DCASE 2026 TASK 6: AUDIO MOMENT RETRIEVAL USING BACK-TRANSLATION AND TIME MASKING FOR DATA AUGMENTATION

Technical Report

Jun-Ting Lu, Chung-Che Wang, Syu-Siang Wang

Department of Electrical Engineering (Program A), Yuan Ze University, Taiwan
s1154601@mail.yzu.edu.tw, geniusturtle6174@gmail.com, sypdbhee@saturn.yzu.edu.tw

ABSTRACT

This technical report describes our methods for Task 6 of the DCASE 2026 challenge: Audio Moment Retrieval from Long Audio. In this work, we build upon the official baseline without modifying its network architecture, and focus on data augmentation to improve the generalization ability of the model. Specifically, we investigate two augmentation methods: back-translation-based paraphrasing of the text queries using the OPUS-MT models, and a time mask applied to the CLAP embedding sequence on the audio side. Both augmentation methods improve the performance over the baseline, and the best configuration applies back-translation in both the pretraining and fine-tuning stages.

***Index Terms*—** Audio moment retrieval, data augmentation, back-translation, time mask

1. INTRODUCTION

The goal of DCASE 2026 Task 6 is audio moment retrieval, where, given a long audio recording and a free-format text query, the system is required to identify the temporal segments in the audio whose semantics match the query, and to output their start and end timestamps. The main challenge of this task lies in capturing the temporal context within long audio recordings and in establishing cross-modal alignment between audio and text. The official baseline of this task [1, 2] adopts a bi-encoder structure: MS-CLAP 2023 [3, 4] with a sliding window is used to convert the audio and text into sequential embeddings, which are then fed into a DETR-based [5] network that directly predicts multiple pairs of start and end timestamps. In this challenge, we keep the network architecture of the baseline unchanged, and instead focus on improving the generalization ability of the model through data augmentation. We investigate two augmentation methods: back-translation-based paraphrasing of the text queries, and applying a time mask to the CLAP embedding sequence on the audio side.

The rest of this report is organized as follows. Section 2 describes details of our approaches and submissions. Section 3 shows experimental results.

2. OUR APPROACHES AND SUBMISSIONS

All of our systems are built upon the official baseline provided by this task, without modifying its network architecture. Instead, we focus on data augmentation to improve the generalization ability of the model under the limited size of the development set. The remainder of this section describes the two data augmentation methods, as well as how they are combined in each of our submitted systems.

For training data, we use the development-training split as defined by the task organizers, namely the training splits of CASTELLA [6] and Clotho-Moment [1]; the validation and testing splits are used for model selection and for reporting internal results, respectively. Following the setup of the official baseline, we also adopt a two-stage training procedure in which the model is first pretrained on Clotho-Moment and then fine-tuned on CASTELLA.

2.1. Back-translation for text queries

To increase the semantic diversity of the text queries, we apply back-translation augmentation to the English captions in the training set: each original English caption is first translated into a pivot language and then translated back into English, producing a paraphrased version that is semantically close to but lexically different from the original caption, and is treated as a training sample equivalent to the original. We experimented with several different pivot languages; the detailed comparison and the final choice are described in Section 3.

The LLMs used for translation are the OPUS-MT models released by Helsinki-NLP [7, 8], whose weights are publicly available. We load and run inference with these models locally through the `MarianMTModel` and `MarianTokenizer` classes in the HuggingFace `transformers` library, without calling any remote API, which complies with the task rule that allows LLMs that "can run in a local environment, and have publicly available weights." For each translation direction, a corresponding unidirectional model is used (for example, en $\rightarrow$ de and de $\rightarrow$ en are two separate models).

The back-translated captions share the same timestamps as the original captions, since paraphrasing does not alter the temporal structure of the audio content. Rather than enlarging the training set, we keep the number of training samples unchanged: at each epoch, for the text query of each training sample, one version is randomly selected from the original caption and all of its back-translated versions and used as the input for that epoch.

2.2. Time mask on CLAP embeddings

On the audio side, inspired by the idea of SpecAugment [9], we apply masking to the audio input during training to enhance the robustness of the model against the loss of local information. Since the audio input used by the baseline is neither the raw waveform nor a mel spectrogram, but a sequence of embeddings pre-extracted

Model	R1@0.5	R1@0.7	mAP (avg)	mAP@0.5	mAP@0.75
Baseline (Clotho-Moment + CASTELLA)	27.91	16.78	12.87	25.08	11.42
System 1	37.78	26.99	20.76	34.87	19.86
System 2	36.93	25.28	19.56	33.66	18.76

Table 1: The evaluation results on the development-testing split (CASTELLA test split) for our submitted systems. Higher is better for all metrics.

by MS-CLAP 2023 with a sliding window, we apply the mask directly on this CLAP embedding sequence; furthermore, the mask is applied only along the temporal axis (hereafter referred to as the time mask), that is, within a randomly selected continuous temporal segment, all dimensions of every frame in that segment are replaced by zero.

The length of the time mask ranges from 1 to 10% of the number of CLAP frames; for the 60-second audio recordings commonly used in this study, this corresponds to 1 to 6 CLAP frames (approximately 1 to 6 seconds). At most one mask is applied to each training sample, and the probability of applying it is 0.15. Considering that the annotations of this task correspond to specific temporal segments within the audio, overly aggressive temporal masking may harm the learning of the ground-truth moments, and therefore the above parameters are deliberately set to be conservative.

2.3. Submitted systems

We submit two systems in total. Both adopt the time mask described in Section 2.2, and they differ only in which stage of the two-stage training procedure the back-translation text augmentation described in Section 2.1 is applied to:

- **System 1**: back-translation augmentation is applied in both the Clotho-Moment pretraining stage and the CASTELLA fine-tuning stage.
- **System 2**: back-translation augmentation is applied only in the Clotho-Moment pretraining stage, while the fine-tuning stage uses only the original CASTELLA captions.

Apart from the above difference, the network architecture, other hyperparameters, and the training procedure of all systems are identical to those of the official baseline.

3. EXPERIMENTAL RESULTS

We evaluate all systems on the development-testing split, i.e., the test split of CASTELLA. The primary evaluation metric is Recall1@0.7 as specified by this task; we also report Recall1@0.5, mAP (avg), mAP@0.5, and mAP@0.75 to provide a comprehensive comparison.

Regarding the training setup, all systems use the AdamW optimizer with a learning rate of 1×10^{-4}. The batch size is 32 in the pretraining stage and 16 in the fine-tuning stage, and training is conducted on an NVIDIA GeForce RTX 4070. The pretraining stage runs for 100 epochs and the fine-tuning stage runs for 200 epochs. All other hyperparameters follow the default values of the official baseline.

The main evaluation results are listed in Table 1. The baseline reported here is the "Clotho-Moment pretrain + CASTELLA fine-tune" configuration retrained in our own environment without any data augmentation, rather than the numbers directly cited from the official task webpage. The results show that both System 1 and System 2 outperform the baseline in terms of Recall1@0.7, and System 1 outperforms System 2, suggesting that applying back-translation augmentation in both the pretraining and fine-tuning stages is more beneficial to the final performance than applying it only in the pretraining stage.

We also experimented with pivot languages other than German for back-translation, including French and Spanish. However, in a preliminary comparison on the development-validation split, the configuration using German as the pivot language achieved the best performance, while the augmentation effects of the other pivot languages were inferior to that of German. Therefore, both System 1 and System 2 in our final submissions use back-translation with German as the pivot language. We also attempted to re-synthesize the Clotho-Moment dataset on our own, but were unable to complete the experiment due to resource limitations.

4. REFERENCES

[1] H. Munakata, T. Nishimura, S. Nakada, and T. Komatsu, "Language-based audio moment retrieval," in *Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2025, pp. 1–5.

[2] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in *Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2024, pp. 336–340.

[3] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[4] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," 2023. [Online]. Available: https://arxiv.org/abs/2309.05767

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[6] H. Munakata, T. Imamura, T. Nishimura, and T. Komatsu, "CASTELLA: Long audio dataset with captions and temporal boundaries," 2026.

[7] J. Tiedemann, M. Aulamo, D. Bakshandaeva, M. Boggia, S.-A. Grönroos, T. Nieminen, A. Raganato Y. Scherrer, R. Vazquez, and S. Virpioja, "Democratizing neural machine translation with OPUS-MT," *Language Resources and Evaluation*, no. 58, pp. 713–755, 2023.

[8] J. Tiedemann and S. Thottingal, "OPUS-MT — Building open translation services for the World," in *Proceedings of the 22nd*

Annual Conferenec of the European Association for Machine Translation (EAMT), Lisbon, Portugal, 2020.

[9] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.