SOUND SCENE SYNTHESIS WITH AUDIOLDM AND TANGO2 FOR DCASE 2024 TASK7

Technical Report

Yu Sun, Zhidong Xie, Haicheng Liu, Xinyu Xin, Xiaoyan Zou

Samsung Electronics China R&D Center

ABSTRACT

This report describes our submission for DCASE2024 Challenge Task 7, a system for sound scene synthesis. Our system is based on AudioLDM, a text-to-audio generation model improved both in generation quality and computational efficiency and Tango2, based on tango while compare to Tango and AudioLDM2, it can generates better result in some metrics. Experiments are conducted on the dataset of DCASE2024 Challenge Task 7. The Frechet Audio Distance (FAD) between the sound generated by our system and the develop set is 60.64.

Index Terms— Sound Scene Synthesis, AudioLDM, Tango2

1. INTRODUCTION

The subject of DCASE2024 Challenge Task 7 is “Sound Scene Synthesis”, which aims to generate a more general environmental sound scene given a text prompts [1], which is a more general next-generation task compared to last year’s Foley sounds text. In this test, the caption of input can be divided into two parts: foreground and background. For rapid development, we use a scheme which combines AudioLDM [2] and Tango2 [3] in a certain criteria which are both state-of-the-art TTA model.

2. METHOD

We use a combined system for this test in the way of combining AudioLDM and Tango which both have great perform in text-to-audio task.

AudioLDM, a TTA system that is built on a latent space to learn continuous audio representations from contrastive language-audio pretraining embeddings. Overview of the AudioLDM system for text-to-audio generation are as follow figure 1.

Tango2 is based on Tango and use diffusion-DPO (direct preference optimization) to get a better result over Tango and AudioLDM2. Overview of the Tango2 for text-to-audio generation are as follow figure 2.

To use the advantage of the two models, we combined the two model in some way. First, we retrain the small model of AudioLDM by audiocaps dataset. Then, we try two ways to combine the two models: combine the mel-spectrum of the two model or replace the audio generated by AudioLDM with Tango2. As the result shows in experiment part, the first mixed way has bad result. In the second way, we select the bad effect case by cos similarity. Next, we replace the bad case by the result inferred by Tango2. As a result, we can get a great balance in FAD score and auditory sensations.

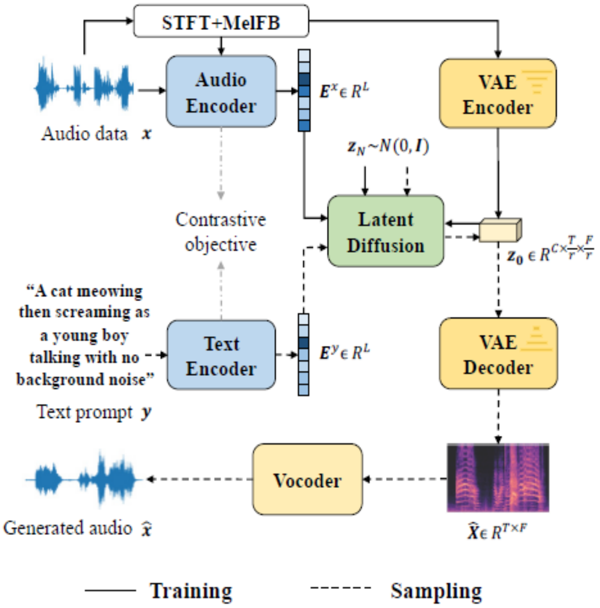

Figure1 AudioLDM diagram

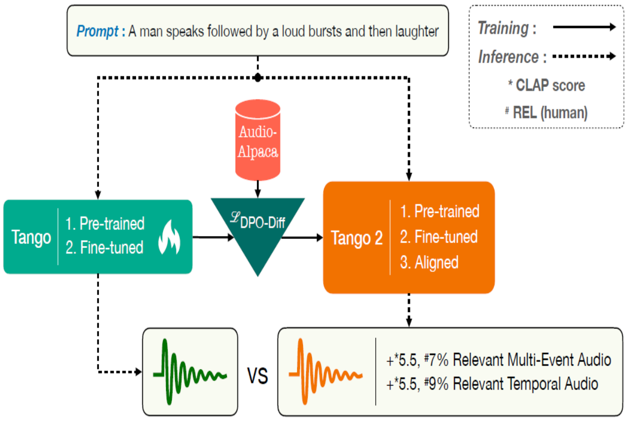

Figure2 tango2 diagram

Base above two models, The overview of our solution's can be seen from following picture.

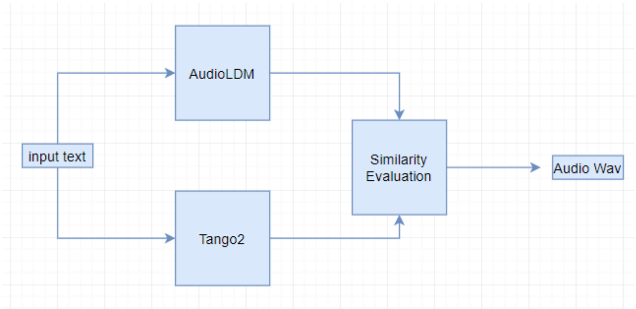

Figure3 Overview diagram

The input text will be sent both AudioLDM and Tango2 model. Then we use Similarity Evaluation Model to make sure the audio's reality and diversity, choice the better one as the final ouput result.

3. EXPERIMENT

We use the develop set of DCASE2024 Challenge Task 7 to conduct our experiments. There are 60 pairs of prompts and audio embeddings in the develop dataset. Table below shows some of the caption.

Table 1

a buzzer is ringing with water in the background
a pig is grunting with water in the background
an alarm of a car door staying open is ringing with crowd in the background
a small dog is whining with water in the background
a car horn is honking with crowd in the background
a zipper is zipping with traffic in the background
a baby is sneezing with crowd in the background
a pen is writing with crowd in the background
a small gun is shooting with traffic in the background
a small dog is barking with crowd in the background

We use the Frechet Audio Distance (FAD) [4] between inferred audio and original audio embeddings.
First, we calculate the first way of mix and the result of the way of mix mel is as follows:

Table 2

	FAD
AudioLDM	48.7082
mix with Tango2	54.5723

We compare different seed of the model AudioLDM and get the result that different seed of the same model may lead to different result. We test three couple seeds such as 0, 1234, 5678 and the FAD result of the three seeds are as follow.

Table 3

seed	FAD
0	48.7082
1234	46.7387
5678	44.8741

Compare the result of baseline which is 61.2761, the FAD score of our system is 45.5548. The result shows that our system can achieve 15.7213 better than the baseline.

4. CONCLUSION

Using self-train AuidoLDM model, FAD model has better result. But the audio has large noise and sometimes the sound is not corresponding with object descripted. Tango2 model could generate high quality audio without noisy and it's audio has better alignment with input text. The weak point is FAD score is not ideal. Maybe the sample's diversity is not rich enough. Based on above facts, we decide to combine this two model's advantages and use the Tango2 as mainly model to generate high quality sample and AudioLDM model as supplement to improve the diversity when Tango2's result has poor similarity with input text. In this report, we use a combined system to submit DCASE2024 Challenge task 7, using AudioLDM and Tango2 for sound scene synthesis. The experiment results show that our combined system have a better perform in development datasets.

5. REFERENCES

[1] https://dcase.community/challenge2024/task-sound-scene-synthesis

[2] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio gen-

eration with latent diffusion models," in Proc. of the International Conference on Machine Learning (ICML). IEEE, 2023.

[3] Navonil Majumder, Chia-Yu Hung and Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, Soujanya Poria,"Tango 2: Aligning Diffusion-based Text-to-Audio Generations through Direct Preference Optimization",in arXiv 2024.

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