VRDMG: Vocal Restoration via Diffusion Posterior Sampling with Multiple Guidance

Carlos Hernandez-Olivan, Koichi Saito, Naoki Murata, Chieh-Hsin Lai, Marco A. Martínez-Ramírez, Wei-Hsiang Liao, Yuki Mitsufuji
University of Zaragoza, Sony AI, Sony Group Corporation

Music Restoration using Unsupervised Diffusion Models

Music restoration framed as inverse problem: Recovering a clean signal from a degraded one (inpainting, declipping, etc.).

Motivation: Explore vocal dry and wet restoration with different sampling methods.

Challenges: Ill-posed problems: Multiple possible solutions exist.

Traditional methods: Require task-specific assumptions and paired data (clean & degraded). Not well generalization to unseen data.

Deep generative models for music restoration: Trained on clean signals only.

Data-driven assumptions about clean music.
- Adaptable to various restoration tasks.
- CQT-Diff [3]: A diffusion-based approach. Good performance on piano declipping, bwe, and inpainting.

Areas for improvement:
- Semantic consistency: DC method may lead to nonsensical results.
- Generalizability: Performance on diverse data not explored.

Score-based Generative Modeling with Diffusion Processes

$$\underbrace{\nabla_{\boldsymbol{x}_\tau} \log p_\tau(\boldsymbol{x}_\tau|\boldsymbol{y})}_{\text{Conditional score}} = \underbrace{\nabla_{\boldsymbol{x}} \log p_\tau(\boldsymbol{x}_\tau)}_{\text{density}} + \underbrace{\nabla_{\boldsymbol{x}_\tau} \log p_\tau(\boldsymbol{y}|\boldsymbol{x}_\tau)}_{\text{likelihood}}$$

Inverse Problem via Posterior Sampling

Recovering a clean vocal signal (x_0) from a degraded observation (y) considering a degradation function ($\mathcal{A}$) and measurement noise.

- **Posterior sampling:** It leverages the relationship between the prior distribution of clean signals and the likelihood of observing the degraded signal given the clean one.
- **Conditional score function:** Combines the score function and the likelihood term based on the degradation function.

Tasks

Model: CQT-Diff [3] trained on vocal dry and wet data: NHSS dataset, NUS, MUSDB18 (vocals). 22.05KHz.

Declipping $\mathcal{A}(\boldsymbol{x}_0) = (|\boldsymbol{x}_0 + c| - |\boldsymbol{x}_0 - c|)/2.$

Bandwidth Extension $\mathcal{A}(\boldsymbol{x}) = \mathrm{LPF}(\boldsymbol{x})$

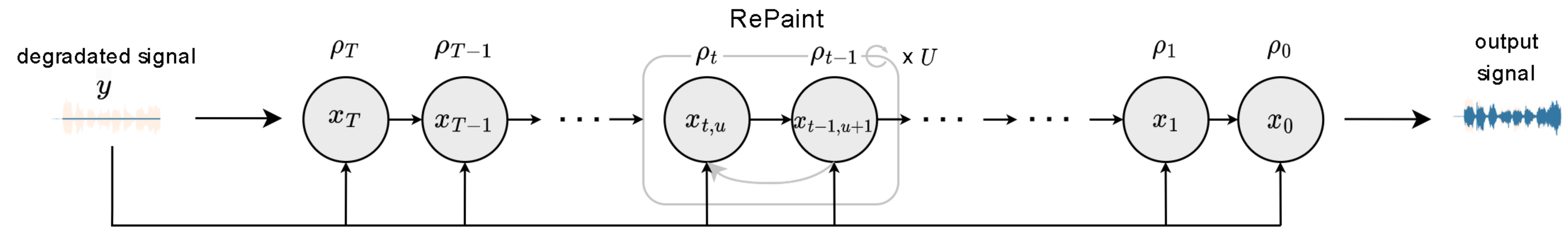
Contributions

Enhanced Diffusion Posterior Sampling (DPS)
- RP strategy with time scheduling for improved semantic consistency.
- Refine Reconstruction Guidance (RG) with time-dependent scaling.
- Integrate Pseudoinverse-Guided Diffusion Models for broader applicability.

Systematic Evaluation
- Conduct experiments to identify the most effective combination of these approaches.
- Achieve comparable performance to the CQT model using the 1D SaShiMi-Diff architecture with a 15x faster inference speed.

$\nabla_{\boldsymbol{x}_\tau}$: Gradient operator
$\mathcal{A}(\cdot)$: Degradation function
$D_{\boldsymbol{\theta}}(\boldsymbol{x}_t; \sigma_t)$: Denoiser

Inverse Problem via Posterior Sampling



Reconstruction Guidance (RG)
- Compute the **likelihood term** in the **conditional score function**.
- Measures the difference between the observed data and the predicted clean signal after applying the degradation function to it.

$$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{y}|\boldsymbol{x}_t) \simeq -\rho(t) \nabla_{\boldsymbol{x}_t} \|\boldsymbol{y} - \mathcal{A}(D_{\boldsymbol{\theta}}(\boldsymbol{x}_t; \sigma_t))\|^2.$$

Pseudo-Inverse Guidance (ΠGDM)
- An alternative method for calculating the **likelihood term**, applicable even for non-differentiable degradation functions.
- It leverages the pseudo-inverse of the degradation function.

$$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{y}|\boldsymbol{x}_t) \simeq \left((h^\dagger(\boldsymbol{y}) - h^\dagger(h(\hat{\boldsymbol{x}}_0)))^\top \frac{\partial \hat{\boldsymbol{x}}_0}{\partial \boldsymbol{x}_t} \right)^\top$$

Improving Data Consistency with RePaint (RP) Strategy
- RP resamples the intermediate prediction during the inference process, promoting semantic consistency. To avoid excessive computation, we propose applying RP cycles during a specific phase inspired by FreeDoM.

$$\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{x}_{t-1}, (\sigma_t^2 - \sigma_{t-1}^2)\mathbf{I}) \quad U = u \cdot \mathbb{1}_{[\phi_1 T/3, \phi_2 T/3]}(t)$$

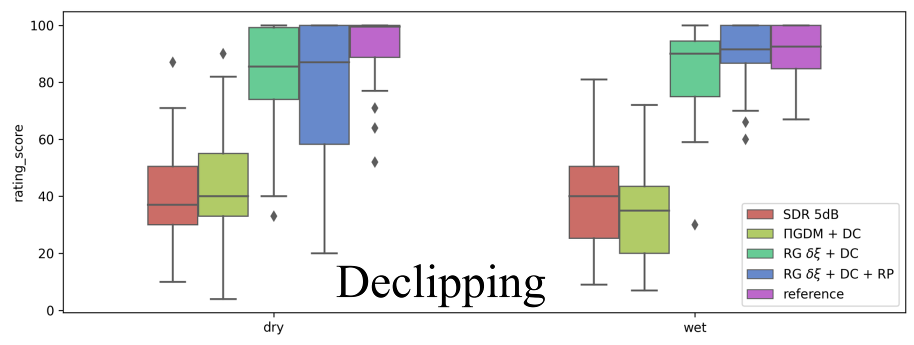
Results

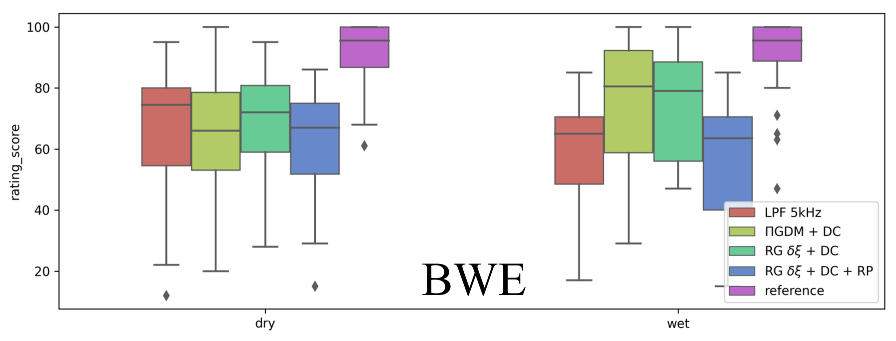
Table 1. Objective metrics for declipping with pre-trained CQT and 1D models.

		SDR = 5dB				SDR = 10dB			
		FAD		SI-SDR		FAD		SI-SDR	
	Method	↓ dry	↓ wet	↑ dry	↑ wet	↓ dry	↓ wet	↑ dry	↑ wet
	Clipped	3.48	2.25	6.42	5.06	1.10	0.87	11.08	9.18
1D	RG [7]	1.84	1.55	9.79	5.36	0.66	1.19	14.39	7.76
	RG + DC	0.94	1.05	10.13	6.15	0.66	0.52	14.48	8.69
	RG $\delta\rho$ + DC	0.88	1.00	10.97	6.03	0.35	0.32	14.79	8.57
	ΠGDM + DC	2.76	1.73	5.44	3.18	1.02	0.76	10.31	7.01
	RG $\delta\rho$ + DC + RP	**0.47**	**0.65**	**11.83**	**7.03**	**0.15**	**0.21**	**15.42**	**9.05**
CQT	RG [7]	1.59	1.05	9.87	5.76	0.37	1.04	14.29	8.20
	RG + DC	1.16	0.80	10.09	6.24	0.37	0.39	14.33	8.91
	RG $\delta\rho$ + DC	**0.52**	**0.34**	10.58	6.51	0.16	**0.19**	14.50	**9.02**
	ΠGDM + DC	2.62	1.57	5.70	3.75	0.74	0.60	10.70	7.22
	RG $\delta\rho$ + DC + RP	0.73	1.02	**11.61**	**7.33**	**0.13**	0.28	**15.03**	8.75

Table 2. Objective metrics for bandwidth extension with pre-trained CQT and 1D models

		f_c = 3KHz				f_c = 5KHz			
		FAD		LSD		FAD		LSD	
Model	Method	↓ dry	↓ wet	↓ dry	↓ wet	↓ dry	↓ wet	↓ dry	↓ wet
	LPF	4.12	3.16	4.26	4.55	2.53	1.79	3.61	3.86
1D	RG [7]	2.38	1.35	**1.87**	1.95	1.72	1.00	1.59	1.86
	RG + DC [7]	1.53	1.44	1.98	**1.84**	0.57	**0.44**	1.67	1.61
	RG $\delta\rho$ + DC	**1.13**	1.38	2.01	1.89	**0.46**	0.48	**1.57**	1.61
	ΠGDM + DC	1.69	**1.26**	2.14	1.86	0.67	0.48	1.93	1.70
	RG $\delta\rho$ + DC + RP	1.15	1.46	2.00	1.88	0.73	0.58	1.56	**1.60**
CQT	RG [7]	1.63	0.75	**1.95**	1.91	1.31	0.82	1.70	1.79
	RG + DC [7]	**1.06**	0.65	2.01	1.82	**0.39**	**0.30**	1.67	**1.59**
	RG $\delta\rho$ + DC	1.65	1.17	2.34	2.10	0.44	0.35	**1.64**	1.61
	ΠGDM + DC	**1.06**	**0.63**	2.00	**1.80**	1.23	0.32	1.68	1.66
	RG $\delta\rho$ + DC + RP	1.78	1.57	2.31	2.03	0.51	0.67	1.68	1.62





Conslusions
- 1D model beats CQT one and it is 15 times faster in inference time.
- Further explore posterior sampling techniques especially for bwe.

References

[1] Song, Y., & Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (Vol. 32)*.

[2] Song J, Vahdat A, Mardani M, Kautz J. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations 2022*.

[3] Moliner, E., Lehtinen, J., & Välimäki, V. (2023). Solving audio inverse problems with a diffusion model. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

[4] Yu, J., Wang, Y., Zhao, C., Ghanem, B., & Zhang, J. (2023). FreeDoM: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023. p. 23174-23184.

[5] Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp 11451-11461)*.

Paper Demos Scholar CV

ICASSP 2024 KOREA