DIFFUSION BASED SOUND SCENE SYNTHESIS FOR DCASE CHALLENGE 2024 TASK 7

Technical Report

Yi Yuan[1], *Haohe Liu*[1], *Xubo Liu*[1], *Mark D.Plumbley*[1], *Wenwu Wang*[1]

[1] University of Surrey, Guildford, United Kingdom

ABSTRACT

Sound scene synthesis aims to generate a variety of environment-related sounds within a specific scene. In this work, we proposed a system for DCASE 2024 challenge task 7. The proposed system is based on the official baseline model AudioLDM, a diffusion-based text-to-audio generation model. The system first trained with large-scale datasets and then downstream into this task via transfer learning. Addressing the challenge of no target audio data, we implemented an automated pipeline to synthesize audio and generate corresponding captions that mirror the semantic structure of the task. Despite the absence of dedicated training and testing sets for this task, our robust audio synthesis model effectively adapts the given conditions, fulfilling all the task requirements. Our system achieved a Fréchet Audio Distance (FAD) score of 55.1, surpassing the baseline system's FAD score of 61.3 calculated by the official evaluation toolkit.

Index Terms— Sound scene generation, Diffusion model, Transfer learning, Language model

1. INTRODUCTION

Nowadays, diffusion-based generative architecture has contributed to remarkable breakthroughs in audio generation [1, 2, 3]. A particular area of interest within this field is the generation of environmental sounds, which include both sound effects and natural vocal sounds. The DCASE 2024 challenge Task 7 is organized to tackle this challenging problem. In this report, we introduce our system submitted to this challenge.

Specifically, Task 7 focuses on synthesizing sounds in various scenarios, involving a combination of foreground and background sounds. The input to the system is structured as a prompt comprising a foreground source described by an action verb and a background source. For example, a prompt might be: "a dog is barking with water in the background." Notably, the categories of foreground and some categories of background sounds are not predefined, which adds complexity to the task and makes it challenging to train the model into an expert in generating accurate audio representations in this particular domain.

State-of-the-art (SOTA) audio generation models typically follow a two-stage generation pipeline. They use an encoder-decoder architecture to compress the waveform and a generative module to produce the audio features. The baseline system for Task 8 [4] employs a similar approach: a latent diffusion-based model to generate audio features within the latent space, a variational autoencoder (VAE) decoder to reconstruct the information into a mel spectrogram, and a pre-trained generative adversarial network (GAN) vocoder [5] to produce the final waveform.

Our proposed system builds on this baseline with several enhancements. We first substitute the previous CLAP model [6] with the T5 model [7] for text embedding, then integrate a cross-attention module to effectively align the input conditions with the system. This change is crucial since the task relies on text prompts and lacks a development dataset, necessitating a robust and adaptable model. Additionally, to improve the quality of audio generation, we replace the HifiGAN vocoder with BigVGAN [8], known for its superior performance at higher frequencies. Furthermore, the task requires generating 32kHz, 4-second audio clips. Most existing audio-language datasets feature 10-second waveforms, and the baseline model is trained on 16kHz audio. To align with the task requirements, our system is initially trained on 32kHz, 10-second audio and subsequently fine-tuned on 32kHz, 4-second clips using transfer learning techniques. To address the scarcity of training data, we also generate synthetic audio and captions by concatenating various audio clips during the fine-tuning stage.

The structure of this report is as follows: Section 2 presents the details of our system architecture. Section 3 describes the training pipeline of our system, including specific training configurations and data processing methods for each stage. Finally, we present our results and conclude the report.

2. SYSTEM METHODOLOGY

Similar to the baseline system, the system also consists of four main sections, a text embedding encoding module, an audio feature generating module, a waveform reconstruction module and a similarity selection module. Our system takes the same diffusion-based backbone for sound generation and a VAE decoder for mel-spectrogram reconstruction. Instead of the CLAP model used for text embedding in the baseline model, we use the T5 [7] model, which is more capable of extracting the semantic feature of textual information. Nevertheless, the previous Hifi-GAN-based vocoder is replaced by BigVGAN-based for better waveform generation on high-frequency domains. As for the inference procedure, we follow the idea of AudioLDM [4] by generating several audios and using the CLAP [6] model to pick the candidate that best matches the prompt.

Like the baseline system, our architecture comprises four primary components: a text embedding encoding module, an audio feature generation module, a waveform reconstruction module, and a similarity selection module. We utilize the same diffusion-based framework for sound generation and a VAE decoder for mel-spectrogram reconstruction.

However, our system introduces several key enhancements. Instead of the CLAP model for text embedding used in the baseline, we employ the T5 [7] model, which shows better performance at capturing the semantic nuances of textual input. For waveform generation, we replaced the previous Hifi-GAN-based vocoder

with BigVGAN, which provides superior performance in generating high-frequency content. During inference, our approach follows the methodology of AudioLDM [4]. We generate multiple audio samples and then use the CLAP [6] model to select the candidate that best aligns with the given prompt. This process ensures that the generated audio is closely matched to the specified textual description. Detailed explanations of these methods are provided in the following section.

2.1. Text encoding

For embedding the input text prompt, we replaced the previous CLAP model with the Flan-T5 model, which has demonstrated superior performance in extracting semantic features from text [9]. Unlike the CLAP encoder, which can process both audio and text, the pre-trained Flan-T5 model is specifically designed to handle text inputs only. To enhance the robustness of our system and compensate for this specialization, we trained the model using a larger audio-caption dataset. This approach ensures that our system can effectively leverage the rich semantic information embedded in textual prompts.

2.2. Audio feature generating

Our system applied the latent diffusion model (LDM) for the generation. In detail, our model takes the textual embedding as the condition and generates the related audio feature as latent tokens. LDM consists of two processes, a forward process that incrementally adds noise $\boldsymbol{\epsilon}$ to the latent vector $\boldsymbol{z}0$, and a reverse process entails the model predicting the transition probabilities $\boldsymbol{\epsilon}\theta$ for each step n. resulting in a sequence of latent vectors $\boldsymbol{z}n$ over N steps, configuring the training loss as:

$$L_n(\theta) = E_{\boldsymbol{z}_0,\boldsymbol{\epsilon},n}||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_n, n, \boldsymbol{E}^x)||_2^2 \quad (1)$$

Previously, the conditioning strategy in our model involved adding the textual features as an additional layer concatenated with the step information within each layer of the U-Net backbone. Our proposed method integrated the conditioning information directly into the network by augmenting the feature maps at every convolutional layer. To more effectively leverage the embedding conditions from the T5 model, we have refined this approach. Instead of simple concatenation, we now employ a cross-attention module after each convolution layer. This enhancement allows the model to incorporate the text embeddings into the audio generation process.

2.3. VAE decoder & HiFi-GAN vocoder

We trained a 32kHz Variational Autoencoder (VAE) to decode the latent feature tokens into mel-spectrograms. Following the approach of the baseline model, our VAE is designed to compress mel-spectrograms into latent space tokens $\boldsymbol{z}0$ and then reconstruct these tokens back into mel-spectrograms. For the final waveform generation, we employed BigVGAN [8], a state-of-the-art vocoder known for its ability to produce high-fidelity audio. BigVGAN excels at generating detailed sound waveforms, particularly in the high-frequency range, which is crucial for achieving clear and natural audio outputs.

2.4. Similarity selection

To further enhance the sound quality, AudioLDM integrated a scoring mechanism to select the most suitable audio outputs. This mechanism leverages the CLAP model, which utilizes a shared latent space for both audio and text embeddings. By calculating the cosine similarity between the generated audio and the target text embeddings, the system can effectively measure how closely the audio matches the intended description. In our approach, we further improve the overall performance by replacing the previous CLAP-LAION model [6] with CLAP-Micro [10]. The CLAP-Micro model has demonstrated greater robustness across various features, making it more effective in assessing the relevance and quality of the generated audio in diverse scenarios.

3. EXPERIMENTS

3.1. Dataset

Challenge official dataset. The official dataset for the challenge provides only 60 audio feature embeddings derived from three different audio encoders, which is insufficient for training a robust and powerful system. To address this limitation, we initially trained our model on a large-scale audio-caption dataset with 32kHz, 10-second data, followed by fine-tuning using 4-second audio-caption pairs.

Pretraining dataset. Our pretraining stage utilized an extensive audio dataset that captured a wide variety of sounds. In detail, we employed AudioSet, the largest available audio dataset, which contains approximately 2.1 million 10-second audio clips annotated with labels. For textual representation processed by the T5 encoder, we used captions automatically generated by Large Language Models (LLMs) to match the audio content.

Fine-tuning dataset. For the fine-tuning stage, we used a dataset comprising 4-second audio samples to align with the task requirements. We sourced this data from three distinct datasets: Wavecap [11], ESC50 [12], and Urbansound8K [13], only collecting the audio clips of less than 4 seconds. To prepare the training data, two different sound categories into each sample, with one serving as a foreground source and the other as a background source. Captions were synthetically generated to reflect the sequence structure described in the task, ensuring coherence with the expected outputs. All the audio and captions are generated randomly when loading the data to enhance the robustness of the model.

3.2. Experimental process

As an ensemble model, our system is developed in three stages. First, the decoder and vocoder are trained with 32Khz audio for more than 100K steps independently. Then, the diffusion model is developed for 50K steps on extensive datasets in 10 seconds. Lastly, the system is fine-tuned through transfer learning on 4-second audio clips for another 50K steps.

3.3. Results

Table 1: The FAD score using different audio embedding models, where P is short for PANNS, M for Micro$_c lapandV forVGG - ish$.

Model	FAD$_P\downarrow$	FAD$_M\downarrow$	FAD$_V\downarrow$
Baseline	61.3	-	-
Proposed System	55.1	312.5	7.5

Although our system has achieved significant improvements in various aspects, enhancing performance for this task remains challenging without access to specific training data. Table **??** illustrates the Fréchet Audio Distance (FAD) scores for different audio embedding strategies as provided by the official evaluation tool [14]. When comparing the audio generated by our system to the audio embeddings from the development set, our system consistently outperforms the baseline across all three audio embedding approaches.

4. CONCLUSION

This technical report details the system we developed for the DCASE 2024 Challenge Task 7. Our approach builds on the baseline model and integrates several advanced techniques to enhance audio quality. Experimental results demonstrate that our system significantly surpasses the baseline model, achieving substantial performance improvements. Tackling the challenge of developing a system without specific development and evaluation sets not only presents a significant research opportunity but also aligns well with the practical demands of real-world applications. In the future, we aim to explore the potential of even more robust and versatile models capable of excelling across diverse tasks.

5. REFERENCES

[1] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[2] J. Huang, Y. Ren, R. Huang, D. Yang, Z. Ye, C. Zhang, J. Liu, X. Yin, Z. Ma, and Z. Zhao, "Make-an-audio 2: Temporal-enhanced text-to-audio generation," *arXiv preprint arXiv:2305.18474*, 2023.

[3] Y. Yuan, H. Liu, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, "Retrieval-augmented text-to-audio generation," in *International Conference on Acoustics, Speech, and Signal Processing*, 2023.

[4] 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," in *International Conference on Machine Learning*, 2023.

[5] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020, pp. 17 022–17 033.

[6] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[7] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[8] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "Bigvgan: A universal neural vocoder with large-scale training," *arXiv preprint arXiv:2206.04658*, 2022.

[9] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction tuned LLM and latent diffusion model," *arXiv preprint arXiv:2304.13731*, 2023.

[10] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP learning audio concepts from natural language supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[11] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *arXiv:2303.17395*, 2023.

[12] K. J. Piczak, "ESC: dataset for environmental sound classification," in *Proceedings of the ACM International Conference on Multimedia*, 2015.

[13] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *International Conference on Multimedia*, 2014.

[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.