21696–21707, 2021.

[2] H. Chung, B. Sim, D. Ryu, and J. C. Ye, "Improving diffusion models for inverse problems using manifold constraints," in *Advances in Neural Information Processing Systems*, 2022.

[3] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," in *International Conference on Learning Representations*, 2021.

[4] C.-Y. Yu, S.-L. Yeh, G. Fazekas, and H. Tang, "Conditioning and sampling in variational diffusion models for speech super-resolution," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023.

UK Research and Innovation

CSTR

AI + MUSIC