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]

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

Queen Mary University of London

AI + MUSIC

CSTR

THE UNIVERSITY of EDINBURGH

UK Research and Innovation

Outline

- Introduction
- Background
  - Speech Super-Resolution
  - Diffusion Generative Models
- Related Works
  - NU-Wave series
- Proposed Methodology
- Experiments
- Conclusions

Introduction

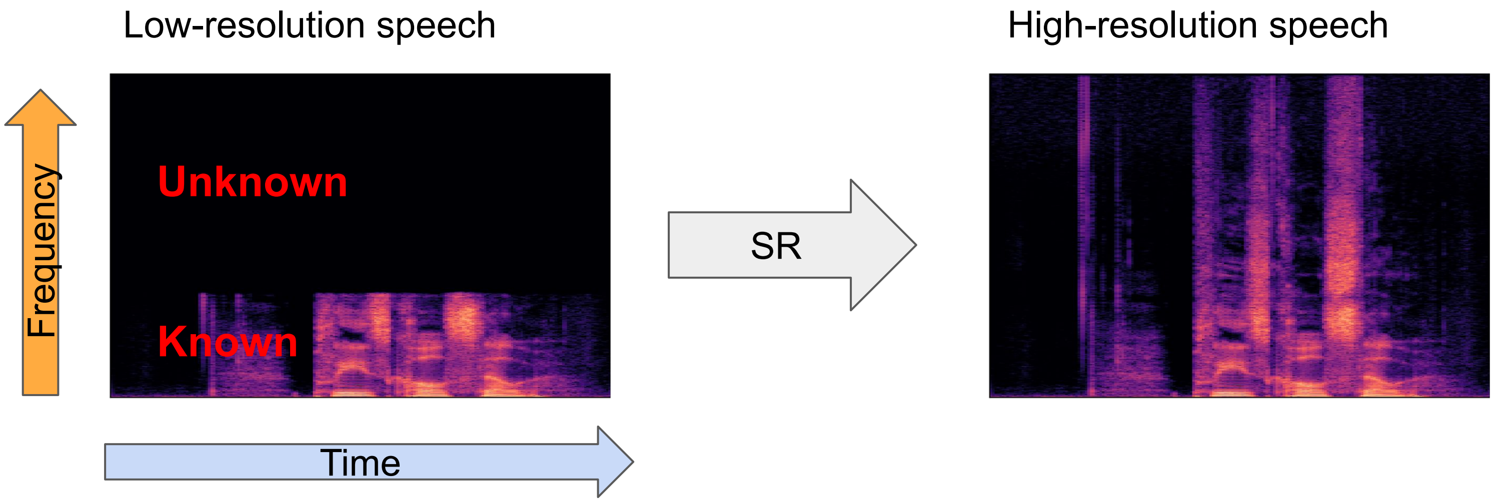
Problems

- Current diffusion-based methods rely on **supervised training** and **cannot adapt** to settings outside its training distribution.
- The condition is only enforced **during denoising.**

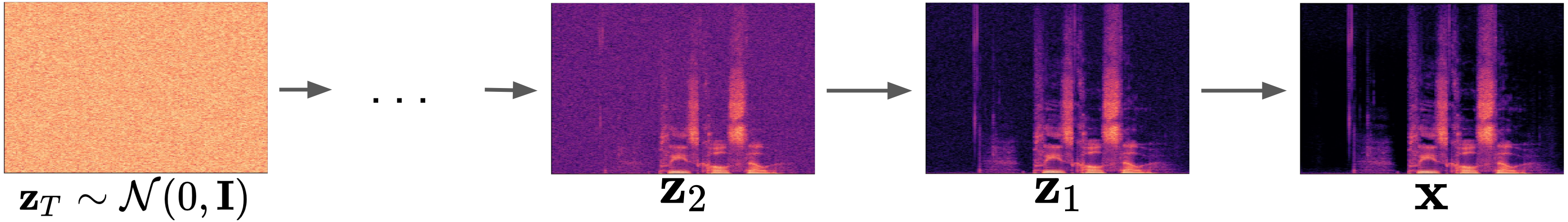
We aim to

- explore possibilities to utilise **pre-trained diffusion models** for **unseen task** (i.e. super-resolution)
- enhance existing methods by conditioning **during diffusion** sampling process

Speech Super-Resolution (Speech SR)

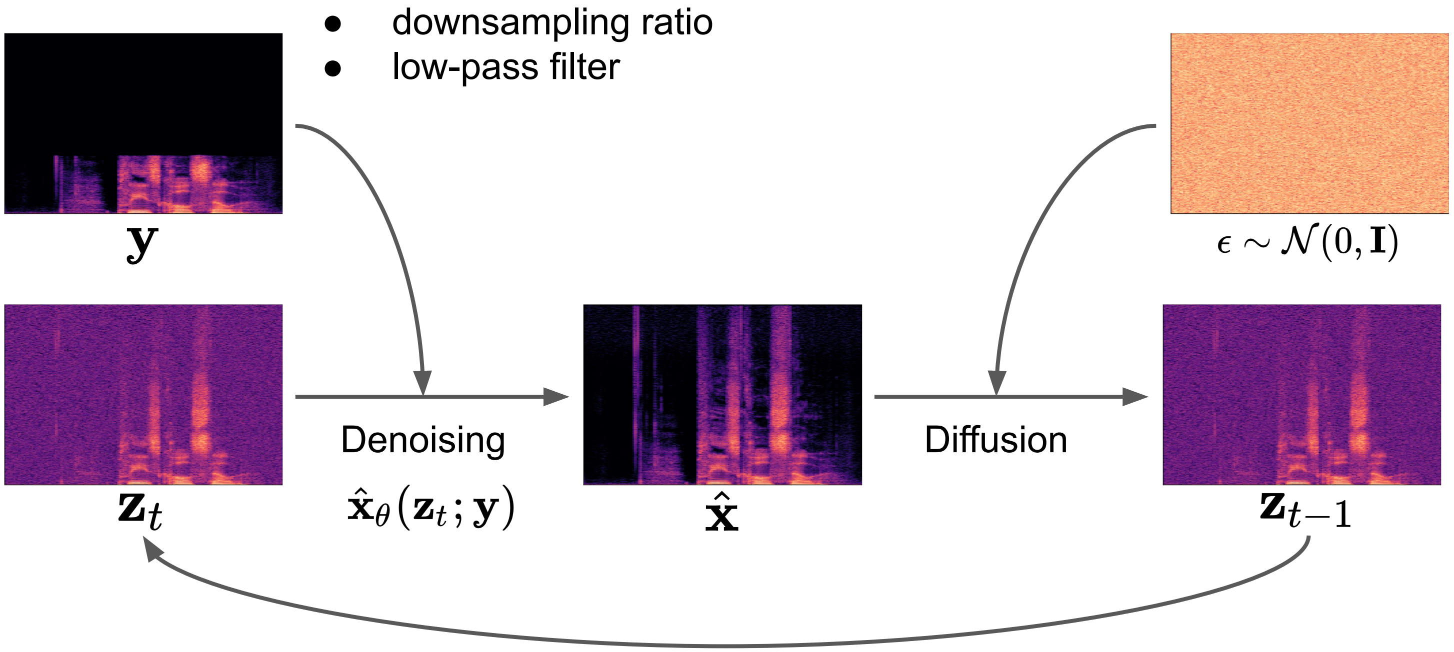


Diffusion Speech Generation



Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in Neural Information Processing Systems 33 (2020): 6840-6851.
Kingma, Diederik, et al. "Variational diffusion models." Advances in neural information processing systems 34 (2021): 21696-21707.

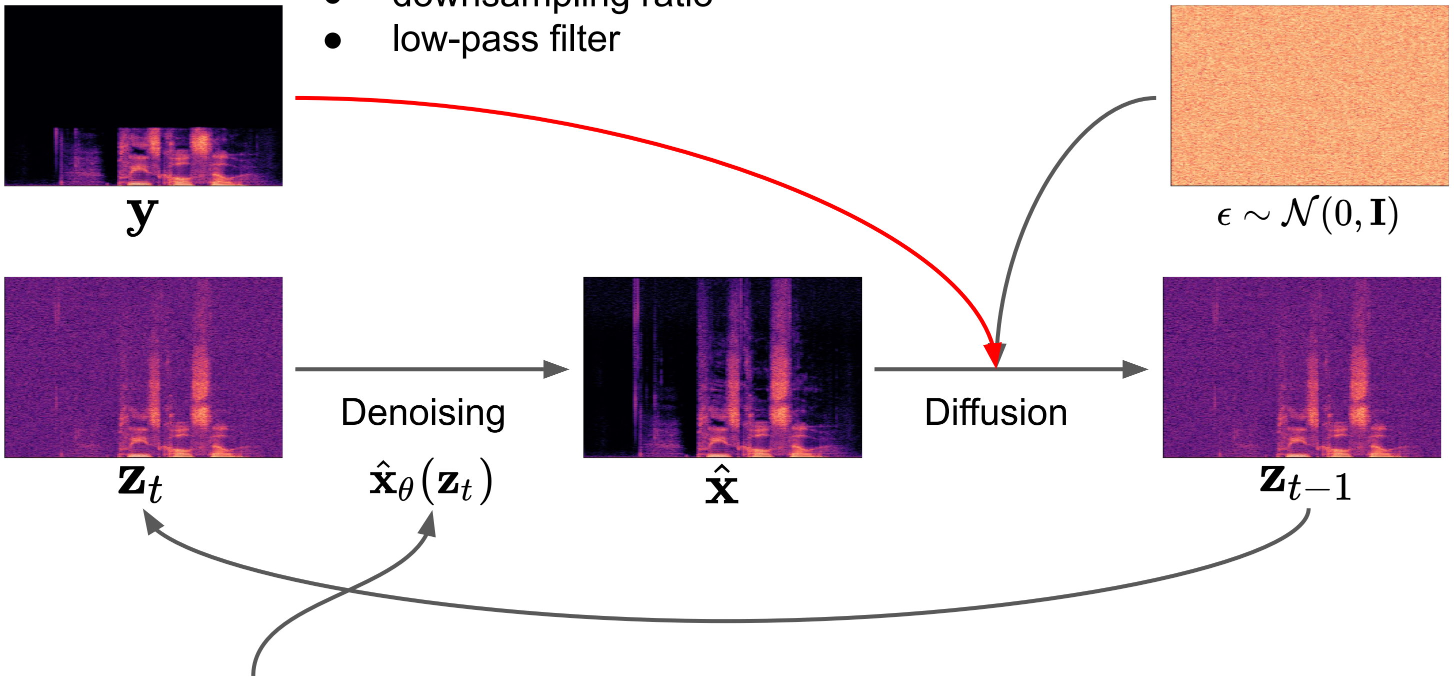
Diffusion-Based SR: NU-Wave/NU-Wave 2



Lee, Junhyeok, and Seungu Han. "NU-Wave: A Diffusion Probabilistic Model for Neural Audio Upsampling}}." *Proc. Interspeech 2021* (2021): 1634-1638.
Han, Seungu, and Junhyeok Lee. "NU-Wave 2: A general neural audio upsampling model for various sampling rates." *arXiv preprint arXiv:2206.08545* (2022).

Proposed: Conditional Diffusion

- downsampling ratio
- low-pass filter



Unconditional pre-trained denoiser

Conditional Diffusion: Downsampling Matrix

Downsampler
(low-pass filtering + dimension reduction)

$$\mathbf{y} = \mathbf{W}\mathbf{x}$$

Upsampler

Low-pass filter

$$\hat{\mathbf{y}} = \mathbf{W}^T\mathbf{y} = \mathbf{W}^T\mathbf{W}\mathbf{x} = \mathcal{F}(\mathbf{x})$$

Same resolution as **x**

Conditional Diffusion in the Frequency Domain

Low-pass filter frequency
response at $\boldsymbol{\omega}$

$$\hat{\mathbf{Y}}^{\omega} = c_{\omega} \mathbf{X}^{\omega}$$

Output from the denoiser

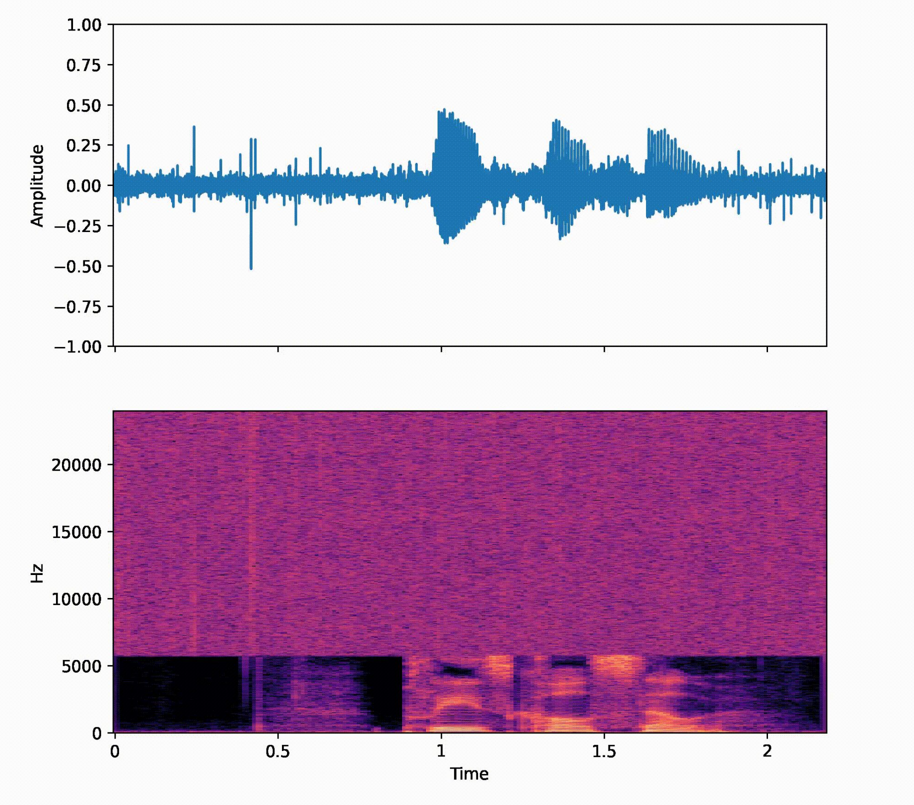
$$p(\mathbf{Z}_{t-1}^{\omega} | \mathbf{Z}_{t}^{\omega}, \hat{\mathbf{Y}}^{\omega}) =$$
$$q(\mathbf{Z}_{t-1}^{\omega} | \mathbf{Z}_{t}^{\omega}, \mathbf{X}^{\omega} = \hat{\mathbf{Y}}^{\omega} + \underline{(1 - c_{\omega})} \hat{\mathbf{X}}_{t}^{\omega})$$

High-pass filter

Conditional Diffusion in the Time Domain

$$p(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{y}) = q(\mathbf{z}_{t-1}|\mathbf{z}_t,$$

$$\mathbf{x} = \mathbf{W}^T\mathbf{y} + \underbrace{(\mathbf{I} - \mathbf{W}^T\mathbf{W})}_{\text{High-pass filter}}\hat{\mathbf{x}}_\theta(\mathbf{z}_t))$$



Manifold Constraint Gradient (MCG)

steps size

$$\eta\underbrace{(\mathbf{I} - \mathbf{W}^T\mathbf{W})}_{\text{High-pass filter}}\frac{\partial}{\partial \mathbf{z}_t}||\underbrace{\mathbf{W}^T\mathbf{y} - \mathbf{W}^T\mathbf{W}\hat{\mathbf{x}}_\theta(\mathbf{z}_t)}_{\text{Differences in low frequencies}}||_2^2$$

Chung, Hyungjin, et al. "Improving diffusion models for inverse problems using manifold constraints." arXiv preprint arXiv:2206.00941 (2022).

Proposed Methods

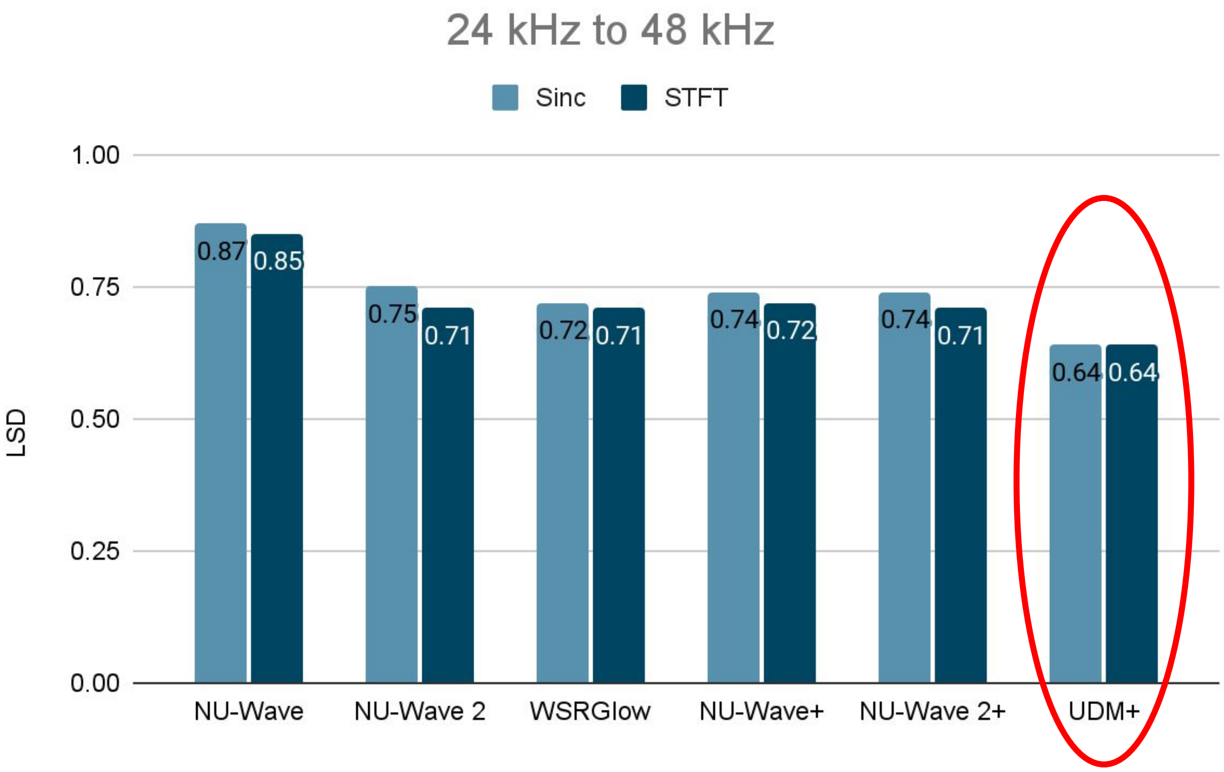
- Pure conditional diffusion
  - Unconditional DiffWave + conditional diffusion (**UDM+**)
- Conditional denoiser + conditional diffusion
  - NU-Wave + conditional diffusion (**NU-Wave+**)
  - NU-Wave 2 + conditional diffusion (**NU-Wave 2+**)

Kong, Zhifeng, et al. "Diffwave: A versatile diffusion model for audio synthesis." arXiv preprint arXiv:2009.09761 (2020).

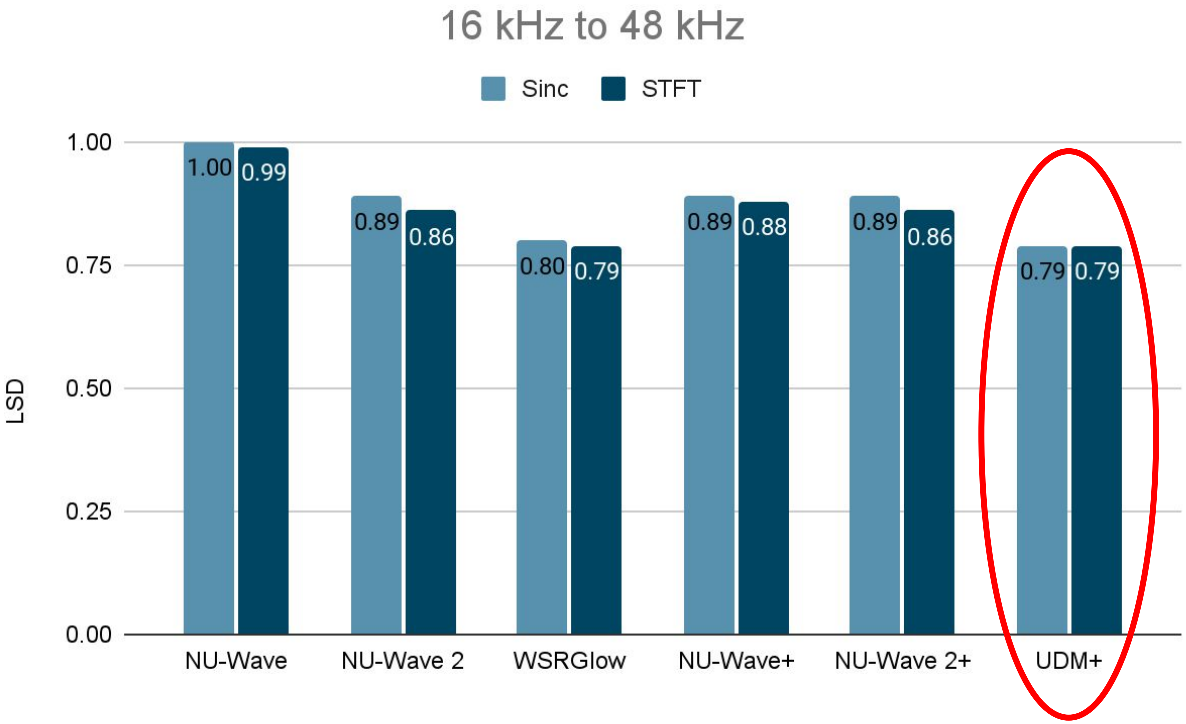
Experimental setup

- Control variables:
  - 2 low pass filters: Sinc, STFT
  - 3 upscaling settings
- VCTK Benchmark
  - 48 kHz & 16 kHz (one UDM for each sampling rate)
- Metrics
  - Log-spectral-distance (LSD)
  - PESQ
- Extra Baselines
  - WaveGlow (48 kHz)
  - NVSR (16 kHz)
- Generation settings
  - 50 steps
  - Linear log-SNR noise schedule

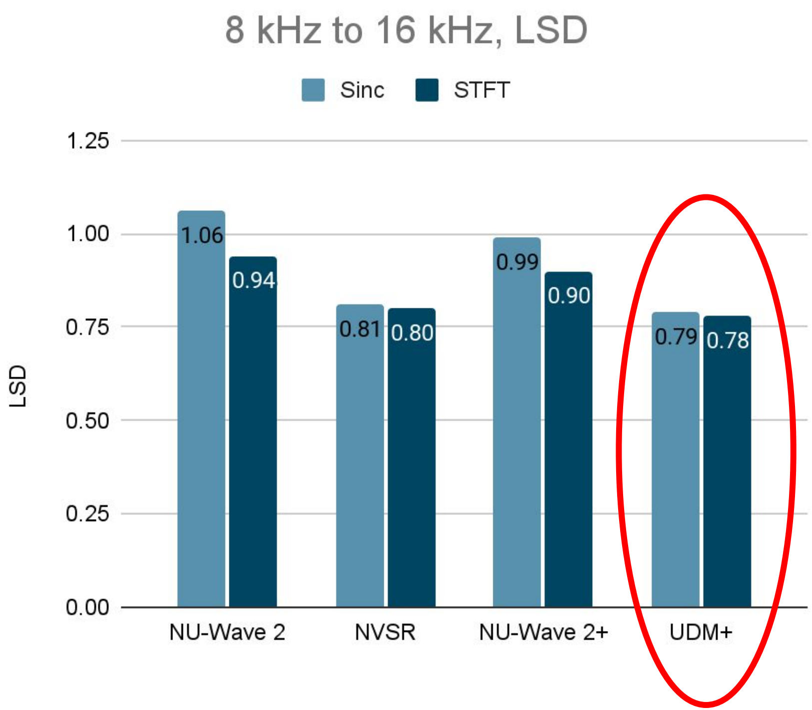
Evaluations on VCTK Benchmark

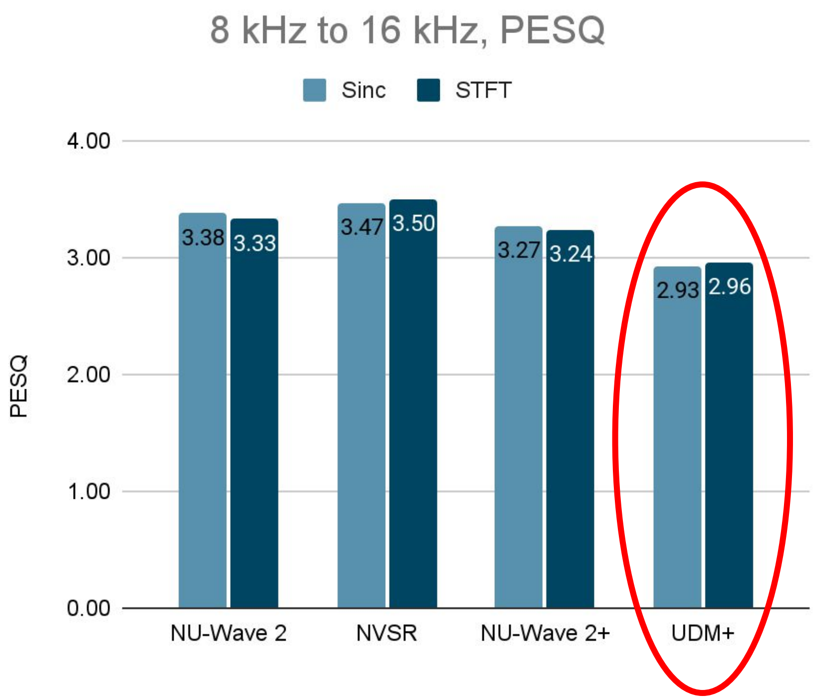


Evaluations on VCTK Benchmark

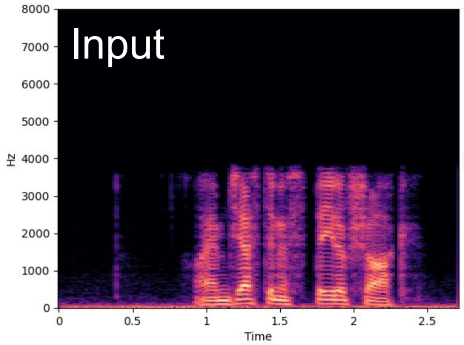


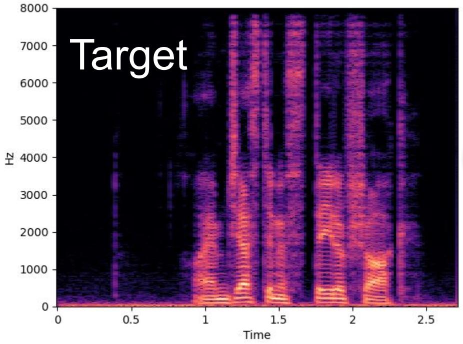
Evaluations on VCTK Benchmark

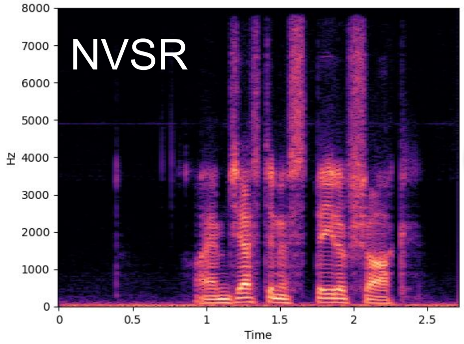


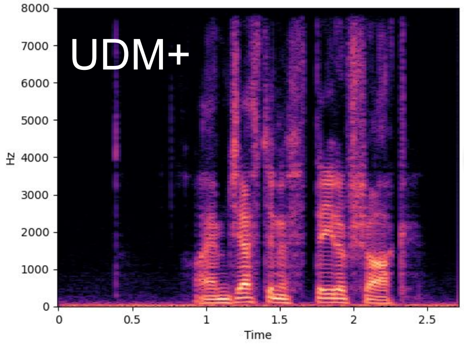


Samples


Input
Hz
8000
7000
6000
5000
4000
3000
2000
1000
0
0
0.5
1
1.5
2
2.5
Time


Target
Hz
8000
7000
6000
5000
4000
3000
2000
1000
0
0
0.5
1
1.5
2
2.5
Time


NVSR
Hz
8000
7000
6000
5000
4000
3000
2000
1000
0
0
0.5
1
1.5
2
2.5
Time


UDM+
Hz
8000
7000
6000
5000
4000
3000
2000
1000
0
0
0.5
1
1.5
2
2.5
Time

Conslusions

- Conditional diffusion improves previous diffusion-based SR significantly.
- UDM+ is robust to various downsampling scheme.

Listening samples, source code, and pre-trained models are available there!

https://yoyololicon.github.io/diffwave-sr/