FOLEY SOUND SYNTHESIS WITH AUDIOLDM FOR DCASE2023 TASK 7

Technical Report

Shitong Fan[1]*, Qiaoxi Zhu*[2]*, Feiyang Xiao*[1]*, Haiyan Lan*[1]*, Wenwu Wang*[3]*, Jian Guan*[1*]

[1]Group of Intelligent Signal Processing (GISP), College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[2]Centre for Audio, Acoustic and Vibration (CAAV), University of Technology Sydney, Ultimo, Australia
[3]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK

ABSTRACT

This report describes our submission for DCASE2023 Challenge Task 7, a system for foley sound synthesis. Our system is based on AudioLDM, a high-quality and computationally efficient text-to-audio generation model. Experiments are conducted on the dataset of DCASE2023 Challenge Task 7. The Fréchet audio distance (FAD) between the sound generated by our system and the actual sound sample is 5.120 in the category "DogBark" and 8.102 in the category "Rain", better than the baseline with a FAD of 7.256 and a FAD of 4.901 distance closer to the actual samples, respectively.

Index Terms— Foley Sound Synthesis, Text-to-Audio Generation, Diffusion Model.

1. INTRODUCTION

The subject of DCASE2023 Challenge Task 7 is "Foley Sound Synthesis" [1], where "Foley" refers to sound effects added to multimedia during post-production to enhance its perceived acoustic properties [2]. The task aims to generate raw audio clips representing a class of sounds, e.g., dog barking and footsteps. The development of this task can be beneficial to fit the growing demand for the automated generation requirement of Foley sounds in virtual environments. As a simple but effective approach, we use AudioLDM [3], the state-of-the-art method in the text-to-audio task, to synthesize Foley sounds for this challenge.

2. METHOD

Our submission is based on AudioLDM for foley sound synthesis. First, we define the description prompt for each category. Then, we input them into AudioLDM to generate synthetic foley sounds corresponding to each category. As a result, we leverage the advantages of AudioLDM to generate high-quality foley sounds. Based on the audio synthesis quality, we have predefined the description prompt for the AudioLDM model. Table 1 provides the descriptive prompt we have predefined for each category.

*Corresponding author.

This work was partly supported by the Natural Science Foundation of Heilongjiang Province with Grant No. YQ2020F010 and LH2022F010, and the GHfund with Grant No. 202302026860.

3. EXPERIMENT

We conduct the experiments on the dataset of DCASE2023 Challenge Task 7, which consists of seven categories with 100 real audio samples for each category. To meet the submission requirement, we generate 100 foley sounds for each category in the experiments. For example, regarding the category "DogBark", we first predefine a description prompt according to the quality of audio synthesis and then input the prompt into AudioLDM for the generation. Finally, the system generates 100 different foley sounds for the category "DogBark".

The evaluation metric is the Fréchet Audio Distance (FAD) [4] between the generated and original audio samples. We calculate each category's FAD metric, and the evaluation results are presented in Table 2. The results show that our method can achieve 5.120 on the "DogBark" and 8.102 on the "Rain", which were 7.256 and 4.901 better than the baseline, respectively.

We also provide a combined system to take advantage of both the baseline and proposed systems. For categories that the proposed system does not perform well, the combined system will turn to the Baseline system. The experimental results are shown in Table 2 as "Combined System". Because of the randomness in generating the 100 audio clips, there is a small fluctuation in the FAD of the audio generated by the same strategy in the same category. The results show that our second system achieved 7.657 on the average FAD, which was 1.893 better than the baseline.

4. CONCLUSION

In this technical report, we presented our submission systems for DCASE2023 Challenge Task 7, using AudioLDM for foley sound synthesis. Evaluation is conducted on the DCASE2023 Challenge Task 7, and the experimental results show that our proposed system outperforms the baseline systems in generating dog barking sounds and the sound of rain.

5. REFERENCES

[1] K. Choi, J. Im, L. Heller, B. McFee, K. Imoto, Y. Okamoto, M. Lagrange, and S. Takamichi, "Foley sound synthesis at the dcase 2023 challenge," *arXiv preprint arXiv:2304.12521*, 2023.

[2] K. Choi, S. Oh, M. Kang, and B. McFee, "A proposal for foley sound synthesis challenge," *arXiv preprint arXiv:2207.10760*, 2022.

Table 1: The descriptive prompts were predefined for each category.

Category	Prompts
DogBark	This is a dog barking in the park.
Footstep	This is the sound of a person walking in the room wearing leather shoes.
GunShot	This is the sound of a gunshot.
Keyboard	This is the sound of a computer keyboard.
MovingMotorVehicle	This is the sound of a motor vehicle revving its engine.
Rain	This is the sound of it raining heavily and raindrops falling on the road.
Sneeze/Cough	The man with an uncomfortable throat is coughing.

Table 2: The Frechet audio distance between synthesized foley sounds and the actual samples.

Methods	DogBark	Footstep	GunShot	Keyboard	MovingMotorVehicle	Rain	Sneeze/Cough	Average
Baseline [5]	12.376	**7.624**	8.794	**4.037**	17.831	13.003	**3.184**	9.550
AudioLDM based System	**5.120**	10.536	**8.709**	4.982	**17.010**	**8.102**	5.344	8.543
Combined System	**5.397**	**6.962**	**8.616**	**4.152**	**17.442**	**7.752**	**3.251**	**7.657**

[3] 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 *Proc. of the International Conference on Machine Learning (ICML)*. IEEE, 2023.

[4] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms." in *Proc. of INTERSPEECH*, 2019, pp. 2350–2354.

[5] X. Liu, T. Iqbal, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Conditional sound generation using neural discrete time-frequency representation learning," *International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2021.