LANGUAGE-QUERIED AUDIO SOURCE SEPARATION ENHANCED BY EXPANDED LANGUAGE-AUDIO CONTRASTIVE LOSS

Technical Report

Hae Chun Chung

KT Corporation
Acoustic Processing Project, AI Tech Lab
Seoul, Republic of Korea
hc.chung@kt.com

Jae Hoon Jung

KT Corporation
Acoustic Processing Project, AI Tech Lab
Seoul, Republic of Korea
hoony.jung@kt.com

ABSTRACT

This technical report outlines the efforts of KT Corporation's Acoustic Processing Project for addressing language-queried audio source separation (LASS), DCASE 2024 Challenge Task 9. The objective of this work is to separate arbitrary sound sources using a text description of the desired source. We propose three systems, each with the same model architecture but different training methods. These systems use the FLAN-T5 model as the text encoder and the ResUNet model as the separator. To train these systems, we introduced three loss functions: L1 loss in the time domain, multi-scale mel-spectrogram loss in the frequency domain, and contrastive loss, with a loss balancer to stabilize the training. Utilizing the Contrastive Language-Audio Pre-training (CLAP) model, we designed three contrastive losses: audio-to-text (A2T-CL), audio-to-audio (A2A-CL), and audio-to-multi (A2M-CL). The first system was trained with A2T-CL, the second with both A2A-CL and A2T-CL, and the third with A2M-CL. These systems achieved signal-to-distortion ratio (SDR) of 7.030, 7.124, and 7.139, respectively, showing nearly a 30% improvement over the baseline SDR of 5.708 provided by the challenge.

Index Terms— Source Separation,

1. INTRODUCTION

In real-world scenarios, unintended and uncontrollable events frequently occur. During on-location content creation, numerous factors are managed to capture the desired material. Nevertheless, unexpected elements often appear in the final output, making the pursuit of perfection both costly and challenging. If an AI system could separate the desired result from an imperfect one, these costs could be significantly reduced. However, this is quite difficult. This challenge is especially pronounced when the target is audio. Consequently, research in this field is limited, and existing performance levels are suboptimal. [1, 2] DCASE 2024 Challenge Task 9: Language-Queried Audio Source Separation (LASS) [3] targets this issue. This task focuses on developing the system that separates the desired sound from a source with extraneous elements, based on a text description about the intended audio. This report details the models and methodologies employed to tackle this task.

Our system comprises two models: a text encoder and a separator. For the text encoder, we use FLAN-T5 [4], an enhanced version of the text-to-text transfer transformer (T5) model [5]. The separator is a ResUNet model [6, 7] that takes an audio waveform (a mixture of a target audio clip and a noise audio clip) and a text embedding as inputs, producing a separated audio waveform conditioned on the text embedding. To train this system, we introduced three loss functions, and utilized a loss balancer [8] to stabilize the training. First, L1 loss was employed to align the separated audio waveform with the target audio waveform in the time domain. Second, to optimize performance in both the time and frequency domains, we utilized multi-scale mel-spectrogram loss [9, 10, 8, 11], applied across multiple time scales in the mel-spectrogram. Lastly, contrastive loss was introduced in addition to L1 loss and spectrogram loss.

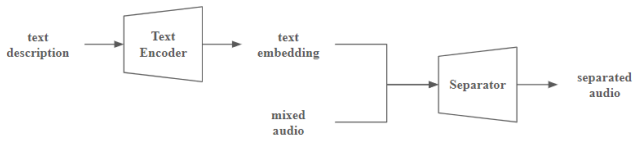

Figure 1: Overall system.

We propose three systems trained with the same model but utilizing different contrastive losses. To implement contrastive loss, we embedded audio and text using a pre-trained Contrastive Language-Audio Pre-training (CLAP) model [12]. We designed three distinct contrastive losses using target audio, noise audio, target text, and noise text for output audio. The first system was trained with audio-to-text contrastive loss (A2T-CL). The second system was trained using audio-to-audio contrastive loss (A2A-CL) and A2T-CL. The third system integrated A2A-CL and A2T-CL into a single audio-to-multi contrastive loss (A2M-CL). These three systems achieved signal-to-distortion ratio (SDR) of 7.030, 7.124, and 7.139, respectively, showing nearly a 30% improvement over the baseline model's SDR of 5.708 provided in the challenge.

2. METHODS

2.1. Overview

Our system consists of two models: a text encoder and a separator. For the text encoder, we utilize FLAN-T5 [4], an enhanced version of the text-to-text transfer transformer (T5) model [5]. FLAN-T5 is initialized with a T5 checkpoint and fine-tuned with instructions

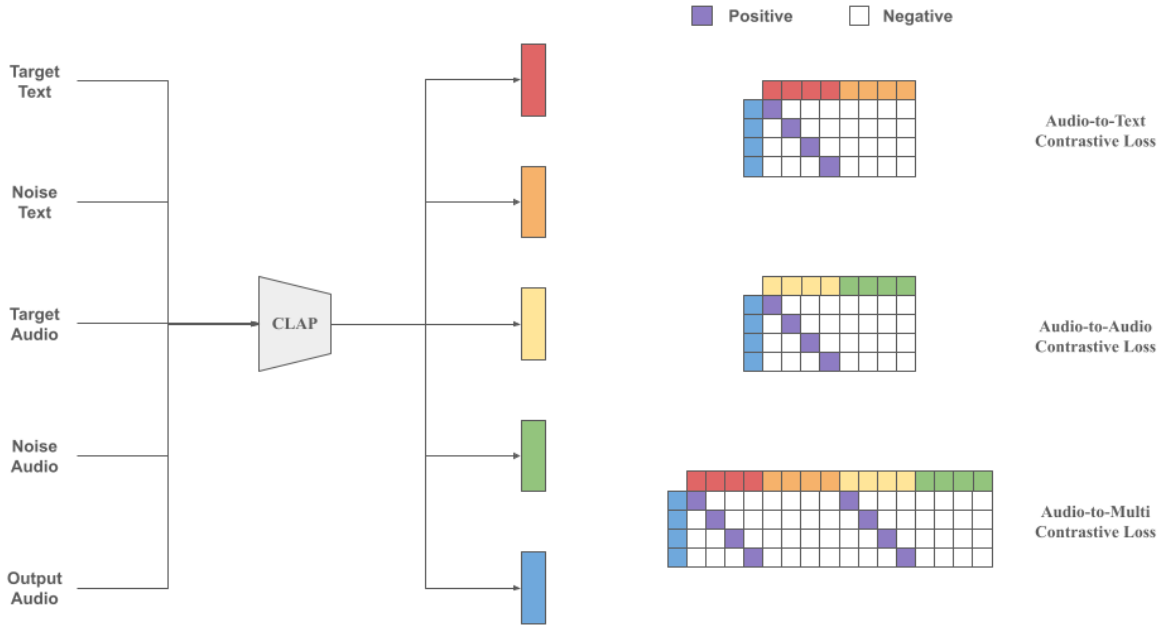

Figure 2: 3 types of contrast loss.

and chain-of-thought reasoning, enabling it to extract robust text embeddings from text descriptions with its strong text representation capacity.

The separator is the ResUNet model [6, 7], an advanced version of the UNet model. The ResUNet model takes a mixed audio waveform and text embedding as input and separates the audio waveform to match the text from the mixed audio. The process begins with applying a short-time Fourier transform (STFT) to the waveform to extract the complex spectrogram, magnitude spectrogram, and phase. The ResUNet model inputs the complex spectrogram and outputs the magnitude mask and phase residual conditioned on the text embedding. The separated complex spectrogram is obtained by multiplying the STFT of the mixture with the predicted magnitude mask and phase residual. Finally, the separated complex spectrogram is converted back into an audio waveform using the inverse short-time Fourier transform (iSTFT).

2.2. Training Loss Terms

From the audio-text paired data, N target pairs (target audio d^{ta} and target text d^{tt}) and N noise pairs (noise audio d^{nt} and noise text d^{nt}) are randomly sampled. For audio, two audio waveforms are combined to create a mixed audio waveform d^{ma} with a signal-to-noise ratio (SNR) ranging from -15 to 15 dB. The target text is input into the text encoder to extract the text embedding. The separator then receives the mixed audio waveform and the text embedding, separating the output audio waveform d^{oa} conditioned on the text from the mixture.

L1 Loss In the source separation task, it is crucial to extract the desired target sound source from a given mixture without altering its original characteristics. In other words, the closer the separated sound source is to the target sound source, the better the performance. To achieve this, minimizing the L1 distance between the target and separated audio over the time domain is commonly used due to its simplicity and effectiveness in universal source separation tasks. We also applied this approach. The equation is as follows:

$$\mathcal{L}_{time} = \left\| d^{ta} - d^{oa} \right\|_1 \tag{1}$$

Spectrogram Loss To optimize performance in both the time and frequency domains, we also employed a multi-scale mel-spectrogram loss [9, 10, 8, 11] applied across multi time scales in the mel-spectrogram. This loss is calculated based on the distance in the mel-spectrogram, which is derived from the short-time Fourier transform (STFT) and converted to a mel scale that better captures human auditory characteristics. This approach enhances the perceptual quality of the output. Additionally, using loss functions on mel-spectrograms across multiple STFT scales enables the model to effectively capture the time-frequency distribution, significantly enhancing its overall performance.

$$\mathcal{L}_{freq} = \frac{1}{|\alpha| + |s|} \sum_{\alpha_i \in \alpha} \sum_{i \in e} \left\| \mathcal{S}_i(d^{ta}) - \mathcal{S}_i(d^{oa}) \right\|_1 + \alpha_i \left\| \log \mathcal{S}_i(d^{ta}) - \log \mathcal{S}_i(d^{oa}) \right\|_2 \tag{2}$$

where $\mathcal{S}_i$ is a 64-bins mel-spectrogram using a normalized STFT with window size of 2^i and hop length of 2^{i-1}, $e = 6, ..., 12$ is the set of scales, and α represents the set of scalar coefficients balancing between the L1 and L2 terms, $\alpha_i = \sqrt{2^{i-1}}$. Here, $|\alpha|$ denotes the sum of the elements of the α set, and $|s|$ is the number of scales.

Audio-to-Text Contrastive Loss The output audio from text-conditioned source separation should match both the target audio and the target text. To achieve this, we implemented an audio-to-text contrastive loss (A2T-CL) using the contrastive language-audio pre-training (CLAP) model [12]. CLAP was trained to align audio

and text by projecting them into a shared feature space. Firstly, we designed the loss so that the output audio attracts its corresponding target text as positive and repels other target texts within the mini-batch as negative in the shared feature space of CLAP model. Contrastive learning becomes more effective as the number of negatives increases. To leverage this, we additionally use noisy texts within the mini-batch as negative examples. This approach encourages the output audio to be distinguishable from various other texts while accurately fitting the target text. The equation is as follows:

$$\mathcal{L}_{a2t} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(f_i^{oa}\cdot f_i^{tt}/\tau)}{\sum_{k=1}^{N}\left\{\exp(f_i^{oa}\cdot f_k^{tt}/\tau)+\exp(f_i^{oa}\cdot f_k^{nt}/\tau)\right\}} \tag{3}$$

where f^{oa} is a feature with output audio embedded using audio encoder of CLAP model, and f^{tt} and f^{nt} are features with target text and noise text embedded using text encoder of CLAP model. And τ is a scalar temperature parameter.

Audio-to-Audio Contrastive Loss In addition, since there are both target audios and noisy audios, the output audio can be matched to target audios and noise audios. Therefore, it is possible to design an audio-to-audio contrastive loss using these.

$$\mathcal{L}_{a2a} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(f_i^{oa}\cdot f_i^{ta}/\tau)}{\sum_{k=1}^{N}\left\{\exp(f_i^{oa}\cdot f_k^{ta}/\tau)+\exp(f_i^{oa}\cdot f_k^{na}/\tau)\right\}} \tag{4}$$

where f^{ta} and f^{na} are features with target audio and noise audio embedded using audio encoder of CLAP model.

Audio-to-Multi Contrastive Loss As aforementioned, contrastive learning shows better performance as the number of negatives increases. To take advantage of this, we integrated audio-to-text contrastive loss and audio-to-audio contrastive loss into a single expanded loss: audio-to-multi contrastive loss, effectively doubling the number of negatives. This causes the output audio to pull closer to the corresponding target text and target audio while pushing away from all remaining target texts, noise texts, target audios, and noise audios within the mini-batch. As a result, the output audio maximizes its similarity to both the target text and target audio.

$$a2t_i = \sum_{k=1}^{N}\left\{\exp(f_i^{oa}\cdot f_k^{tt}/\tau)+\exp(f_i^{oa}\cdot f_k^{nt}/\tau)\right\} \tag{5}$$

$$a2a_i = \sum_{k=1}^{N}\left\{\exp(f_i^{oa}\cdot f_k^{ta}/\tau)+\exp(f_i^{oa}\cdot f_k^{na}/\tau)\right\} \tag{6}$$

$$\mathcal{L}_{a2m} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left\{\log\frac{\exp(f_i^{oa}\cdot f_i^{tt}/\tau)}{a2t_i+a2a_i}+\log\frac{\exp(f_i^{oa}\cdot f_i^{ta}/\tau)}{a2t_i+a2a_i}\right\} \tag{7}$$

Loss Balancer Encodec [8] introduced a loss balancer to stabilize the training by adjusting the loss weights based on various scales of gradients from the model. We used a loss balancer to stabilize the model training with various losses. The gradient $\frac{\partial l_i}{\partial d^{oa}}$ of the loss based on the output d^{oa} is recalculated using the following equation, incorporating the weights λ_i for the loss and reference norm R.

$$\tilde{g}_i = R\frac{\lambda_i}{\sum_j \lambda_j}\cdot\frac{g_i}{\left\langle\|g_i\|_2\right\rangle_\beta} \tag{8}$$

where $\left\langle\|g_i\|_2\right\rangle_\beta$ is the exponential moving average of g_i. We take $R = 1$ and $\beta = 0.999$. All the model losses fit into the balancer. The model is then backpropagated to $\sum_i \tilde{g}_i$ instead of the original $\sum_i \lambda_i g_i$.

2.3. Proposed Systems

We propose a total of three systems. The process by which data is preprocessed and fed forward to the model in all systems is the same as mentioned in Section 2.1. The primary difference between each system lies in the configuration of losses during the training process, particularly the type of contrastive loss. The configuration of the losses for each system in our training was defined as follows. All weights λ for the losses are set 1.

$$System_1 = \lambda_1\mathcal{L}_{time} + \lambda_2\mathcal{L}_{freq} + \lambda_3\mathcal{L}_{a2t} \tag{9}$$

$$System_2 = \lambda_1\mathcal{L}_{time} + \lambda_2\mathcal{L}_{freq} + \lambda_3\mathcal{L}_{a2t} + \lambda_4\mathcal{L}_{a2a} \tag{10}$$

$$System_3 = \lambda_1\mathcal{L}_{time} + \lambda_2\mathcal{L}_{freq} + \lambda_3\mathcal{L}_{a2m} \tag{11}$$

3. SETTING

3.1. Training Data

A total of four datasets were used for model training: AudioCaps [13], WavCaps [14], Clotho v2 [15], and FSD50K [16]. For the WavCaps dataset, only data belonging to AudioSet were used. The combined dataset comprises a total of 216,398 audio clips, amounting to approximately 580 hours. The following procedure was employed to generate mixed audio:

1. Random Selection: Target and noise audio clips were randomly selected to ensure no overlap within the entire dataset.
2. Mono Conversion: If an audio clip had 2 channels, the average of the two channels was calculated to convert it into a mono clip.
3. Resampling: Audio clips with a sampling rate different from 16 kHz were resampled to 16 kHz.
4. Length Adjustment: If an audio clip exceeded 10 seconds in length, it was randomly truncated to 10 seconds. If it was shorter than 10 seconds, zero padding was added to the end to make it 10 seconds long.
5. Mixing: The pre-processed target audio clip and a noise audio clip were mixed with signal-to-noise ratios (SNR) ranging from -15 dB to 15 dB to produce a mixed audio clip.

3.2. Model

The text encoder for embedding the text is used pre-trained FLAN-T5 model [4], and all parameters were frozen. AdamW optimizer [17] with a learning rate of 0.0003 is used for training the separator with the batch size of 25. τ was all set to 0.1 for the contrastive loss.

3.3. Test Data

To evaluate the performance of the model, validation (synth) dataset provided in DCASE2024 Challenge Task9 [3] was used.

3.4. Metric

We evaluate the performance of the source separation system using the signal-to-distance ratio (SDR).

4. RESULTS

	Baseline	System1	System2	System3
SDR	5.708	7.030	7.124	7.139

Table 1: The comparison of SDR score.

While the baseline provided for the challenge achieved a signal-to-distortion ratio (SDR) score of 5.708, our systems achieved SDR scores of 7.030, 7.124, and 7.139, respectively. This represents a remarkable performance improvement of over 30% compared to the baseline. In language-queried audio source separation (LASS), it is crucial to precisely match the output audio to the target audio. Additionally, we demonstrate that aligning the output audio more closely with both the target text and target audio in the feature space using contrastive learning enhances performance. Contrastive learning is more effective the more negatives are. We also show the effectiveness of the audio-to-multi contrastive loss, which leverages the characteristics of contrastive learning by integrating audio-to-text and audio-to-audio contrastive losses. This approach leverages the advantage of having more negatives, significantly improving the model's effectiveness.

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," *arXiv preprint arXiv:2203.15147*, 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] https://dcase.community/challenge2024/task-language-queried-audio-source-separation.

[4] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[5] 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. 140, pp. 1–67, 2020.

[6] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, "Decoupling magnitude and phase estimation with deep resunet for music source separation," *arXiv preprint arXiv:2109.05418*, 2021.

[7] Q. Kong, K. Chen, H. Liu, X. Du, T. Berg-Kirkpatrick, S. Dubnov, and M. D. Plumbley, "Universal source separation with weakly labelled data," *arXiv preprint arXiv:2305.07447*, 2023.

[8] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[9] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[10] A. Gritsenko, T. Salimans, R. van den Berg, J. Snoek, and N. Kalchbrenner, "A spectral energy distance for parallel speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 062–13 072, 2020.

[11] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "Audiogen: Textually guided audio generation," *arXiv preprint arXiv:2209.15352*, 2022.

[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 *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[13] 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.

[14] 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 preprint arXiv:2303.17395*, 2023.

[15] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.

[16] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[17] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.