Diffusion Denoising Probabilistic Model- Speech Enhancement on Wave

Tushar Dhyani and Yung-Ching Yang

Institut für Maschinelle Sprachverarbeitung, Uni Stuttgart

tushar.dhyani, yung-ching.yang{@ims.uni-stuttgart.de}

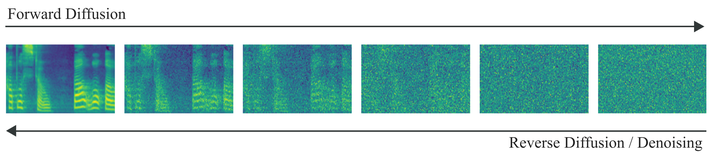
Institut für Maschinelle Sprachverarbeitung

1. Introduction

Project Overview

Diffusion Probabilistic Models [6], inspired by non-equilibrium thermodynamics, are parameterized on Markov chains that slowly add random noise to data and then learns to reverse the diffusion. Hence, they can be used as **generative models**.
This projects aims at using diffusion models for speech enhancement task and improving speech intelligibility.

$$q(x_t|x_{t-1}) = \mathcal{N}(x_T; x_{t-1}\sqrt{1-\beta_t}, I\beta_t))$$



$$q(x_{t-1}|x_t) = \mathcal{N}(x_T; f_\mu(x_t, t), f_\Sigma(x_t, t))$$

Speech enhancement is a crucial component of many user-oriented system such as Text-To-Speech (TTS) systems. In Figure 2, we showcase a basic pipeline for *TTS* system and indicate where our method can be integrated.

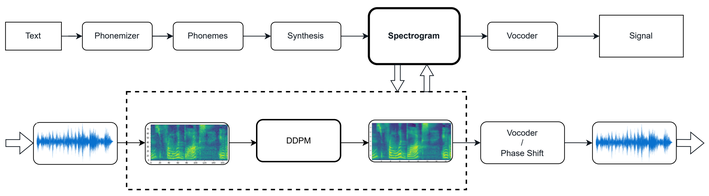


Fig. 2: Pipeline of our method along with an example application/use-case.

Datasets

We use subset of datasets shown in table below. We used only female voices to remove any bias for learning. For evaluation, specifically, we created one variant of dataset containing Gaussian noise at various intensities.

Training Dataset			
Datasets	**amount**	**# speakers**	**noise type**
LJSpeech	∼ 896 minutes	single	Gaussian
LibriSpeech	∼ 680.5 minutes	multiple	Gaussian
Evaluation Dataset			
Datasets	**amount**	**# speakers**	**noise type**
Valentini	30 minutes	multiple	environmental & Gaussian noise

2. Discussion Question

- Does Denoising Diffusion Models benefit from Transfer Learning?
- Does the model trained on Gaussian noise have the ability to reduce the non-Gaussian noise and to what extent?

3. Methods

Baseline Model
In baseline, we implement DDPM with noise schedules as mentioned in [4] with some changes in the design of building blocks of U-Net model. We employed a ConvNeXT block [3] suggested by Liu et al.(2022) in their implementation.
We employed *l1 loss*, *l2 loss* and *huber loss*. In the following experiment section, all experiments are done using l1 loss.

Variance Schedule

We implement linear, cosine, and quadratic schedules, increasing between 0.02 and 10^{-4}.

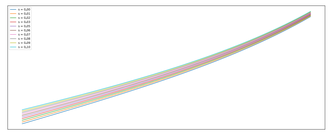


Fig. 3: cosine schedule

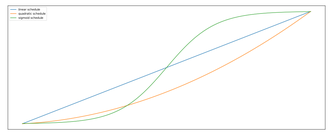


Fig. 4: Other schedules

Experiments
To compare the performance of model with and without pretrained weights and different variance schedules, we ran several experiments using same configurations and took the average of steps at which a clear spectrogram was obtained for reporting.

#	schedule	pretrained?	# Steps to converge
1	cosine	No	-
2	quadratic	No	598
3	quadratic	**Yes**	**171**
4	linear	No	861
5	linear	**Yes**	**186**

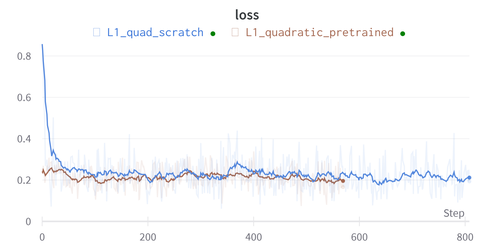


Fig. 5: Losses from pretrained and non-pretrained models compared in the figure.

Here we also report the generated outputs captured at interval of 100 timesteps.

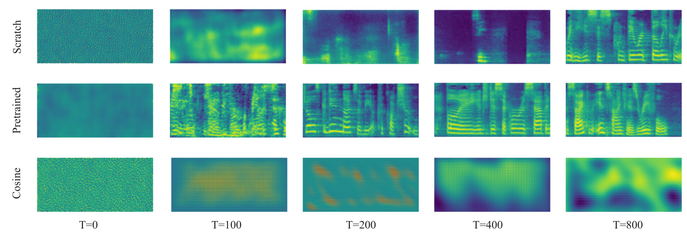


Fig. 6: Output from training with and without the pre-trained weights.

4. Results

We compare the performance of our model with the three baselines, Spectral Gating (noisereduce) [5], Speech Separation (Denoiser) [2], and Speech Enhancement[1]. We compare the results on denoising non-Gaussian and Gaussian noise using Valentini-Botinhao, Cassia. (2017) dataset.

We calculate Signal-to-Noise-Ratio (SNR) and Signal-Invariant-Signal-to-Noise-Ratio (SISNR) for blind estimation and Short-Time Objective Intelligibility (STOI) and Perceptual Evaluation of Speech Quality (PESQ) to calculate Intelligibility and Perception quality of our resulting signals.

Model (Gaussian)	**SNR**	**SISNR**	**STOI**	**PESQ**
NoiseReduce	2.787	6.889	0.782	1.031
Denoiser	4.502	7.731	0.833	2.605
Speech Enhancement	2.945	7.306	0.786	1.361
Our Model	3.251	7.257	0.814	1.901

Model (non-Gaussian)	**SNR**	**SISNR**	**STOI**	**PESQ**
NoiseReduce	2.241	6.616	0.735	1.091
Denoiser	4.345	7.204	0.806	2.682
Speech Enhancement	1.361	6.902	0.762	1.077
Our Model	7.072	7.072	0.799	1.754

5. Future work

1. As we have waveform as input, we can extract phase information to recover waveform after denoising and compare with the neural vocoder's performance.
2. We can condition on real-world noises instead of Gaussian noise to calculate the performance on speech denoising.
3. As spectrograms are analogous to images, could imagenet trained backbone networks improve the overall learning of the model.

Take home message

- DDPM has shown good performance on denoising spectrograms from Gaussian noise.
- Pretrained network from another dataset has shown faster convergence of the model.
- Schedules have shown to impact the overall performance of the training on spectrogram generation.

6. References

[1] Vincent Belz. *Speech-enhancement*. `https://github.com/vbelz/Speech-enhancement`. 2020.

[2] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. "Real Time Speech Enhancement in the Waveform Domain". In: *Interspeech*. 2020.

[3] Zhuang Liu et al. *A ConvNet for the 2020s*. Jan. 2022.

[4] Alex Nichol and Prafulla Dhariwal. *Improved Denoising Diffusion Probabilistic Models*. 2021. DOI: 10.48550/ARXIV.2102.09672. URL: https://arxiv.org/abs/2102.09672.

[5] Tim Sainburg. *timsainb/noisereduce: v1.0*. Version db94fe2. June 2019. DOI: 10.5281/zenodo.3243139. URL: https://doi.org/10.5281/zenodo.3243139.

[6] Jascha Sohl-Dickstein et al. "Deep Unsupervised Learning using Nonequilibrium Thermodynamics". In: *CoRR* abs/1503.03585 (2015). arXiv: 1503.03585. URL: http://arxiv.org/abs/1503.03585.