SOUND SCENE SYNTHESIS BASED ON FINE-TUNED LATENT DIFFUSION MODEL FOR DCASE CHALLENGE 2024 TASK 7

Technical Report

Gaurav, Sagnik Ghosh, Shivesh Singh, Shubham Sharma, Siddharath Narayan Shakya

Indian Institute of Technology Mandi, India

ABSTRACT

With the advancements in generative AI, text-to-audio systems have become increasingly popular, transforming audio generation across various domains such as music and speech. These systems enable the generation of high-quality audio from textual descriptions, offering freedom and control when producing a variety of audio. This technical report explores advancements in deep learning applied to sound generation specifically focusing on environmental sound scene generation. Our approach leverages a Text-to-Audio (TTA) system with Contrastive Language-Audio Pretraining (CLAP), Conditional Latent Diffusion Models, a Variational Autoencoder (VAE) decoder, and a HiFi-GAN vocoder where LDM learn continuous audio representations from CLAP embeddings, enhancing synthesis control through natural language prompts. Also finetuned the diffusion model with the custom dataset created using two audio dataset in order to improve generation quality.

Index Terms— Sound scene generation, Diffusion Model, Custom dataset creation

1. INTRODUCTION

The recent advancements in deep learning have ushered in remarkable breakthroughs in various domains, including sound generation [1], [2], [3], [4]. Starting from the foley sound generation, where foley sound which is basically refers to sound effects created in order to convey or enhance events in a narrative, like radio or film for example: dog bark, moving car, rain sound etc is generated and this foley sound generation forms the fundamental aspect of creating realistic, individual sounds that mimic specific actions and objects within a scene for example: a dog barking with the rain in background. However, sound scene synthesis goes beyond foley by combining various foley-generated sounds into unified audio landscapes that mimic imagined or real-world circumstances.

Therefore, sound scene generation, in particular, plays an important role in enriching the overall auditory experience in movies, music, videos, and other multimedia content. The integration of sound scene synthesis systems holds tremendous promise in simplifying traditional sound generation processes, thereby reducing the reliance on manual recording and mixing by human artists.

The Task 7 of DCASE 2024 Challenge [5] is about generating a enviornmental sound given the textual description and this task also expands the scope of foley sound synthesis to more general case. The official baseline system for Task 7 [5] is a Text-to-Audio (TTA) system, which utilizes a Contrastive Language-Audio Pretraining (CLAP), Conditional Latent Diffusion Models, Variational Auto Encoder (VAE) decoder and HiFi-GAN Vocoder.

This report presents a method, we submitted to Task 7 of DCASE 2024 challenge [6]. The task involves synthesizing the environmental sound given a textual description. Environmental sounds synthesis system encompass any non-musical and unintelligible vocal sounds and also adds controllability with natural language in the form of text prompts. In previous TTA works, a potential limitation for generation quality is the requirement of large-scale high-quality audio-text data pairs, which are usually not readily available, and where they are available, are of limited quality and quantity which poses a challenging task, therfore to address this challenge, we adopt the concept of AudioLDM that is built on a latent space to learn continuous audio representations from CLAP[6] embeddings and the pretrained CLAP models enable us to train LDMs with audio embeddings while providing text embeddings as the condition during sampling. To achieve better result we adopt the technique of fine-tuning the LDM using the custom dataset which is created using the audios of the two audio datasets namely ESC-50 dataset [7] and Acoustic Scene dataset [8]. More details about the generation process are mention in the section 3 and 4 of this report.

The following sections of this technical report are organized as follows: Section 2 offers an insight into the proposed system. The methodology utilized by the network is elaborated upon in Section 3. Section 4 outlines the experimental setup employed. Results will be showcased in Section 5. Lastly, Section 6 encapsulates this endeavor, offering a summary and drawing conclusions.

2. OVERVIEW OF THE PROPOSED SYSTEM

The proposed system mirrors the baseline architecture, specifically AudioLDM, functioning as a Text-To-Audio (TTA) model. It leverages a diffusion model to acquire continuous audio representations through CLAP embeddings where CLAP model utilized is trained on LAION-Audio-630K, AudioSet, Music and Speech dataset. Pretrained CLAP models facilitate the training of Latent Diffusion Models (LDMs) with audio embeddings, simultaneously incorporating text embeddings as conditional inputs during the sampling process. The architecture of AudioLDM revolves around four key components: a U-Net based Latent diffusion Model, CLAP, a VAE decoder, and a High Fidelity Generative adversarial network (Hifi-GAN) vocoder. This system is then further improved with fine-tuning on the custom dataset which is generated by mixing the two audios from ESC-50 dataset and Acoustic Scene dataset. In-depth discussions on these methodologies are presented in the subsequent section. The overview of system is presented in Figure 1.

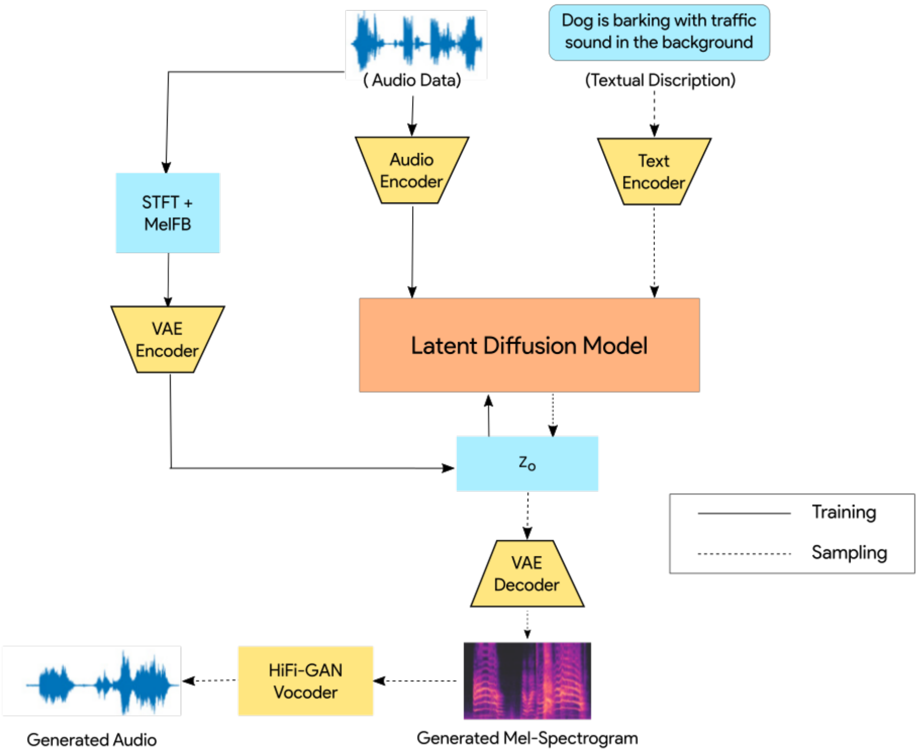

Figure 1: Overview of the System.

3. METHODOLOGY

3.1. Contrastive Language - Audio Pretraining (CLAP)

CLAP [6] which is based on the concept of Contrastive Language - Image Pretraining (CLIP) [9] is used as embedding encoder which is generating the input embeddings. CLAP consist of audio encoder $f_{\text{audio}}(.)$ and a text encoder $f_{\text{text}}(.)$. An audio encoder convert audio descriptions denoted as x into audio embeddings $\mathbf{E}^{x}$, and a text encoder, convert text description denoted as y into text embeddings $\mathbf{E}^{y}$. Both encoders undergo training using symmetric cross-entropy loss and this process yields a latent space characterized by consistent dimensionality for both audio and text embeddings.

3.2. Conditional Latent Diffusion Models

The Latent Diffusion Model serves as a generator, trained on audio embeddings derived from the trained CLAP audio encoder. During the sampling process, it utilizes text embedding as a conditional input. Diffusion process is consist of two process: 1) Forward Process: It transform the data distribution into the noise using a predefined noise schedule. 2) Reverse Process: It gradually generates the data sample from the complete noise according to the inference schedule. Throughout the training phase, the model is optimized with a re-weighted objective [10], denoted as $L_n()$, formulated as follows:

$$L_n(\theta) = E_{z_0,\epsilon,n}||\epsilon - \epsilon_\theta(z_n, n, E_x)||_2^2$$

Using a Gaussian noise sample x_o, the reverse transition probability learned during training, and the text condition (y) from CLAP, the model generates an outcome x during the sampling phase.

3.3. Decoder and Vocoder

During training Variational Autoencoder (VAE) is used to decode the generated latent token into a mel-spectrogram and during the training phase, the VAE learns to encode the mel-spectrograms, represented as $\mathbf{X}$ into a latent space vector z and subsequently reconstructs the mel-spectrogram back to $\hat{\mathbf{X}}$. Hifi-GAN is utilized as the vocoder which generate the sound $\hat{x}$ from the reconstructed mel-spectrogram denoted as $\hat{\mathbf{X}}$

3.4. Fine-Tuning with Custom Dataset

First the custom dataset is created using the two audio dataset namely, **ESC-50** dataset and **TAU Urban Acoustic Scenes 2020 Mobile development dataset**. And the LDM is fine-tune by utilizing this custom dataset. In-depth discussions on this is presented in the subsequent section.

4. EXPERIMENTS

4.1. Dataset Creation

Although the CLAP [6] is pretrained on large dataset like AudioSet and UrbanSound8k, the LDM trained finetuned by us is trained on a custom dataset prepared by mixing sounds from **ESC-50 dataset [7]** which is a labeled collection of 2000 environmental audio recordings and **TAU Urban Acoustic Scenes 2020 Mobile development dataset [8]** which consists of 10-seconds audio segments from 10 acoustic scenes. Few Samples were collected from the TAU Urban Acoustic Scenes containing 10 different Acoustic Scenes from 12 different cities. The custom dataset was created by overlaying two audio samples, one from each of these above mentioned dataset, let say, x $_1$ be the audio from ESC-50 dataset and x$_2$ be the audio from TAU Urban Acoustic Scenes 2020 Mobile development dataset and the new audio ,say x got after overlapping is formulated as:

$$x = \lambda_1 x_1 + \lambda_2 x_2$$

where λ_1 λ_2 are scaling factor which is equal to one , thereby creating a total of around 51k training sound samples. Experiments were conducted to amplify the foreground sound overlaying with a suppressed background audio to create audio samples resembling an Audio Scene, for example 'snoring with the sound of park in background'.

4.2. FAD Calculation

As per the given Evaluation metric, the Frechet Audio Distance (FAD) is calcualted based on the embedding vectors from two groups of audios calculated using features extracted from PANN [11] and CNN14 Wavegram-Logmel models. Here, the FAD score represents the similarity between ground truth audios and the generated audios. The lower the FAD Score, the better is the quality of generation.

4.3. Experimental Setup

The ESC50 [7] dataset and TAU Urban Acoustic Scene [8] dataset used was sampled at 44kHz each and overlayed on each other. A library called pydub was used in python environment to create the mix of these sounds. Further the train and test json files were constructed as per the file format used in AudioCaps [12] dataset. The dataset root file was constructed as part of metadata of our custom generated audio dataset and placed in metadata folder. We have used this dataset root json to train our LDM model.

4.4. System Information

The baseline LDM underwent training using four NVIDIA RTX™ A6000 GPU nodes, reaching a maximum step value of 52000 during the training process. Validation occurred after every five epochs, during which FAD scores were computed using VGGish, PANN, and MS-CLAP embeddings. The training, testing, validation and Variational Lower Bound (VLB) losses were monitored and recorded using the Weight & Biases server, and their respective curves were stored for analysis

5. RESULTS

The system's performance on the validation set is shown in Table 1. The FAD scores [13] [14] is utilized the evaluation metric which is formulated as follow:

$$F(N_r, N_g) = \|\mu_r - \mu_g\|_2^2 + \mathrm{tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}})$$

where $\mathcal{N}_r(\mu_r, \Sigma_r)$ and $\mathcal{N}_g(\mu_g, \Sigma_g)$ are, from the generated samples and the evaluation set, the multivariate Gaussians distribution of the VGGish embeddings, respectively.

Table 1: Results of AudioLDM models on development and external datasets

Model	FAD (dev dataset)	FAD (ext dataset)
AudioLDM (Baseline)	61.2761	61.2761
AudioLDM - finetuned	28.1756	31.5788

6. CONCLUSION

The solution that we presented for DCASE 2024 challenge task 7 is essentially described in this technical report. Our system basically utilizes state-of-the-art diffusion-based models and incorporates fine-tuning technique to get better results. Comparing with the baseline system our fine-tuned system significantly leverages the audio generation, achieving a low FAD score of 28.1756 on development dataset.

7. REFERENCES

[1] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," 2023.

[2] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *arXiv preprint arXiv:2207.09983*, 2022.

[3] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," *arXiv preprint arXiv:2301.12661*, 2023.

[4] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Defossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "Audiogen: Textually guided audio generation," in *International Conference on Learning Representations*, 2023.

[5] http://dcase.community/challenge2024/.

[6] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2023, pp. 1–5.

[7] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[8] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2021, pp. 626–630.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[10] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Conference on Neural Information Processing Systems*, 2020.

[11] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, “Panns: Large-scale pretrained audio neural networks for audio pattern recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[12] C. D. Kim, B. Kim, H. Lee, and G. Kim, “Audiocaps: Generating captions for audios in the wild,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[13] A. Gui, H. Gamper, S. Braun, and D. Emmanouilidou, “Adapting frechet audio distance for generative music evaluation,” *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1331–1335, 2024.

[14] M. Tailleur, J. Lee, M. Lagrange, K. Choi, L. M. Heller, K. Imoto, and Y. Okamoto, “Correlation of fréchet audio distance with human perception of environmental audio is embedding dependent,” *arXiv:2403.17508*, 2024.