Zero-Shot Duet Singing Voices Separation with Diffusion Models

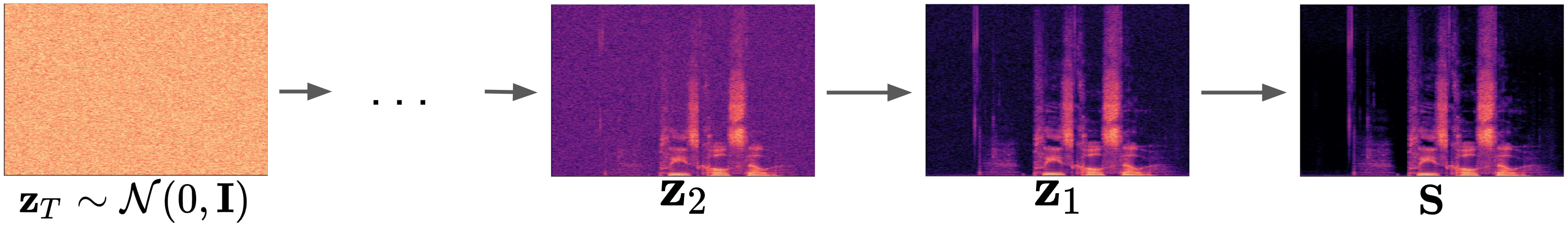
Chin-Yun Yu[1], Emilian Postolache[2], Emanuele Rodolà[2], György Fazekas[1]

Queen Mary University of London[1]
Sapienza University of Rome[2]

SAPIENZA UNIVERSITÀ DI ROMA

Queen Mary University of London

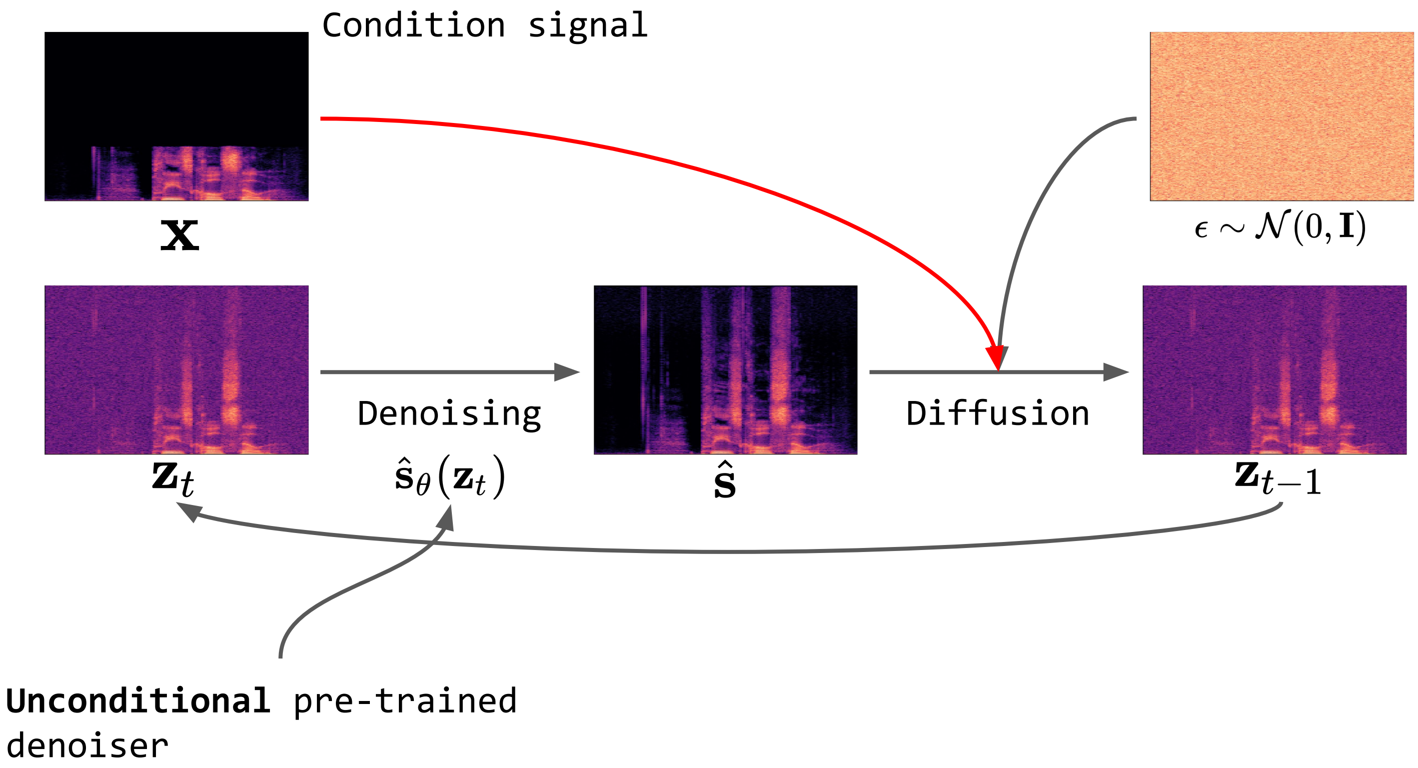
Unconditional Diffusion Generation



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.

Posterior Sampling in Diffusion Models



General Source Separation Problem

Transfer Function

M-channel mixture

$$\mathbf{x}(n,f) = \mathbf{H}(n,f) \begin{bmatrix} s_1(n,f) \\ s_2(n,f) \\ \vdots \\ s_N(n,f) \end{bmatrix} + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, \sigma_x^2)$$

Measurement Noise

N sources

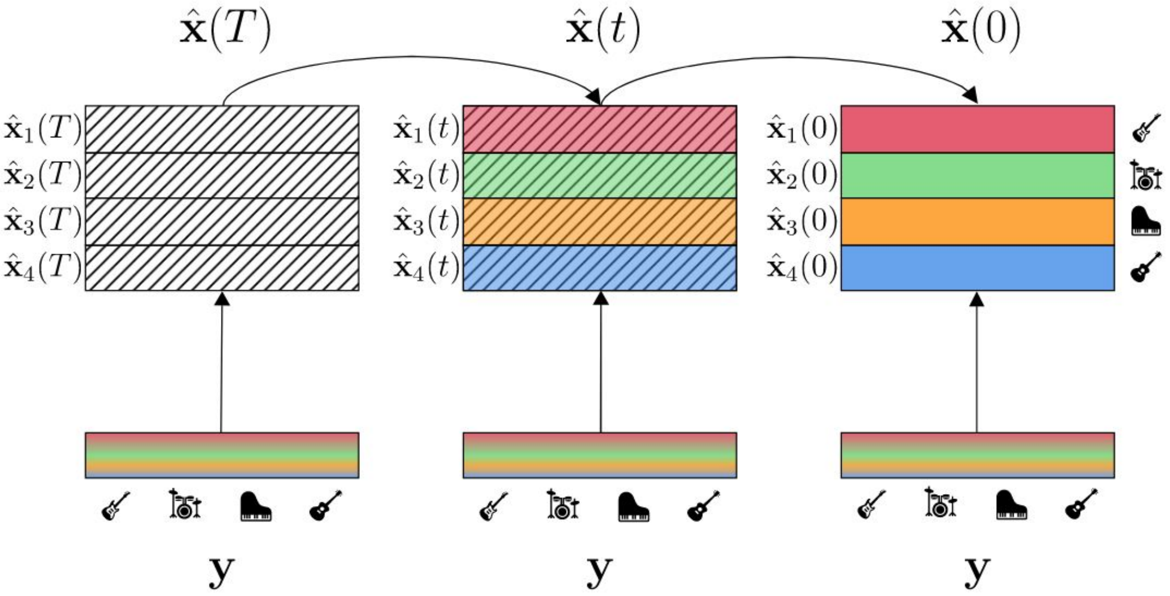
Audio Inverse Problems (N = 1)

- Bandwidth Extension (**H** is known)
  - VRDMG (Hernandez-Olivan et al., 2023)
  - CQTDiff (Moliner et al., ICASSP 2023)
  - UDM (Yu et al., ICASSP 2023)
- Dereverberation (**H** is unknown)
  - GibbsDDRM (Murata et al., 2023)
  - Saito et al., ICASSP 2023

$$p(s_1|\mathbf{x}, \mathbf{H})$$

$$p(s_1, \mathbf{H}|\mathbf{x})$$

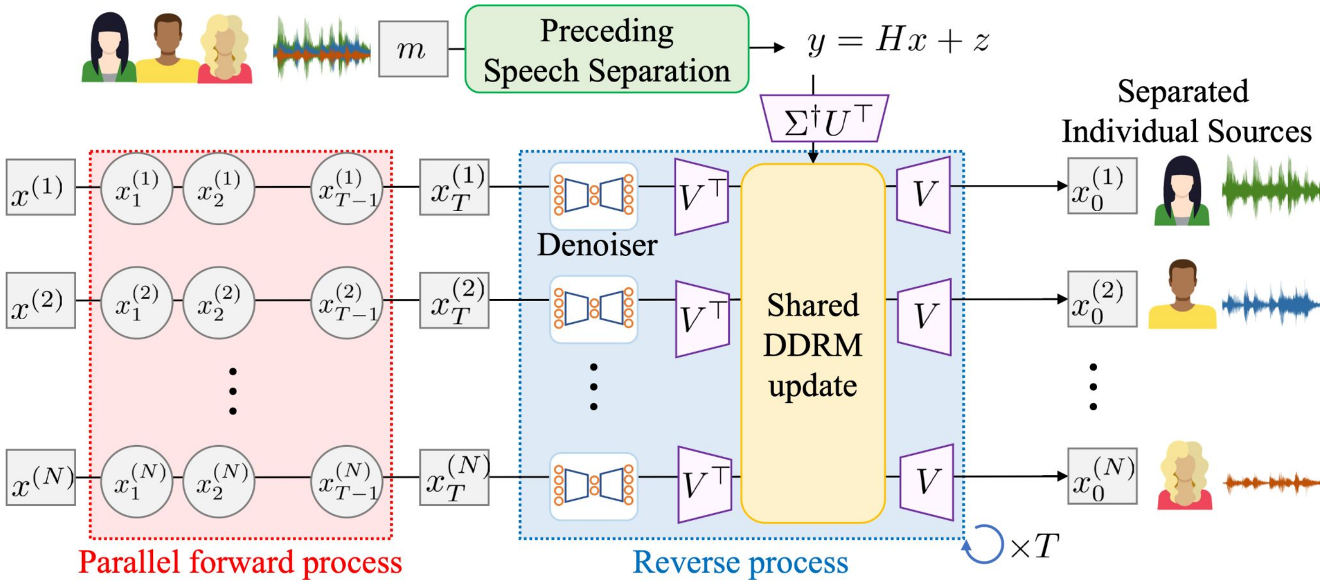
Music Source Separation (N = 4)



Source separation

Mariani, Giorgio, et al. "Multi-source diffusion models for simultaneous music generation and separation." arXiv preprint arXiv:2302.02257 (2023).

Multi Speaker ~~Separation~~ Refinement (N > 1)



Hirano, Masato, et al. "Diffusion-based Signal Refiner for Speech Enhancement." arXiv preprint arXiv:2305.05857 (2023).

Problem with Monotimbral Source Separation

- Definition
  - s_1, s_2,..., s_N have very **similar timbre**
  - All sources are drawn from the **same diffusion model**
- Problem
  - The learned prior is not enough to maintain **temporal coherency** (i.e., singer identity)

Source 1

Predict 1

Mix

Source 2

Predict 2

Proposed Methodology

Dirac Score Posterior Function (Mariani et al.)

Conditional score

$$\nabla_{\mathbf{s}_i(t)} \log p(\mathbf{s}_i(t)|\mathbf{x})$$

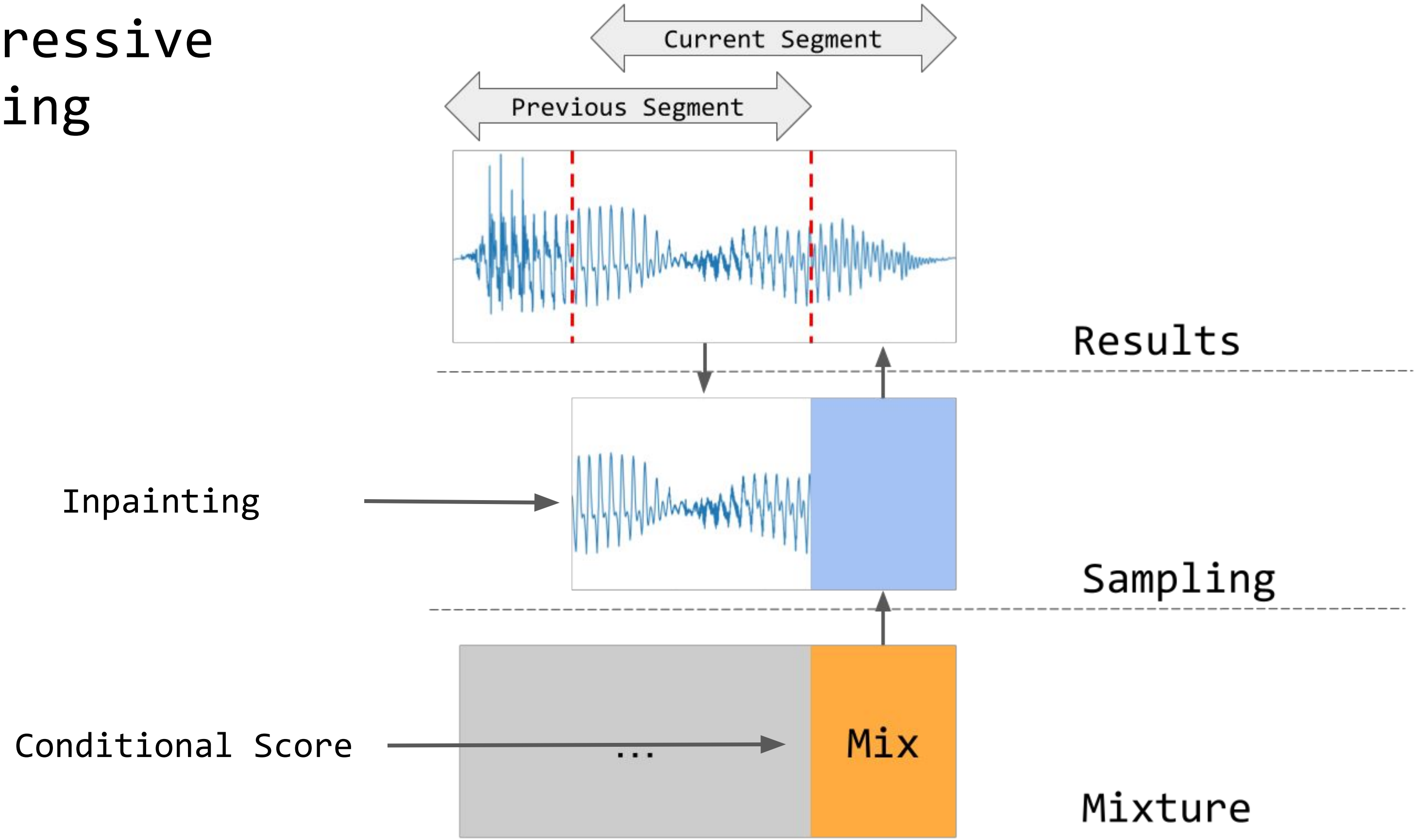
$$\approx \nabla_{\mathbf{s}_i(t)} \log p(\mathbf{s}_i(t)) - \underline{\nabla_{\mathbf{s}_i(t)} \log p(\mathbf{x} - \textstyle\sum_{i=2}^{N} \mathbf{s}_i(t))}$$

Unconditional score

$$\hat{\mathbf{s}}_1(t) = \mathbf{x} - \textstyle\sum_{i=2}^{N} \mathbf{s}_i(t)$$

Mariani, Giorgio, et al. "Multi-source diffusion models for simultaneous music generation and separation." arXiv preprint arXiv:2302.02257 (2023).

Autoregressive Inpainting



Experiment

- Score prediction model: 1D Unet Model from Mousai
  - batch size 32, 1M steps
- Training data: 8 singing datasets combined (>104 hours)
  - 24 kHz
  - 131072 samples per segment (~= 5.46 seconds)
- Test data: MedleyVox duet subset (N = 2)
- Metrics
  - SDRi
  - SI-SDRi

Sampling methods

1. **Naive**: conditional score w/o AR inpainting
2. **AR**: conditional score w/ AR inpainting
3. **Segmented:** non-overlapping chunks + conditional score
4. **AR w/ TF:** ground truth as inpainting context (not from the previous generation)
    a. Similar to teacher forcing

Note: <u>we generated three variations in each step and pick the lowest loss one</u>

Results

Methods	SI-SDRi	SDRi
iSRNet (Jeon et al., 2023)	15.10	14.20
NMF	5.12	5.97
Naive	6.61 ± 0.25	7.60 ± 0.21
Segmented	11.14 ± 0.48	11.77 ± 0.47
AR (proposed)	$\mathbf{11.24} \pm 0.40$	$\mathbf{11.89} \pm 0.34$
AR w/ TF	11.75 ± 0.38	12.34 ± 0.39

Source 1

Predict 1

Mix

Source 2

Predict 2

Source code & model weights:
https://github.com/yoyololicon/duet-svs-diffusion

The Holy Grail

A general sampling method for ***Arbitrary H*** *using* ***diffusion models on individual sources***

Potential problems:

1. On the sources
    a. Temporal coherency with monotimbral sources
2. On the transfer function
    a. Unknown multi-channel **H**
    b. Evaluation datasets?