Text-Guided Zero-Shot Audio Style Transfer

Yiran Wang

Tsinghua Shenzhen International Graduate School

Abstract

Text-to-audio (TTA) systems have gained attention for their ability to synthesize general audio based on text descriptions. Current TTA method can not only generate audio with precise content, but also a timbre with specialization and we would like to transfer the powerful models into style transfer tasks. We reproduced AudioLDM, a TTA system that is built on a latent space to learn continuous audio representations from contrastive language-audio pretraining (CLAP) embeddings, and utilized its pretrained diffusion model to achieve text-guided audio style transfer in a zero-shot fashion without finetuning the model on a specific task.

Introduction

We chose AudioLDM as the basement of this work.AudioLDM learns to generate the representation in a latent space encoded by a mel-spectrogram-based VAE. A latent diffusion model (LDM) conditioned on a CLAP embedding is developed for VAE latent generation. We then realize that AudioLDM also enables zero-shot text-guided audio manipulations in the sampling process. By corrupting the timbre information during a forward diffusion process and injecting the text information during the reverse diffusion process using the LDM pretrained.

Method

During training, latent diffusion models (LDMs) are conditioned on an audio embedding E^x and trained in a continuous latent space z_0 learned by VAE. The sampling process uses text embedding E^y as the condition.

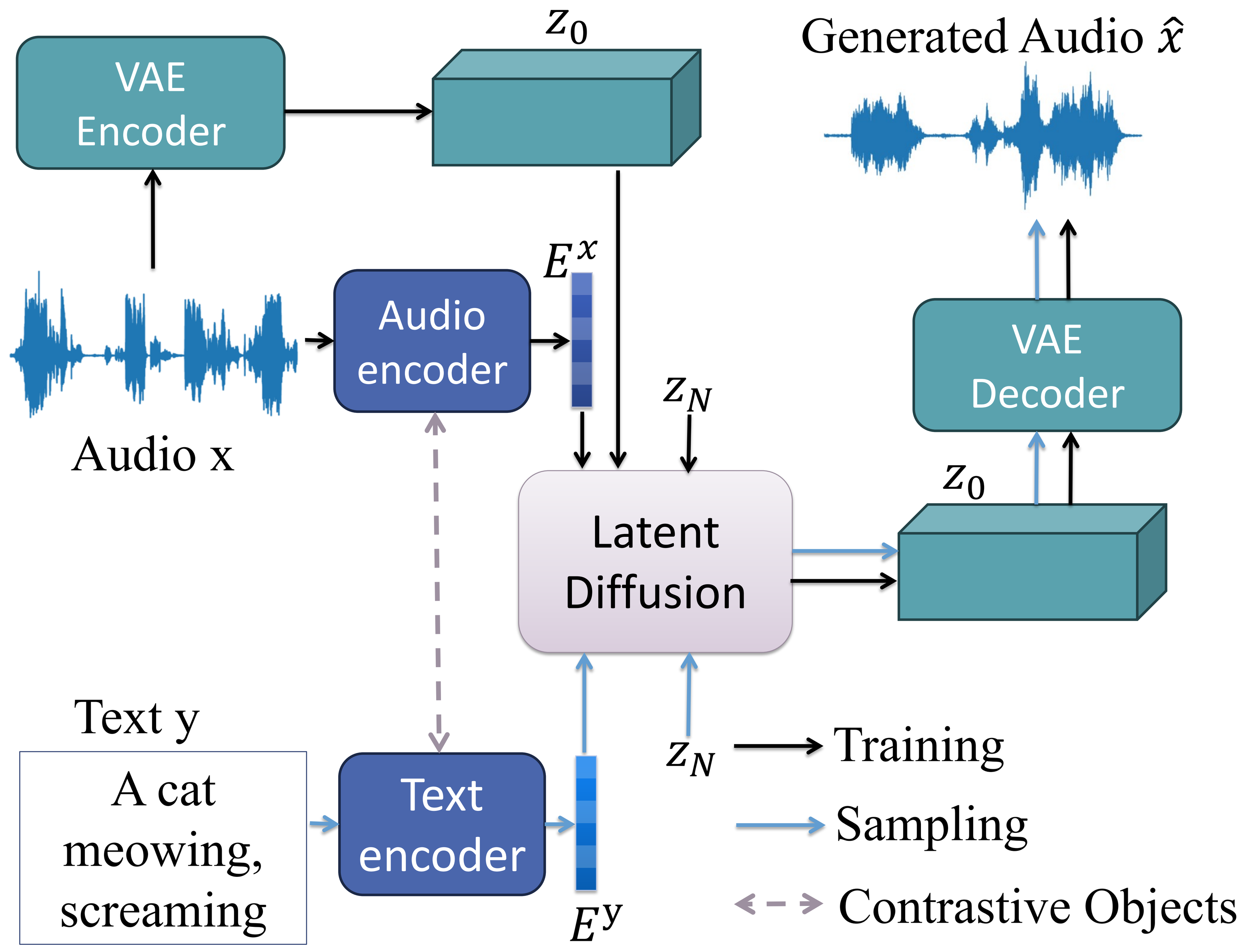
$$L_n(\theta) = \mathbb{E}_{z_0,\epsilon,n} \| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_n, n, \boldsymbol{E}^x) \|_2^2$$

$$p_\theta(\boldsymbol{z}_{0:N} \mid \boldsymbol{E}^y) = p(\boldsymbol{z}_N) \prod_{t=n}^{N} p_\theta(\boldsymbol{z}_{n-1} \mid \boldsymbol{z}_n, \boldsymbol{E}^y)$$

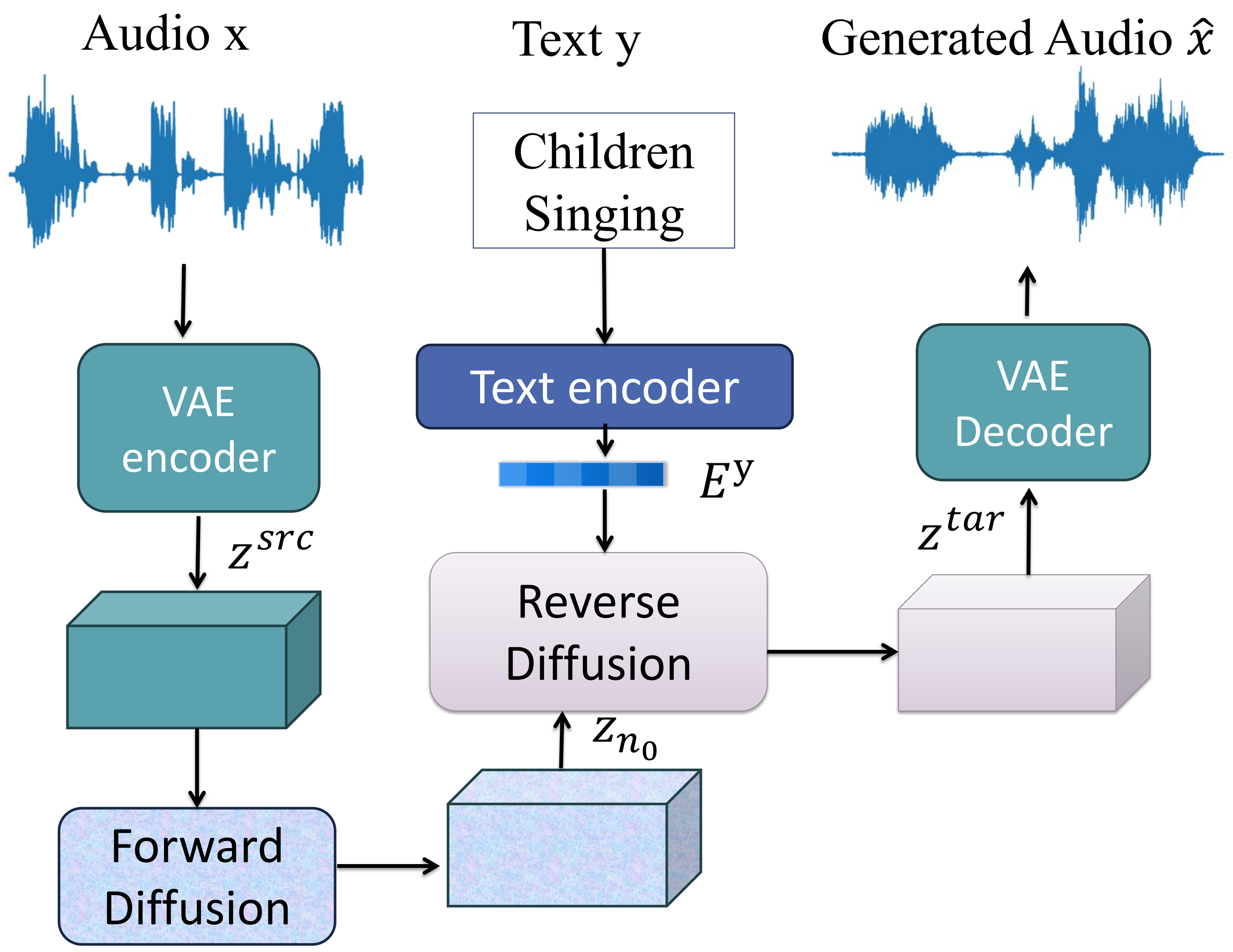
$$p_\theta(\boldsymbol{z}_{n-1} \mid \boldsymbol{z}_n, \boldsymbol{E}^y) = N(\boldsymbol{z}_{n-1}; \boldsymbol{\mu}_\theta(\boldsymbol{z}_n, n, \boldsymbol{E}^y), \boldsymbol{\sigma}_n^2 \boldsymbol{I})$$

$$\boldsymbol{\sigma}_n^2 = \frac{1 - \bar{\alpha}_{n-1}}{1 - \bar{\alpha}_n} \beta_n$$

$$\boldsymbol{\mu}_\theta(\boldsymbol{z}_n, n, \boldsymbol{E}^y) = \frac{1}{\sqrt{\alpha_n}} \left(\boldsymbol{z}_n - \frac{\beta_n}{\sqrt{1 - \bar{\alpha}_n}} \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_n, n, \boldsymbol{E}^y) \right)$$



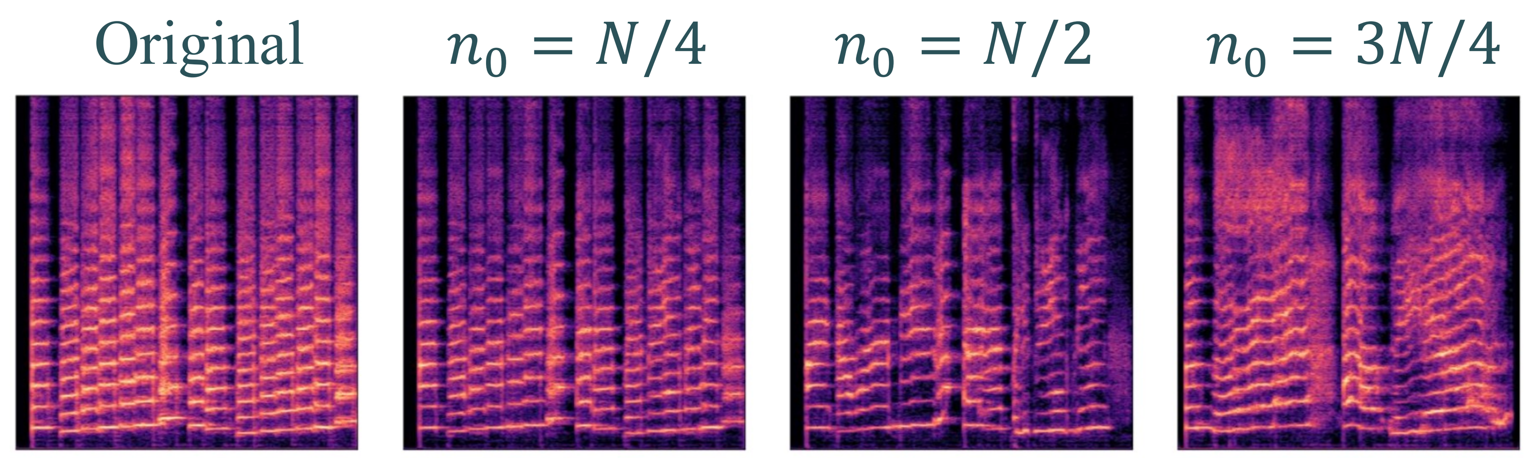
Given a pretrained LDM, zero-shot style transfer can be realized in the reverse diffusion process of LDM. The block "Forward Diffusion" denotes the process that corrupt data with gaussian noise.



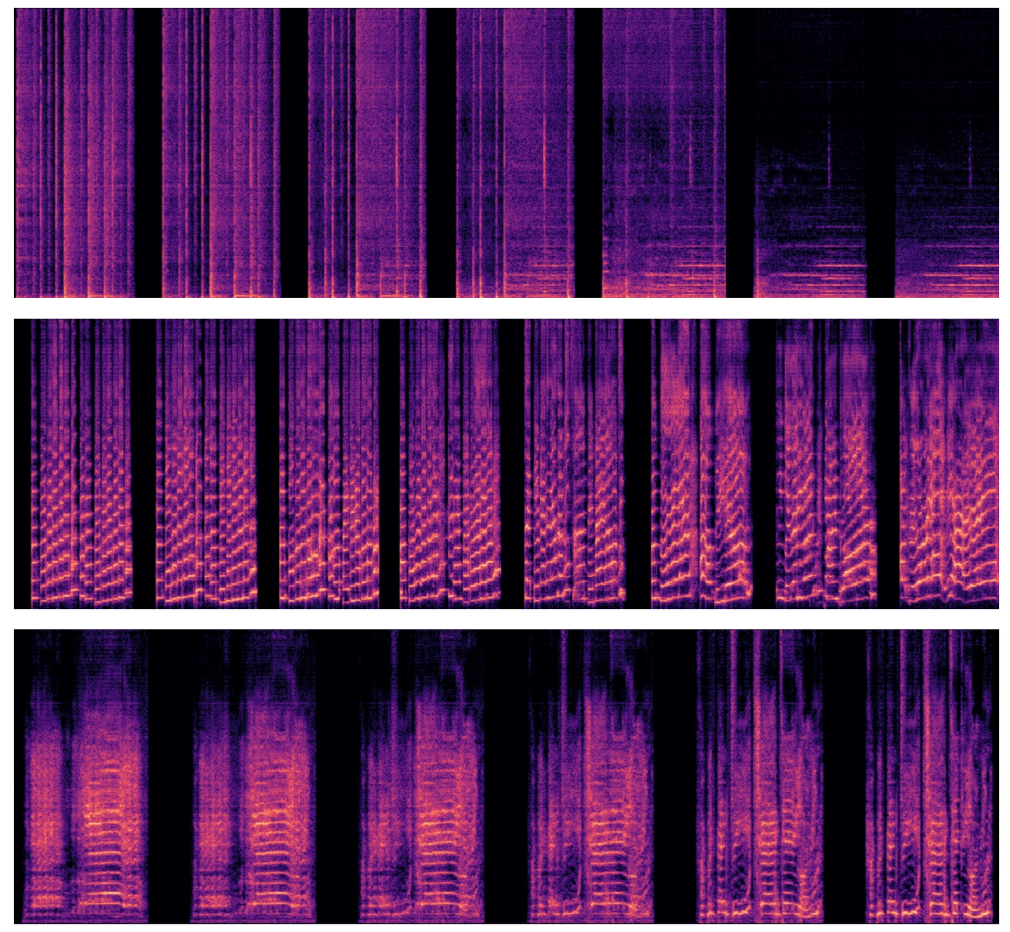
$$p_\theta(\boldsymbol{z}_{0:n_0} \mid \boldsymbol{E}^y) = p(\boldsymbol{z}_{n_0}) \prod_{n=1}^{n_0} p_\theta(\boldsymbol{z}_{n-1} \mid \boldsymbol{z}_n, \boldsymbol{E}^y)$$

Results

The manipulation result with different starting points n_0 of the shallow reverse process. The original signal is Trumpet, and the text prompt for style transfer is Children Singing.



Example mel-spectrogram of audio style transfer:



From drum beats to ambient music.

From trumpet to children singing.

From sheep vocalization to narration, monologue.

Conclusions

We explored another possible application of TTA systems: text-guided audio manipulations without finetuning the model on a specific task. And we provide showcases to prove the effectiveness of the design and the potential of TTA systems.

References

[1] Liu, Haohe et al. "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models." *International Conference on Machine Learning* (2023).

[2] Pascual, Santiago et al. "Full-Band General Audio Synthesis with Score-Based Diffusion." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2022): 1-5.