Diffusion-based Generative Speech Source Separation

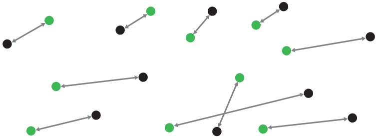
Robin Scheibler (LINE)
Youna Ji, Soo-Whan Chung, Jaeuk Byun,
Soyeon Choe, Min-Seok Choi (NAVER Cloud)

ICASSP 2023

LINE

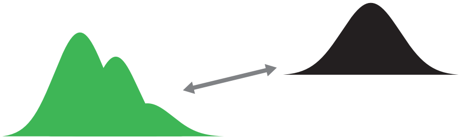
Speech Separation: Discriminative vs Generative

Discriminative



$\min \sum_k \mathcal{L}(\mathbf{s}_k, \hat{\mathbf{s}}_k)$

Generative (proposed)



e.g. GAN, Flow, **Diffusion**. . .

Proposed Method

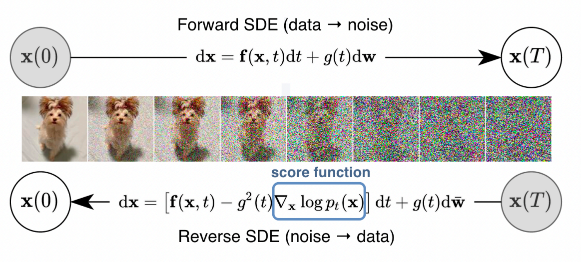
Generative separation of sources **with the same distribution**, i.e., speech

Frameworks for Score-based Generative Modelling

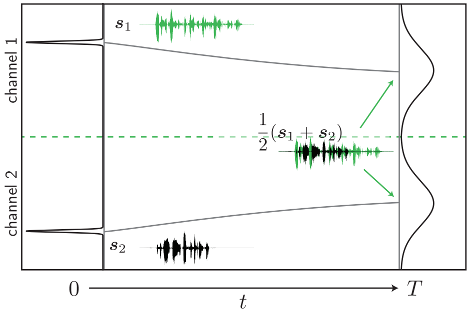
Stochastic Differential Equations [Song2021]

- Continuous-time
- Reverse-time SDE [Anderson1982]
- Model score function

$$\nabla \log p_t(\mathbf{x})$$



New: Diffusion-Mixing Process from Sources → Mixture



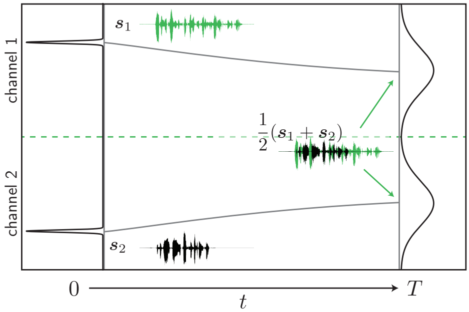
- **2 channels** SDE
- removes difference of sources

$$d\mathbf{x}_t = -\gamma(\mathbf{I} - \mathbf{P})\mathbf{x}_t + g(t)d\mathbf{w}, \quad \mathbf{P} = \frac{1}{2}\mathbb{1}\mathbb{1}^\top, \quad \mathbf{x}_0 = [\mathbf{s}_1 \quad \mathbf{s}_2]^\top$$

Marginal is Gaussian $\mathbf{x}_t \sim \mathcal{N}(\mu_t, \mathbf{\Sigma}_t)$

$$\mu_t = (1 - e^{-\gamma t})\bar{\mathbf{s}} + e^{-\gamma t}\mathbf{x}_0, \qquad \mathbf{\Sigma}_t = \lambda_1(t)\mathbf{P} + \lambda_2(t)(\mathbf{I} - \mathbf{P})$$

New: Diffusion-Mixing Process from Sources → Mixture



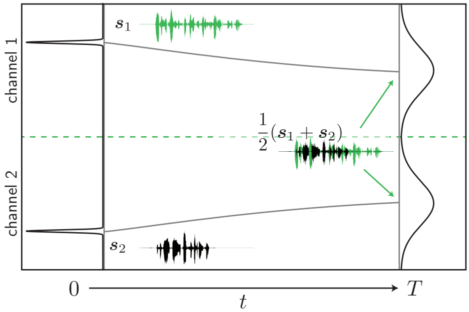
- **2 channels** SDE
- removes difference of sources

$$d\mathbf{x}_t = -\gamma(\mathbf{I} - \mathbf{P})\mathbf{x}_t + g(t)d\mathbf{w}, \quad \mathbf{P} = \frac{1}{2}\mathbb{1}\mathbb{1}^\top, \quad \mathbf{x}_0 = \begin{bmatrix}\mathbf{s}_1 & \mathbf{s}_2\end{bmatrix}^\top$$

Marginal is Gaussian $\mathbf{x}_t \sim \mathcal{N}(\mu_t, \mathbf{\Sigma}_t)$

$$\mu_t = (1 - e^{-\gamma t})\bar{\mathbf{s}} + e^{-\gamma t}\mathbf{x}_0, \qquad \mathbf{\Sigma}_t = \lambda_1(t)\mathbf{P} + \lambda_2(t)(\mathbf{I} - \mathbf{P})$$

New: Diffusion-Mixing Process from Sources $\rightarrow$ Mixture



- **2 channels** SDE
- removes difference of sources

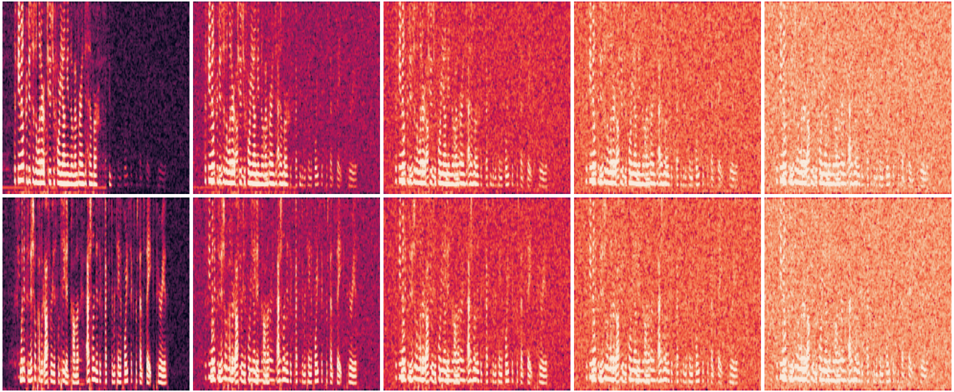
$$d\mathbf{x}_t = -\gamma(\mathbf{I} - \mathbf{P})\mathbf{x}_t + g(t)d\mathbf{w}, \quad \mathbf{P} = \frac{1}{2}\mathbb{1}\mathbb{1}^\top, \quad \mathbf{x}_0 = \begin{bmatrix}\mathbf{s}_1 & \mathbf{s}_2\end{bmatrix}^\top$$

Marginal is Gaussian $\mathbf{x}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$

$$\boldsymbol{\mu}_t = (1 - e^{-\gamma t})\bar{\mathbf{s}} + e^{-\gamma t}\mathbf{x}_0, \qquad \boldsymbol{\Sigma}_t = \lambda_1(t)\mathbf{P} + \lambda_2(t)(\mathbf{I} - \mathbf{P})$$

Illustration of Process

$$\mathbf{x}_0 \longrightarrow \mathrm{d}\mathbf{x} = -\gamma(\mathbf{I} - \mathbf{P})\mathbf{x} + g(t)\mathrm{d}\mathbf{w} \longrightarrow \mathbf{x}_T$$



$$\mathbf{x}_0 \longleftarrow \mathrm{d}\mathbf{x} = \gamma(\mathbf{I} - \mathbf{P})\mathbf{x} - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) + g(t)\mathrm{d}\bar{\mathbf{w}} \longleftarrow \mathbf{x}_T$$

Training Procedure and Objective

Score-based Generative Modelling Idea

Replace score $\nabla \log p_t(\mathbf{x})$ by neural network $\mathbf{q}_\theta(\mathbf{x}, \mathbf{y})$

Training

The marginal distribution is **Normal**, i.e., $p_t(\mathbf{x}) \sim \mathcal{N}(\mathbf{\Pi}\boldsymbol{\mu}_t, \mathbf{\Sigma}_t)$, for permutation of source $\mathbf{\Pi}$, the score has a **closed-form** expression

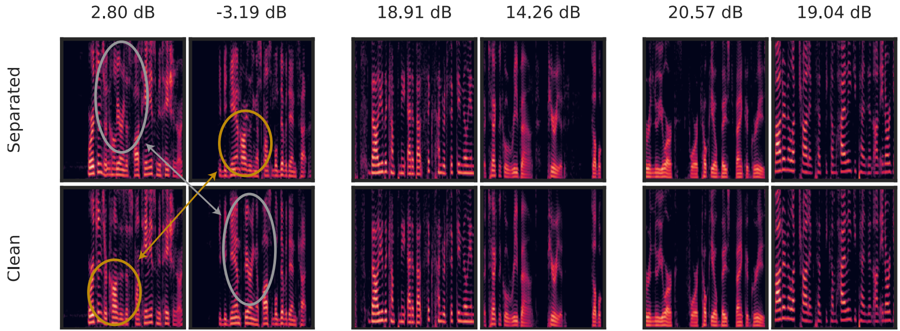
$$\nabla \log p_{t,\mathbf{\Pi}}(\mathbf{x}) = -\mathbf{\Sigma}_t^{-1}(\mathbf{x}_t - \mathbf{\Pi}\boldsymbol{\mu}_t) \tag{1}$$

1. Sample time $t \sim \mathcal{U}[t_\epsilon, t_{max}]$, permutation of sources $\mathbf{\Pi}$
2. Sample $\mathbf{x}_t \sim \mathcal{N}(\mathbf{\Pi}\boldsymbol{\mu}_t, \mathbf{\Sigma}_t)$
3. Gradient step wrt loss

$$\mathcal{L}(\theta) = \min_{\mathbf{\Pi}'} \mathbb{E} \left\| \mathbf{\Sigma}_t^{1/2} \mathbf{q}_\theta(\mathbf{x}_t, t, \mathbf{y}) - \nabla \log p_{t,\mathbf{\Pi}'}(\mathbf{x}_t) \right\|^2$$

Examples

Low Medium High



mix

tgt

enh

Results: Separation

- Dataset: WSJ0_2mix (train/test)
- Model: Noise Conditional Score Network [Song2021]
- OVRL: DNSMOS P.835 non-intrusive metric

Dataset	Model	SI-SDR	PESQ	ESTOI	OVRL
WSJ0_2mix	Conv-TasNet [Luo2019]	16.0	3.29	0.91	3.21
(matched)	DiffSep (proposed)	14.3	3.14	0.90	**3.29**

Results: Enhancement

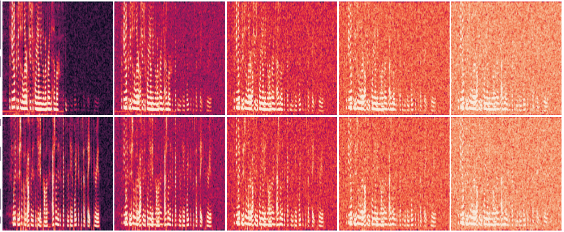
Method is applicable to **enhancement** by letting $\mathbf{s}_2 = \mathbf{n}$.

Dataset: VCTK-DEMAND

Model	SI-SDR	PESQ	ESTOI	OVRL
Discriminative				
Conv-TasNet [Luo2019]	18.3	2.88	0.86	3.20
Generative				
CDiffuse† [Lu2022]	12.6	2.46	0.79	—
SGMSE+† [Richter2022]	17.3	2.93	0.87	—
DiffSep (proposed)	17.5	2.56	0.84	3.09

† results reported in [Richter2022].

Conclusion



- New speech source separation method using diffusion process
- Formulation based on stochastic differential equation

Future Work

- Improve performance
- Speech specific models

Code/Contact

- fakufaku/diffusion-separation
- fakufakurevenge