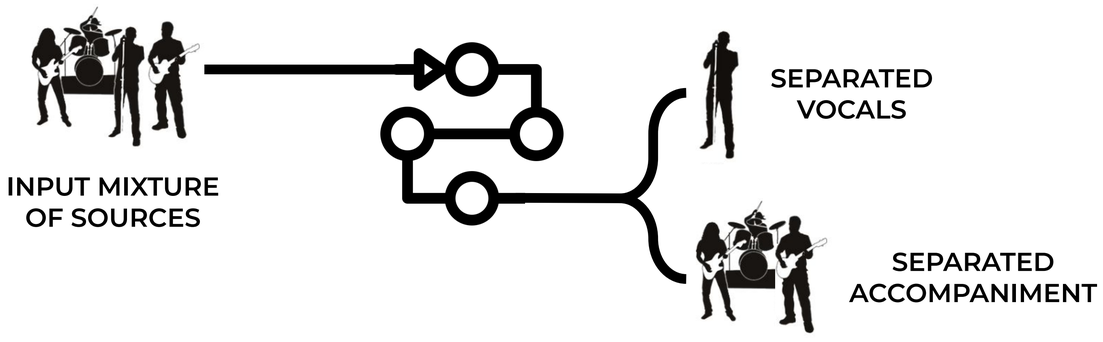
A DIFFUSION-INSPIRED TRAINING STRATEGY FOR SINGING VOICE EXTRACTION IN THE WAVEFORM DOMAIN

Genís Plaja-Roglans, Marius Miron and Xavier Serra
Universitat Pompeu Fabra, Music Technology Group, Barcelona, Spain

ISMIR 2022 BENGALURU

CONTEXT

SINGING VOICE EXTRACTION



DIFFUSION MODELS

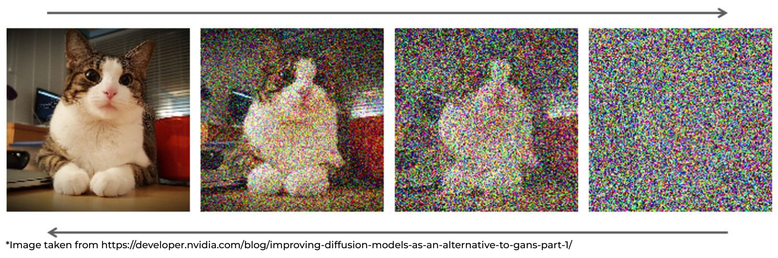
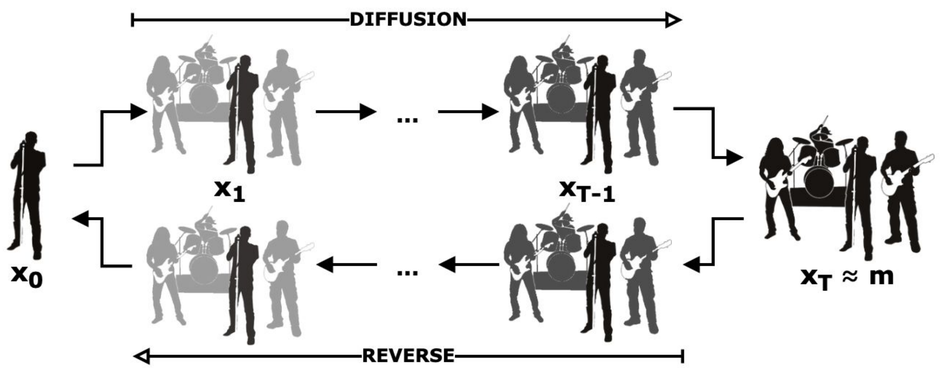


*Image taken from https://developer.nvidia.com/blog/improving-diffusion-models-as-an-alternative-to-gans-part-1/

$$q(x_1, ..., x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$ DIFFUSION

$$p_\theta(x_0, ..., x_{T-1}|m) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$$ REVERSE PROCESS

PROPOSED APPROACH



- We propose to **use a deterministic signal to perturb the input signal** for diffusion.
- By gradually converting a singing voice into its corresponding mixture, we train a small model to conduct the reverse operation.

DIFFUSION PROCESS

- Given the parametrization in [1], we compute each diffusion step as:

$$q(x_t|x_0) = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}m \quad \begin{cases} \alpha_t = 1 - \beta_t \\ \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s \end{cases}$$

NOISE SCHEDULES

- We build our schedules following [1].
  - ➔ Using 20 steps: $\beta_{20} = \{\beta_0 = 1^{-4}, \beta_T = 0.2\}$
  - ➔ Using 8 steps: $\beta_8 = \{\beta_0 = 1^{-4}, \beta_T = 0.5\}$

NETWORK

- We train a non-autoregressive WaveNet [2] to learn the reverse process.
- This network has been previously used for music source separation [3].
- We rely on the DiffWave [4] implementation.

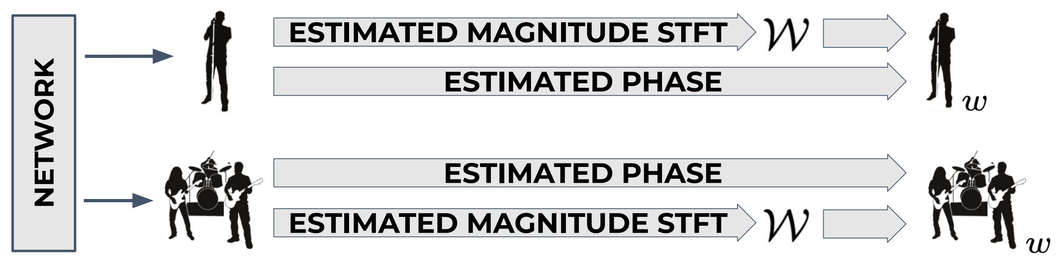
REVERSE PROCESS

- We parametrize the reverse process as:

$$x_{t-1} = \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon(x_t, t)\right)\frac{1}{\sqrt{\alpha_t}}$$

this is an adaptation of the original formula in [1] that yields improved results.

WIENER FILTERING



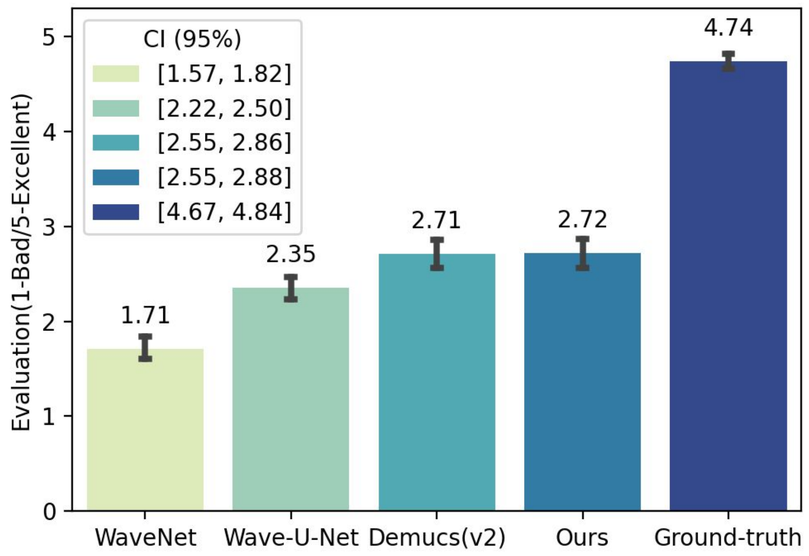
TRAINING OBJECTIVE

$$\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}m, t)\right\|$$

We use the **accompaniment** as target!

EVALUATION

Model	Params	Diff. steps	Singing Voice SDR	Singing Voice SIR	Singing Voice SAR	Accompaniment SDR	Accompaniment SIR	Accompaniment SAR
WaveNet	≈ 3.3M	-	4.49	13.52	6.17	11.39	16.37	13.49
Wave-U-Net	≈ 10.2M	-	4.97	13.98	4.41	11.11	15.30	11.44
ConvTas-Net	≈ 8.75M	-	6.43	-	-	-	-	-
Demucs (v1)	-	-	5.44	-	-	-	-	-
Demucs (v2)	≈ 450M	-	6.84	-	-	-	-	-
Ours (vocal)	≈ 750K	1	4.81	9.21	8.09	-	-	-
Ours (vocal)	≈ 750K	8	5.63	10.55	8.86	-	-	-
Ours (vocal)	≈ 750K	20	5.59	10.78	8.89	-	-	-
Ours (vocal) + Wiener	≈ 750K	20	5.66	11.60	8.49	-	-	-
Ours (accomp)	≈ 750K	100	-	-	-	11.12	13.11	16.44
Ours (vocal & accomp) + Wiener	≈ 750K + 750K	20 + 100	6.07	12.77	8.61	11.72	14.44	16.81



[1] J. Ho, A. Jain, and P. Abbeel, **"Denoising diffusion probabilistic models"**. In Proc. of the 33th Advances in Neural Information Processing Systems (NeurIPS), Online, pp. 6840–6851, 2020.
[2] D. Rethage, J. Pons, and X. Serra, **"A wavenet for speech denoising"**. In Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Paris, France, vol. 2018-April, pp. 5069–5073, 2018.
[3] F. Lluís, J. Pons, and X. Serra, **"End-to-end music source separation: Is it possible in the waveform domain?"**. In Proc. of the Annual Conf. of the Int. Speech Communication Association (INTERSPEECH), Graz, Austria, pp. 4619–4623, 2019
[4] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, **"DiffWave: A Versatile Diffusion Model for Audio Synthesis"**. In Proc. of the 9th Int. Conf. on Learning Representations (ICLR), Vienna, Austria, 2021.

This work was carried out under the projects Musical AI – PID2019-111403GB-I00/AEI/10.13039/501100011033 and NextCore – RTC2019-007248-7 funded by the Spanish Ministerio de Ciencia, Innovación y Universidades (MCIU) and the Agencia Estatal de Investigación (AEI). We would like to thank the 40 participants that took the perceptual test for this work.

upf. Universitat Pompeu Fabra Barcelona — MTG Music Technology Group — MUSICAL AI_ — nextcore — GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES