LANGUAGE-QUERIED AUDIO SOURCE SEPARATION WITH GPT-BASED TEXT AUGMENTATION AND IDEAL RATIO MASKING

Technical Report

Feiyang Xiao[1]*, Wenbo Wang*[2]*, Dongli Xu*[3]*, Shuhan Qi*[4]*, Kejia Zhang*[1]*, Qiaoxi Zhu*[5]*, Jian Guan*[1*]

[1]College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[2]Faculty of Computing, Harbin Institute of Technology, Harbin, China
[3]Independent Researcher, China
[4]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China
[5]University of Technology Sydney, Ultimo, Australia

ABSTRACT

This technical report details our submission systems for Task 9 (language-queried audio source separation) of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2024 Challenge. Our four proposed systems utilize the large language model GPT-4 for data augmentation and apply the ideal ratio masking strategy in the latent feature space of the text-to-audio generation model, AudioLDM. Additionally, our systems incorporate the pre-trained language-queried audio source separation model, AudioSep-32K, which leverages extensive pre-training on large-scale data to separate audio sources based on text queries. Experimental results demonstrate that our systems achieve better separation performance compared to the official baseline method on objective metrics. Furthermore, we introduce a novel evaluation metric, audio-text similarity (ATS), which measures the semantic similarity between the separated audio and the text query without requiring a reference target audio signal.

Index Terms— Language-queried audio source separation, GPT-based augmentation, ideal ratio masking, latent diffusion

1. INTRODUCTION

Language-queried audio source separation focuses on separating a target audio source from a mixture of audio signals based on a text query, such as an audio caption, enhancing the clarity of semantic and temporal relationships of sound events in the separated output [1, 2]. This technical report presents our submitted systems for the language-queried audio source separation task in DCASE 2024 Challenge Task 9.

Our submitted systems are based on three key points, i.e., the text augmentation based on GPT-4, the ideal ratio masking strategy on the latent diffusion feature space, and the up-sampling process to incorporate the pre-trained language-queried audio source separation model, i.e., AudioSep-32K [2].

Furthermore, we introduce a novel evaluation metric for language-queried audio source separation, termed audio-text similarity (ATS). Unlike the official metrics that require a reference audio signal to measure the separation performance, the proposed ATS metric evaluates the semantic similarity between the separated audio signal and the text query without needing a reference audio signal. This makes our ATS metric particularly useful for real-world applications where a reference audio signal may not be available.

2. OUR SUBMITTED SYSTEMS

We present three individual language-queried audio source separation models as our systems 1-3, respectively. The details of these systems are given as below.

2.1. System 1: Using GPT-based Text Augmentation

We employ the large language model, i.e., GPT-4 [3], to augment the quantity of the available text queries in the training set, and then use the augmented training set to train the baseline system [1].

Specifically, following the previous text-audio retrieval studies [4, 5], we use the predefined prompt, the ground truth text query and keywords of the target audio signal as the input of GPT-4, which asks GPT-4 to generate 5 audio captions describing the sound events corresponding to the target audio signal for each text query in the development split of FSD50K dataset [6]. The predefined prompt is illustrated in Table 1.

With the augmented text queries, more diverse text information can be learnt to separate the corresponding audio information of the target audio signal using the separation model. Here, we use the augmented training set to train the baseline model [1] as our system 1, which is a ResUNet based separation model [7].

2.2. System 2: Using pre-trained AudioSep-32K

In this system, we employ the pre-trained language-based audio source separation model, i.e., AudioSep-32K, as our system 2. The AudioSep-32K model works for the audio signal with the sampling rate of 32 kHz, whereas the sampling rate of the evaluation set in DCASE 2024 Challenge Task 9 is set as 16 kHz. To use the AudioSep-32K for separation on the evaluation set in DCASE 2024 Challenge Task 9, we up-sampled the audio signals in the evaluation set from 16 kHz to 32 kHz and passed them to AudioSep-32K. Then, the separated audio signals are down-sampled from 32 kHz to 16 kHz.

*Corresponding author.

Table 1: Illustration of the GPT-based augmentation. The content in blue is the ground truth text query, and the contents in orange are the keywords for the target audio signal.

Prompt	Response of GPT-4
I will give a description of an audio recording and tags associated with the audio recording. Generate 5 audio captions describing the sound events. Each audio caption should be one sentence with less than 20 words. Use grammatical subject-verb-object sentences. Do not include names of cities, countries, and persons. Do not include the time. Do not include recording device names. Do not write introductions or explanations. Do not use "heard", "recorded". Start each output sentence with its index. "The act of breathing creates audible respiratory sounds." "breathing, respiratory sounds"	1. Air flows in and out, producing soft, rhythmic respiratory sounds. 2. Gentle inhalations and exhalations generate a calming pattern of noise. 3. The chest rises and falls with each breath, emitting quiet sounds. 4. Deep breaths cause a distinct, soothing sound of air movement. 5. Shallow breathing results in faint, yet discernible respiratory noises.

2.3. System 3: Using Ideal Ratio Masking in Generation Model

Previous audio source separation studies show that the ideal ratio masking strategy [8] is an effective choice to separate the target audio signal from the multi-sources mixture audio signal. Therefore, we introduce the ideal ratio masking strategy to separate the target audio signal with the guidance of the text query, denoted as IRM-LDM. Moreover, to reduce the requirement of the computation and memory, we employ the ideal ratio masking strategy on the latent feature space of the pre-trained text-to-audio generation model, i.e., AudioLDM [9].

Specifically, we firstly extract the latent audio feature of the mixture audio signal $\mathbf{z}_{\text{mix}}$, with the help of the variational auto-encoder (VAE) encoder in AudioLDM. Then, an ideal ratio masking model is introduced to predict the mask for the latent audio feature of the mixture audio signal, where the mask denotes the ratio of the component of the target audio source. With the predict mask, the latent feature of the target source is predicted, denoted as $\hat{\mathbf{z}}$. The ideal ratio masking model consists of 5 similar ResUNet blocks in [1]. The predicted latent feature of the target audio source is converted into the separated waveform $\hat{\mathbf{s}}$ with the VAE decoder and the vocoder structure in AudioLDM.

For model optimization, apart from the original loss function $\mathcal{L}_{\text{LDM}}$ of the latent diffusion model in AudioLDM, we also employ the mean squared error (MSE) loss to measure the distance between separated waveform $\hat{\mathbf{s}}$ and the provided target audio signal $\mathbf{s}$ in the training set. It can be represented as

$$\mathcal{L}_{\text{MSE}} = \|\hat{\mathbf{s}} - \mathbf{s}\|_2^2, \tag{1}$$

where $\mathcal{L}_{\text{MSE}}$ denotes the loss value of the MSE loss function. The overall loss can be represented as

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{LDM}}. \tag{2}$$

Therefore, our IRM-LDM system can obtain the separation ability to separate the target audio signal from the audio mixture with the guidance of text query, and obtain diverse separated audio signals with generation ability of latent diffusion.

3. EXPERIMENTAL RESULTS

3.1. Dataset

We conduct experiments on the Clotho [10] and FSD50K [6] datasets. Two random selected audio signals are mixed together as the mixture audio signal. One of these two audio signals is considered as the target audio signal, and the caption of this audio signal is set as the text query for the target audio signal. Therefore, we can obtain the mixture audio signal, the target audio signal and the text query related to the target audio signal for model training.

3.2. Setting up

The learning rate for our systems is set as 0.001. All of our systems are updated with the AdamW optimizer [11]. The warm up strategy is used on the the first 10000 training steps to avoid the gradient explosion.

3.3. Evaluation Metrics

To evaluate the performance of the language-queried audio source separation models, we employ the official evaluation metrics, i.e., signal-to-distortion ratio (SDR), improvement in signal-to-distortion ratio (SDRi) and scale-invariant signal-to-distortion ratio (SI-SDR) to evaluate the quality of the separated audio signal. Moreover, we introduce an additional evaluation metric for Task 9, i.e., the audio-text similarity (ATS), to measure the semantic similarity between the separated audio signals and the text queries.

Official Metrics: SDR, SDRi and SI-SDR are the widely used blind source separation metrics. SDR measures the ratio of the power of the desired signal to the power of the distortion introduced by the separation process. SDRi is an improvement metric that measures the difference in SDR before and after applying an audio source separation algorithm. SI-SDR normalizes the signals to make the evaluation independent of their amplitude, which is more robust for varying scales. Higher value of these metrics indicates better separation performance.

Proposed ATS Metric: To measure how well the separated audio matches the text queries, we introduce the audio-text similarity (ATS) metric. This metric helps us see how closely the content of the separated audio aligns with the text query. A higher ATS score means that the separated audio's content is more similar to the text query, indicating better performance in separating audio based on the text query.

The calculation of the proposed ATS metric is based on the contrastive language-audio pretraining (CLAP) module [12]. Specifically, the audio embedding of the separated audio signal and the text embedding of the text query are obtained with the CLAP module. Then, the cosine similarity between the audio embedding and the text embedding is calculated as the ATS metric to measure the semantic similarity between the separated audio and the text query.

Table 2: Performance of our systems, compared with the baseline.

Method	SDR	SDRi	SI-SDR	ATS
baseline	5.708	5.673	3.862	0.229
System 1	5.937	5.902	4.191	0.231
System 2	**8.192**	**8.157**	**6.680**	**0.249**
System 3	-4.757	-4.792	-42.346	0.193

Notably, the calculation of the ATS only needs the separated audio signal and the text query. It does not require referent target audio signal for calculation, as that for the official metrics. In this case, the proposed ATS metric can be more applicable in the real-world application that does not have the referent target audio signals.

3.4. Results

The results of our systems and the baseline are shown in Table 2. It can be found that our system 2 achieves the best performance. With the augmented text queries, our system 1 outperforms the baseline system. This illustrates that the GPT-based augmentation is an effective way to improve the language-queried audio source separation performance. The system 3 has worse performance on SDR, SDRi and SI-SDR metrics. The reason maybe that the generation process in the latent diffusion introduces unexpected noise.

4. CONCLUSION

In this technical report, we present our submission systems for DCASE 2024 Challenge Task 9, the language-queried audio source separation task. In our systems, we introduce the GPT-based text augmentation strategy, the pre-trained separation model, and an ideal ratio masking strategy in the generation model, i.e., the AudioLDM model, to build our systems, and achieve improved separation performance. Moreover, except the official evaluation metrics, we propose a novel evaluation metric, i.e., audio-text similarity, for this task. The proposed metric can measure the semantic similarity between the separated audio signal and the text query, without the referent target audio signal. It can be beneficial for the real world application where the referent target audio signal is unavailable.

5. REFERENCES

[1] X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Separate what you describe: Language-queried audio source separation," in *Proc. INTERSPEECH 2022*, 2022.

[2] X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, R. Xia, Y. Wang, M. D. Plumbley, and W. Wang, "Separate anything you describe," *arXiv preprint arXiv:2308.05037*, 2023.

[3] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[4] P. Primus, K. Koutini, and G. Widmer, "CP-JKU's submission to task 6b of the DCASE2023 challenge: Audio retrieval with PaSST and GPT-augmented captions," DCASE2023 Challenge, Tech. Rep., June 2023.

[5] P. Primus, K. Koutini, and G. Widmer, "Advancing natural-language based audio retrieval with PaSST and large audio-caption data sets," in *Proc. Detect. Classif. Acoust. Scenes Events (DCASE) Workshop*, Tampere, Finland, September 2023, pp. 151–155.

[6] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 829–852, 2021.

[7] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, "Decoupling magnitude and phase estimation with deep ResUNet for music source separation," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, 2021, pp. 342–349.

[8] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. IEEE, 2013, pp. 7092–7096.

[9] 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. Int. Conf. Mach. Learn. (ICML)*, 2023.

[10] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. IEEE, 2020, pp. 736–740.

[11] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[12] 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 *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. IEEE, 2023, pp. 1–5.