ENHANCING LANGUAGE-BASED AUDIO RETRIEVAL WITH PARTIAL FINE-TUNING AND ATTENTION POOLING

Technical Report

Filomeno G. *Kandah Y.* *Spiessberger F.*

students at Johannes Kepler University
Altenberger Str. 69, 4040 Linz, Austria
{k12315325, k12204692, l12211061}@students.jku.at

ABSTRACT

This technical report describes our submission to the language-based audio retrieval task of the DCASE 2025 Challenge (Task 6). Building upon our previous work, we retain the dual-encoder architecture that projects audio recordings and textual descriptions into a shared embedding space. This year we focus on architectural and training-level refinements within a single model framework. Specifically, we fine-tune only the upper transformer layers of a PaSST audio encoder, apply attention-based segment pooling, and replace CLS token extraction in RoBERTa with masked mean pooling. Additionally, we introduce time-frequency spectrogram augmentation and reduce the hop size to capture more segment detail. Our improved system achieves a mAP@10 of 36.005 on the ClothoV2 test set, outperforming the official DCASE 2025 baseline without relying on external caption generation or model ensembles. The result for mAP@16 as requested this year is 36.661 (without new annotations). All code and trained models are available on GitHub[1].

***Index Terms*—** Audio-text retrieval, dual encoder, PaSST, attention pooling, fine-tuning, ClothoV2

1. INTRODUCTION

Task 6 of the DCASE 2025 Challenge [1] focuses on language-based audio retrieval, where the goal is to retrieve audio recordings from a database given a natural language description. This retrieval setting is appealing because it enables users to flexibly query for arbitrary acoustic phenomena (such as sound events, background textures, or combinations thereof) without being constrained to a predefined taxonomy of sound labels.

From a technical perspective, however, this task is challenging as it requires bridging raw audio and natural language representations. Most state-of-the-art systems adopt a *dual encoder* approach [2, 3], where audio and text inputs are independently encoded into a shared embedding space, and similarity between modalities is computed via a dot-product or cosine similarity.

The best performing system last year [4] focused on improving retrieval by applying knowledge distillation from an ensemble of pretrained models to produce soft targets for contrastive learning. While this proved effective, our 2025 system builds upon that baseline and explores a complementary direction: refining architectural and training components within a single model setup.

Specifically, we introduce the following key improvements:

- **Fine-tuning of PaSST:** Instead of using the pretrained PaSST model as a frozen feature extractor, we enable end-to-end training of the higher transformer layers (8–11), while keeping the lower layers frozen for stability and generalization.
- **Segment attention pooling:** Rather than averaging segment embeddings, we apply a learnable attention mechanism to aggregate information over overlapping temporal segments.
- **Improved audio preprocessing:** We introduce spectrogram-level augmentation via time and frequency masking, and use a smaller hop size during segmentation to better capture acoustic dynamics.
- **Text encoder enhancements:** Instead of relying on the `[CLS]` token from RoBERTa, we apply masked mean pooling over the entire sequence, leveraging the full contextual representation of the caption.
- **Training and optimization tweaks:** We include moderate weight decay, increase matmul throughput via `torch.set_float32_matmul_precision("medium")`, and enable distributed training robustness using `ddp_find_unused_parameters_true`.

Compared to previous editions, the 2025 task introduces multiple textual annotations for each audio in the public evaluation set, enabling a richer evaluation via multiple query formulations. In our experiments, we do not explicitly leverage these extra annotations, and instead report results using only the original caption per audio as provided in Clotho-evaluation.

Together, these architectural and training modifications lead to substantial gains in retrieval performance, without requiring synthetic captions or model ensembles. We report results on the ClothoV2 benchmark [5] and discuss each component in detail in the following sections.

2. MOTIVATION

While many captions in datasets like ClothoV2 and WavCaps describe overlapping acoustic scenes, our system does not explicitly model such relationships. However, recognizing this redundancy motivates the use of stronger language and audio encoders capable of generalizing across semantically similar descriptions [2, 4].

In this year's system, we build on top of the 2024 winning submission [4] but shift our focus from knowledge distillation toward improving the underlying model architecture and training pipeline. Our motivation stems from the hypothesis that careful

[1] `https://github.com/FaSchpie/DCASE-Task-6`

control over **fine-tuning**, **segment-level attention**, and **modality-specific pooling strategies** can significantly improve retrieval performance—without the need for external ensembles or synthetic captions.

Furthermore, we conduct systematic experiments with various combinations of audio and text encoders. These experiments reveal that *architecture choice and compatibility between modalities* strongly impact final retrieval accuracy (see Section 3). This suggests that performance gains can be achieved not only through data augmentation or loss engineering, but also through deliberate architectural design.

3. EXPLORING AUDIO AND TEXT ENCODER VARIANTS

We further investigate the impact of encoder selection on retrieval performance by testing combinations of PaSST, CLAP, Whisper, and different sentence encoders (e.g., RoBERTa, MPNet, GTR-T5, BGE). Results are summarized in Table 1.

We observe that PaSST combined with `roberta-large` performs best overall, while models based on GTR-T5 underperform significantly, suggesting a mismatch with the contrastive learning setup or non-optimal training setup.

Table 1: Retrieval performance on the DCASE Task 6 test set for various model configurations.

Model	R@1	R@10	R@5	mAP@10
whisper_all-mpnet-base-v2	0.0513	0.2857	0.1834	0.1108
whisper_gtr-t5-large	0.0010	0.0096	0.0048	0.0028
whisper_bge-base-en-v1.5	0.0524	0.2936	0.1901	0.1124
whisper_deberta-v3-large	0.0500	0.2630	0.1679	0.1016
whisper_roberta-large	0.0620	0.3087	0.2027	0.1232
data2vec_bge-base-en-v1.5	0.0036	0.0337	0.0184	0.0106
clap_all-mpnet-base-v2	0.1234	0.4802	0.3478	0.2204
clap_gtr-t5-large	0.0010	0.0096	0.0048	0.0028
clap_bge-base-en-v1.5	0.1305	0.4804	0.3458	0.2250
clap_roberta-large	0.1298	0.4901	0.3545	0.2266
passt_all-mpnet-base-v2	0.1625	0.5638	0.4153	0.2733
passt_gtr-t5-large	0.0010	0.0096	0.0048	0.0028
passt_bge-base-en-v1.5	0.1732	0.5747	0.4274	0.2828
passt_roberta-large	0.1797	0.5878	0.4467	0.2924

4. PROPOSED METHOD

Our retrieval system follows the standard dual-encoder architecture where audio recordings and text captions are independently embedded into a shared multimodal space using a pair of encoders $\phi_a(\cdot)$ and $\phi_c(\cdot)$. At inference time, the similarity between an audio query a_i and a caption c_j is computed via their cosine similarity:

$$C_{ij} = \frac{\phi_a(a_i)^\top \phi_c(c_j)}{\|\phi_a(a_i)\|_2 \|\phi_c(c_j)\|_2} \tag{1}$$

The model is trained with a contrastive loss based on the temperature-scaled cross-entropy [6]. For each training pair (a_i, c_i), we treat all other examples in the batch as negatives and minimize:

$$\mathcal{L}_{\text{sup}} = \mathcal{H}(p_a, q_a) + \mathcal{H}(p_c, q_c) \tag{2}$$

where $q_a(a_i \mid c_j)$ and $q_c(c_j \mid a_i)$ are softmax distributions over similarity scores, and p_a, p_c are target distributions assuming one positive per caption and audio (i.e., $p_a(a_i \mid c_j) = 1_{i=j}$).

Unlike last year's approach [4], we do not rely on an ensemble for generating soft alignment targets. Instead, our contribution focuses on enhancing the model's architecture and training efficiency to boost performance with a simpler setup.

Audio Encoder. We use PaSST [7] as our base encoder. Differently from prior work, we fine-tune only the top 4 transformer layers (8–11), freezing the lower layers to retain general-purpose audio features while adapting to the retrieval task. Additionally, we increase segment coverage by reducing the hop size to 5s and apply time and frequency masking on the input mel-spectrograms for regularization.

Segment Pooling. To combine multiple audio segments into a fixed-size embedding, we replace mean pooling with an attention mechanism that learns segment relevance weights. Given a sequence of segment embeddings $\{z_1, ..., z_T\}$, we compute attention weights α_i and use them to form a weighted sum:

$$\text{AttnPool}(z_1, ..., z_T) = \sum_{i=1}^{T} \alpha_i z_i \tag{3}$$

where $\alpha_i = \text{softmax}(W_2 \tanh(W_1 z_i))$.

Text Encoder. We use RoBERTa-large [8] to embed the caption and, instead of using the `[CLS]` token as in earlier works, we apply masked mean pooling over the valid tokens. This makes the embedding more robust to sentence structure and punctuation noise.

Optimization. We optimize the model using AdamW with cosine learning rate decay and light regularization (weight_decay $= 1e{-}4$). We also use `torch.set_float32_matmul_precision("medium")` for faster training and apply the `ddp_find_unused_parameters_true` strategy for stability with partial fine-tuning.

5. RESULTS

We evaluate our system on the official ClothoV2 test set using the standard metrics for retrieval: Recall at rank 1, 5, and 10 (R@1, R@5, R@10) as well as mean Average Precision at rank 10 (mAP@10) and 16 (mAP@16), without the new annotations. We chose not to include the additional annotations in our evaluation to maintain consistency with the official Clotho-evaluation setup and to isolate system performance on non-redundant textual queries. We pre-trained all models on AudioCaps, WavCaps, and ClothoV2. Table 2 summarizes the performance of our model compared to the provided baseline system.

Table 2: Retrieval performance on the ClothoV2 test set.

Model	mAP16	mAP@10	R@1	R@5	R@10
Baseline (DCASE 2025)	-	35.23	23.29	**52.17**	64.78
Ours (PaSST-v2 + attention)	**36.661**	**36.005**	**24.072**	51.78	**65.327**

Our improved model outperforms the baseline on most metrics, particularly in terms of mAP@10 and top-1 and top-10 recall. This suggests that the combination of partial fine-tuning, attention-based segment pooling, and improved preprocessing leads to more precise and consistent retrieval results. The R@5 score remains essentially unchanged, indicating that most gains are concentrated at the top and bottom ranks.

6. ACKNOWLEDGMENT

We would like to thank Shah Nawaz. The computational results presented in this work have been achieved using the Vienna Scientific Cluster (VSC). All authors contributed equally and they are presented alphabetically.

7. REFERENCES

[1] "Language-based audio retrieval, task description website," https://dcase.community/challenge2025/task-language-based-audio-retrieval, 2025, accessed: 2025-06-07.

[2] A. S. Koepke, A. Oncescu, J. a. F. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries: A benchmark study," *IEEE Transactions on Multimedia*, vol. 25, pp. 2675–2685, 2023.

[3] S. Lou, X. Xu, M. Wu, and K. Yu, "Audio-text retrieval in context," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022.

[4] P. Primus and G. Widmer, "A knowledge distillation approach to improving language-based audio retrieval models," in *DCASE2024 Workshop on Detection and Classification of Acoustic Scenes and Events*, Vienna, Austria, 2024.

[5] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. of the 37th Int. Conf. on Machine Learning (ICML)*, 2020.

[7] K. Koutini, J. Schluter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Proc. Interspeech*, 2022.

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.