EDMSound: Spectrogram Based Diffusion Models for Efficient and High-Quality Audio Synthesis

Ge Zhu[1], Yutong Wen[1], Marc-André Carbonneau[2], Zhiyao Duan[1]
[1]University of Rochester, [2]Ubisoft La Forge

Introduction

Diffusion models have showcased their capabilities in audio synthesis. Existing models often operate on the cascaded modules to reconstruct waveform. This potentially introduces challenges in generating high-fidelity audio. In addition, diffusion models may unintentionally replicate training data which was examined in computer vision.

In this paper,

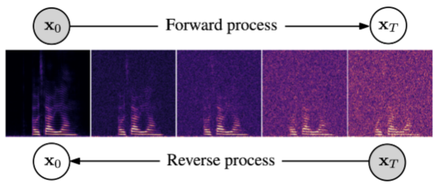
- We propose an end-to-end diffusion-based generative model in complex spectrogram domain under the framework of elucidated diffusion model, named EDMSound.
- We propose a method to examine the content replicaiton issue on a range of audio generation models.

Complex Spectrogram Diffusion

Amplitude transformation on complex spectrogram

$$\tilde{c} = \beta|c|^{\alpha}e^{i\angle c}$$

The forward and backward process of the diffusion in complex sprectrogram domain.

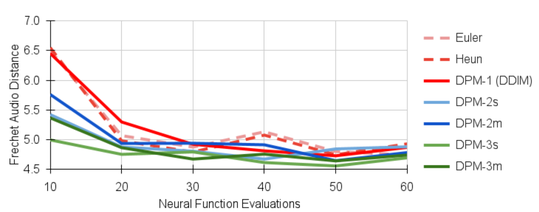


Generalized probability flow ODE under EDM describes the backward diffusion process

$$\mathrm{d}\boldsymbol{x} = \left[\frac{\dot{s}(t)}{s(t)}\,\boldsymbol{x} - s(t)^2\,\dot{\sigma}(t)\,\sigma(t)\,\nabla_{\boldsymbol{x}}\log p\left(\frac{\boldsymbol{x}}{s(t)};\sigma(t)\right)\right]\,\mathrm{d}t$$

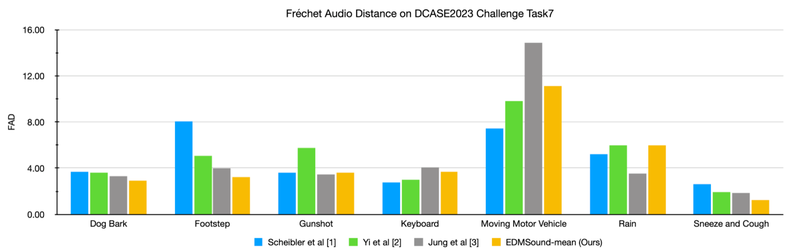
High order DPM-solver

$$x_{t_{i-1}\to t_i} = \frac{\alpha_{t_i}}{\alpha_{t_{i-1}}}x_{t_{i-1}} - \sigma_{t_i}(e^{h_i}-1)\epsilon_\theta(x_{t_{i-1}},t_{i-1}) - \frac{\sigma_{t_i}}{r_2}\left(\frac{e^{h_i}-1}{h}-1\right)D_{2i}$$

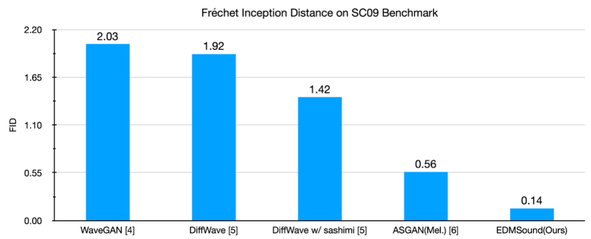


Comparison of FAD scores using different ODE samplers on DCASE2023 Task 7, which focuses on the foley sound generation.

Audio Generation Evaluation



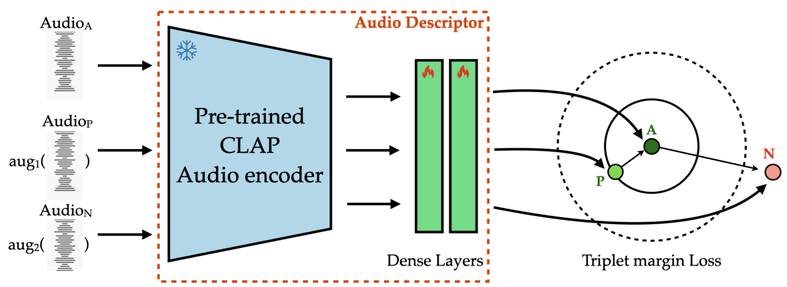
FAD scores of conditional generation results from our system and the top-performed systems from this task.



Dataset SC09 contains speech command zero to nine. The plot shows the FID scores of unconditional generation results from our model and the baseline systems.

Content Replicaton Detection

Definition: Content replicaiton is defined as generated samples that duplicate or closely match with training data.



Training: We freeze the pre-trained CLAP audio encoder and fine-tune the audio descriptor with the triplet margin loss.

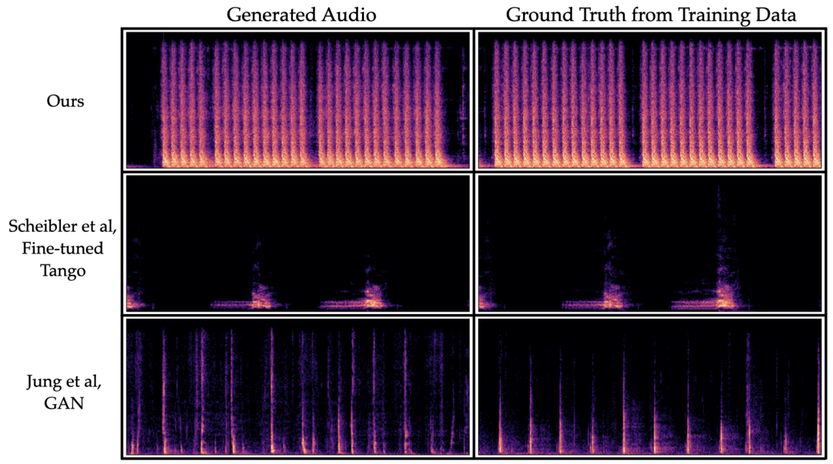
- Anchor
- Positive: augmented anchor
- Negative: augmented sample within the same class of anchor

The data augmentation includes random injection of Gaussian noise, amplitude scaling, and temporal shifting.

Inference: We compute the audio embeddings using the audio descriptor, and the *similarity score* between two audio samples is computed by their cosine similarity. We find the matched audio for a given sample based on its top-1 *similarity score*.

Content Replicaiton Observations

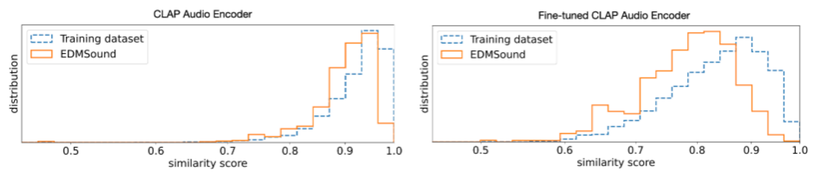
We show examples of content replication in the generated audio samples from our system and the top-performed systems from DCASE2023 task7.



Observation 1:

- We found that our system only generates audio samples very similar to the training data but **not exact copies**.
- System proposed by Scheibler et al[1] generates exact copies of the training data. The bandwidth is limited due to Tango.

Top-1 *similarity score* distributions of the generated audio from our system and the training set using the audio descriptor before and after fine-tune.



Observation 2:

- The fune-tuning process makes the audio embeddings more discriminative.

Reference

[1] R. Scheibler, T. Hasumi, Y. Fujita, T. Komatsu, R. Yamamoto, and K. Tachibana. Class-conditioned latent diffusion model for dcase 2023 foley sound synthesis challenge. Technical report, Tech. Rep., June 2023.
[2] Y. Yuan, H. Liu, X. Liu, X. Kang, M. D. Plumbley, and W. Wang. Latent diffusion model based foley sound generation system for dcase challenge 2023 task 7. arXiv preprint arXiv:2305.15905, 2023.
[3] H. C. Chung, Y. Lee, and J. H. Jung. Foley sound synthesis based on gan using contrastive learning without label information. Technical report, Tech. Rep., June, 2023.
[4] C. Donahue, J. McAuley, and M. Puckette. Adversarial audio synthesis. arXiv preprint arXiv:1802.04208, 2018.
[5] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761, 2020.
[6] M. Baas and H. Kamper. Gan you hear me? reclaiming unconditional speech synthesis from diffusion models. IEEE Spoken Language Technology Workshop (SLT), 2022.