

# Text-Based Audio Retrieval by Learning From Similarities Between Audio Captions

Huang Xie , Khazar Khorrami , Okko Räsänen , *Senior Member, IEEE*, and Tuomas Virtanen , *Fellow, IEEE*

**Abstract**—This letter proposes to use similarities of audio captions for estimating audio-caption relevances to be used for training text-based audio retrieval systems. Current audio-caption datasets (e.g., Clotho) contain audio samples paired with annotated captions, but lack relevance information about audio samples and captions beyond the annotated ones. Besides, mainstream approaches (e.g., CLAP) usually treat the annotated pairs as positives and consider all other audio-caption combinations as negatives, assuming a binary relevance between audio samples and captions. To infer the relevance between audio samples and arbitrary captions, we propose a method that computes non-binary audio-caption relevance scores based on the textual similarities of audio captions. We measure textual similarities of audio captions by calculating the cosine similarity of their Sentence-BERT embeddings and then transform these similarities into audio-caption relevance scores using a logistic function, thereby linking audio samples through their annotated captions to all other captions in the dataset. To integrate the computed relevances into training, we employ a listwise ranking objective, where relevance scores are converted into probabilities of ranking audio samples for a given textual query. We show the effectiveness of the proposed method by demonstrating improvements in text-based audio retrieval compared to methods that use binary audio-caption relevances for training.

**Index Terms**—Audio-caption relevance, audio retrieval, listwise ranking, textual similarity.

## I. INTRODUCTION

TEXT-based audio retrieval, aiming at retrieving audio data based on textual queries, has drawn increasing attention in recent years [1], [2], [3]. It has great potential in real-world applications, e.g., search engines and multimedia databases. Recent works have focused on contrastive learning approaches (e.g., CLAP [4]) utilizing large audio-caption datasets (e.g., WavCaps [5], Auto-ACD [6]). For instance, Primus et al. [7] built their system upon large-scale contrastive learning with audio-caption pairs, resulting in enhanced performance in DCASE 2023 Challenge [8].

Contrastive learning [4], [7] operates with positive and negative audio-caption pairs, assuming a binary relevance between audio samples and captions. Audio-caption pairs are considered positive if the caption accurately describes the paired audio sample; otherwise, they are deemed negative. Those approaches

TABLE I  
CAPTION EXAMPLES FROM AUDIOCAPS

Audio Caption	Textual Similarity
A cat is crying and a person is speaking	1.00
A man is talking and a cat crying	0.90
A cat is crying	0.80
A woman speaks then a cat sighs	0.78
People speak to each other, and a cat wails angrily	0.77

Each caption describes an individual audio sample in AudioCaps [9]. The caption “A cat is crying and a person is speaking” is used as the reference for similarity calculation. Textual similarity is measured with the cosine similarity between Sentence-BERT [11] embeddings of captions.

aim to learn audio and caption representations in a shared embedding space, which will allow measuring audio-caption relevances to be used in retrieval. Likewise, current audio-caption datasets [5], [6], [9], [10] comprise pairs of audio samples and their annotated captions, i.e., only positive pairs. To obtain large quantities of negative pairs for contrastive learning, all other audio-caption combinations in the dataset are utilized as negative pairs. It is likely that audio-caption datasets contain semantically similar captions for different audio samples. For instance, Table I showcases multiple captions from AudioCaps [9], each describing an audio sample that contains cat meows and (or) human speech. These captions exhibit notable textual similarities, e.g., high cosine similarities between their Sentence-BERT [11] embeddings. Constructing negative audio-caption pairs from these captions and their respective audio samples may lead to false negatives, thereby potentially hindering the performance of trained systems. Therefore, instead of binary relevances, employing non-binary measures (e.g., graded relevance [12], [13]) or to model partial relevance is essential for accurately portraying the relationship between audio samples and captions.

On the other hand, limited research has been conducted on assessing the relevance between audio samples and captions in current audio-caption datasets. Our previous works [14], [15] graded audio-caption relevances for a limited subset of Clotho [10] via human crowdsourced assessments, which are usually labor-intensive. Recent work [16] estimated audio-caption relevances by aggregating predictions from multiple audio-language models trained with contrastive learning (i.e., positive and negative pairs), at the cost of increased computational complexity.

In this work, we propose a method for computing non-binary audio-caption relevances within audio-caption datasets based on the textual similarities of audio captions, and then utilize the computed relevances to train models for text-based audio retrieval. Specifically, we measure textual similarities of audio captions by calculating the cosine similarity of their

Received 13 October 2024; revised 30 November 2024; accepted 1 December 2024. Date of publication 4 December 2024; date of current version 17 December 2024. The associate editor coordinating the review of this article and approving it for publication was Dr. Ee Leng Tan. (*Corresponding author: Huang Xie.*)

The authors are with the Faculty of Information Technology and Communication Sciences, Tampere University, 33720 Tampere, Finland (e-mail: huang.xie@tuni.fi; khazar.khorrami@tuni.fi; okko.rasanen@tuni.fi; tuomas.virtanen@tuni.fi).

Digital Object Identifier 10.1109/LSP.2024.3511414

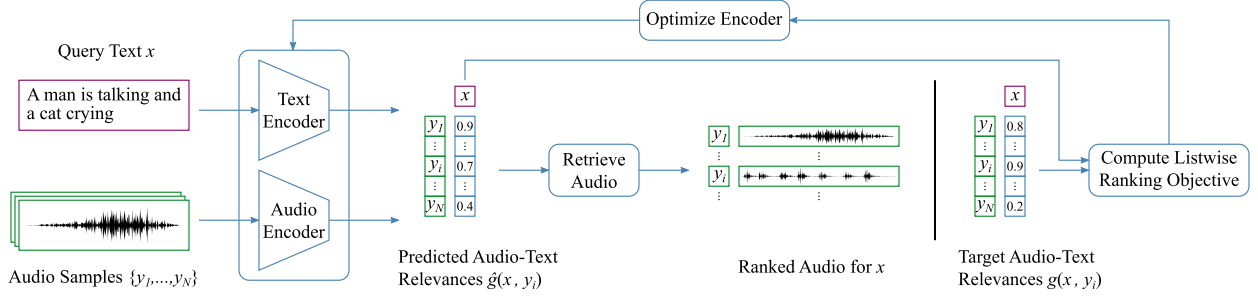


Fig. 1. A model-agnostic dual-encoder framework for text-based audio retrieval. The query text  $x$  and audio samples  $y_i \in \{y_1, \dots, y_N\}$  are projected into a shared embedding space by utilizing a dual-encoder model, and the cosine similarity of their embeddings is utilized as a prediction of their relevance, denoted as  $\hat{g}(x, y_i)$ . Audio retrieval is performed by ranking  $y_1, \dots, y_N$  by their predicted relevance to  $x$ . The dual-encoder model is trained by optimizing a listwise ranking objective, which is computed based on the predicted relevances  $\hat{g}(x, y_i)$  and the target relevances  $g(x, y_i)$ .

Sentence-BERT [11] embeddings, and then transform textual similarities into audio-caption relevance scores using a logistic function. Subsequently, we utilize a listwise ranking objective (i.e., ListNet [17]) to integrate the computed relevances into training. We show the effectiveness of the proposed method by demonstrating improvements in text-based audio retrieval compared to methods that use binary relevances for training. The proposed method can be applied to any captioning dataset (e.g., computing relevances between images and captions in image-caption datasets for visual-textual learning [18], [19]) and, in addition to retrieval, could also be used in other tasks that are based on cross-modal similarity learning.

The remainder of this letter is organized as follows. Section II formalizes the problem of text-based audio retrieval. The proposed method is then presented in Section III, followed by experimental setup in Section IV and corresponding results in Section V. Finally, we conclude this letter in Section VI.

## II. TEXT-BASED AUDIO RETRIEVAL

Text-based audio retrieval refers to retrieving relevant audio samples from a dataset given a textual query. In practice, solving the problem is usually done by estimating a relevance score for each sample to the query and ranking them by relevance. We formalize text-based audio retrieval as follows. Given a query text  $x$  and a set of  $N$  audio samples  $Y = \{y_1, \dots, y_N\}$ , the task is to arrange  $y_i \in Y$  in descending order of their relevance to  $x$ . We define the relevance of  $y_i$  to  $x$  as  $g(x, y_i)$ , and our aim is to develop a model that can accurately predict  $g(x, y_1), \dots, g(x, y_N)$ .

As illustrated in Fig. 1, we employ a model-agnostic dual-encoder framework. A dual-encoder model is trained to learn representations of  $x$  and  $y_i$  in a shared embedding space, where cosine similarity is calculated as the prediction of  $g(x, y_i)$ , denoted as  $\hat{g}(x, y_i)$ . During training, we optimize a listwise ranking objective, which is computed between the target relevances  $g(x, y_i)$  and their predictions  $\hat{g}(x, y_i)$ , for  $i = 1, \dots, N$ . For audio retrieval, we rank  $y_1, \dots, y_N$  by their predicted relevance to  $x$ .

Similar to previous studies [4], [7], we utilize audio-caption datasets for training. Suppose that the training dataset is  $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$ , where  $x_i$  is the caption of  $y_i$ , for  $i = 1, \dots, N$ . During training, each caption  $x_i$  is used as a query, and the dual-encoder model is trained to produce the target relevances  $g(x_i, y_j)$ , for  $j = 1, \dots, N$ .

Previous studies [4], [7] assumed a binary relevance between  $x_i$  and  $y_j$ :  $g(x_i, y_j) = 1$  if  $i = j$  and otherwise  $g(x_i, y_j) = 0$ . However, the training dataset can include multiple audio samples matching with a single caption (and vice versa), i.e.,  $g(x_i, y_j) = 1$  for some  $i \neq j$ . Besides, there can be varying levels of relevance between  $x_i$  and  $y_j$ , ranging from fully relevant, e.g.,  $g(x_i, y_j) = 1$ , to partially relevant, e.g.,  $g(x_i, y_j) = 0.5$ . Therefore, using  $g(x_i, y_j) \in \{0, 1\}$  for training may result in a suboptimal solution. In this work, we compute  $g(x_i, y_j) \in [0, 1]$  based on the textual similarity of  $x_i$  and  $x_j$ , and then integrate it into training by using a listwise ranking objective (Section III-B).

## III. PROPOSED METHOD

In this section, we first introduce the computation of audio-caption relevances based on the textual similarities of audio captions, and then present the listwise ranking objective, which integrates the computed relevances into training.

### A. Computed Audio-Caption Relevance

Current audio-caption datasets (e.g., Clotho [10]) consist of audio samples paired with annotated captions, but lack relevance information about audio samples and captions beyond the annotated ones. To infer this relevance information, we leverage the textual similarity of audio captions.

In practice, audio captions exhibiting high mutual textual similarity can be appropriate descriptions of each other's audio contents. By analyzing the textual similarity between captions, we can determine potential agreement about audio content across audio-caption pairs. This enables us to identify implicit relevance between audio samples and captions from different pairs. For instance, consider the audio-caption pairs  $(x_i, y_i)$  and  $(x_j, y_j)$  with  $i \neq j$ . High textual similarity of  $x_i$  and  $x_j$  would then suggest that both pairs contain similar audio content, implying relevance between  $x_i$  and  $y_j$  (as well as between  $x_j$  and  $y_i$ ). Therefore, we explore computing  $g(x_i, y_j)$  based on the textual similarity of  $x_i$  and  $x_j$ .

With the success of Sentence-BERT in semantic textual similarity tasks [11], we utilize it to gauge the textual similarity between audio captions. Specifically, we compute 768-dimensional caption embeddings and calculate their cosine similarity as a measure of textual similarity. We denote the textual similarity of  $x_i$  and  $x_j$  as  $h(x_i, x_j)$ .

To score the relevance of  $y_j$  to  $x_i$ , we define  $g(x_i, y_j)$  as a function of  $h(x_i, x_j)$ , written as

$$g(x_i, y_j) = f(h(x_i, x_j)), \quad (1)$$

where  $f$  is a monotonically non-decreasing function over the interval  $[-1, 1]$ . In practice, we experimented with linear and logistic functions for  $f$  (Section V-A).

### B. Listwise Ranking Objective

To integrate the computed relevances into training, we employ a listwise ranking objective, i.e., ListNet [17]. Given the computed relevances  $G_i = \{g(x_i, y_j) | y_j \in Y\}$  and the model-predicted relevances  $\hat{G}_i = \{\hat{g}(x_i, y_j) | y_j \in Y\}$  for  $x_i \in D$ , we calculate two probability distributions over  $Y$  and then compute their cross-entropy as the ListNet loss.

For  $G_i$ , we define a probability distribution  $P$ , written as

$$p(y_j | x_i) = \frac{e^{g(x_i, y_j)/\omega}}{\sum_{k=1}^N e^{g(x_i, y_k)/\omega}}, \quad (2)$$

where  $\omega$  is a temperature parameter and  $y_j \in Y$ , for  $j = 1, \dots, N$ . The  $p(y_j | x_i)$  is interpreted as the probability of ranking  $y_j$  highest among  $Y$  for  $x_i$  [15], [17]. Similarly, we compute a probability distribution  $Q$  from  $\hat{G}_i$ , written as

$$q(y_j | x_i) = \frac{e^{\hat{g}(x_i, y_j)/\tau}}{\sum_{k=1}^N e^{\hat{g}(x_i, y_k)/\tau}}, \quad (3)$$

where  $\tau$  is a temperature parameter for  $\hat{G}_i$ . The ListNet loss of  $G_i$  and  $\hat{G}_i$  is calculated as

$$L(P, Q) = - \sum_{j=1}^N p(y_j | x_i) \log q(y_j | x_i). \quad (4)$$

At the training stage, (4) is minimized to ensure that the model can accurately predict  $G_i$  for every  $x_i \in D$ . During audio retrieval, the trained model produces relevance scores for audio samples with respect to a given query text.

## IV. EXPERIMENTS

The aim of the experiments was to test the proposed audio-caption relevance scoring in text-based audio retrieval. For this purpose, we used audio-caption datasets: AudioCaps [9], Clotho [10], and WavCaps [5], with a dual-encoder model from [16].

### A. Audio-Caption Datasets

AudioCaps [9] consists of 51,308 audio samples and 57,188 captions, split into three subsets: a training set with 49,838 audios, a validation set with 495 audios, and a testing set with 975 audios. All audios are drawn from YouTube videos, and their captions are crowdsourced from human annotators. One human-annotated caption is provided for each audio in the training set, and five captions are provided for each in the validation and test sets.

We collected audios of AudioCaps from their original YouTube videos. Due to unavailable YouTube videos, we have 45,522 audios (91.3%) for the training set, 449 audios (90.7%) for the validation set, and 940 audios (96.4%) for the test set.

Clotho [10] comprises 5,929 audio samples, each accompanied by five human-written captions, totaling 29,645 captions. All audios are sourced from the FreeSound [20], and their captions are crowdsourced using a three-step framework [10]. Clotho is partitioned into three subsets: a development set with 3,839 audios, a validation set with 1,045 audios, and an evaluation set with 1,045 audios.

WavCaps [5] is a large-scale, weakly-labeled audio-caption dataset that contains over 400,000 audio samples, each accompanied by a GPT-generated caption. The audio samples are collected from FreeSound [20], BBC Sound Effects [21], SoundBible [22], and AudioSet Strongly-Labeled Subset [23]. We excluded the overlapping audio samples between WavCaps and Clotho, resulting in 401,195 audio-caption pairs.

### B. Dual-Encoder Model

We used audio and text encoders from the best-ranked system [16] in DCASE 2024 Challenge [24]. PaSST [25] was used as the audio encoder, while RoBERTa large [26] served as the text encoder, each followed by two linear layers with a ReLU non-linearity in between, projecting audio samples and captions into a shared embedding space. The dual-encoder model was previously trained using binary audio-caption relevances with contrastive learning (e.g., InfoNCE [27]). In contrast, this work trained it using computed non-binary relevances with listwise ranking.

We followed the training stage one (without a model ensemble) from [16]. We first trained the dual-encoder model on AudioCaps and Clotho as baselines. We divided the training sets into mini-batches of 32 audio-caption pairs and trained the model with an Adam optimizer for 25 epochs. The learning rate was decayed from  $2 \times 10^{-5}$  to  $10^{-7}$  using cosine annealing [28]. We experimented with large-scale pretraining by merging WavCaps with the training sets of AudioCaps and Clotho as a large training set and then fine-tuned the model on AudioCaps and Clotho. For pretraining and fine-tuning, we used the same configuration as for baselines. The temperature parameters  $\omega$  and  $\tau$  were set to 0.05.

### C. Audio-Based Text Retrieval

Besides text-based audio retrieval, we adapted the listwise ranking objective (Section III-B) to audio-based text retrieval (i.e., retrieving captions for audio samples) by swapping  $x$  and  $y$  in (2), (3), and (4). We trained the dual-encoder model for both retrieval scenarios.

### D. Evaluation Metrics

Retrieval performance is evaluated in terms of mean Average Precision at 10 (mAP@10) and Recall at  $k$  (R@ $k$  with  $k \in \{1, 5, 10\}$ ), as done in [16]. The mAP@10 is calculated as the mean of Average Precision (AP) scores across all queries, with AP being the average of precisions at positions where relevant items appear in the ranked list of the top-10 retrieved items for a query. The R@ $k$  is defined as the proportion of relevant items among the top- $k$  items relative to the total relevant items of a query, which is then averaged over all queries. For both metrics, higher values indicate better performance. The evaluation is repeated five times, and the averaged metrics and their standard deviations are reported.

TABLE II  
RETRIEVAL PERFORMANCE ON AUDIOCAPS AND CLOTHO. MEAN AND SD ( $\pm$ ) ARE REPORTED ACROSS FIVE INDEPENDENT RUNS

Dataset	Method	Text-Based Audio Retrieval				Audio-Based Text Retrieval			
		mAP@10	R@1	R@5	R@10	mAP@10	R@1	R@5	R@10
AudioCaps	InfoNCE	54.5 $\pm$ 0.5	39.6 $\pm$ 0.7	75.4 $\pm$ 0.4	86.9 $\pm$ 0.2	37.0 $\pm$ 0.5	9.7 $\pm$ 0.2	38.0 $\pm$ 0.7	54.4 $\pm$ 0.5
	ListNet <sub>audio</sub>	<b>55.0<math>\pm</math>0.3</b>	<b>39.9<math>\pm</math>0.3</b>	<b>76.1<math>\pm</math>0.5</b>	<b>87.6<math>\pm</math>0.4</b>	31.4 $\pm$ 0.1	8.5 $\pm$ 0.2	32.5 $\pm$ 0.3	48.7 $\pm$ 0.2
	ListNet <sub>text</sub>	48.7 $\pm$ 0.3	33.7 $\pm$ 0.5	69.5 $\pm$ 0.2	82.8 $\pm$ 0.4	<b>37.8<math>\pm</math>0.2</b>	<b>10.0<math>\pm</math>0.2</b>	<b>38.6<math>\pm</math>0.3</b>	54.8 $\pm$ 0.4
	ListNet <sub>audio+text</sub>	54.7 $\pm$ 0.1	39.8 $\pm$ 0.2	75.4 $\pm$ 0.4	87.0 $\pm$ 0.3	37.4 $\pm$ 0.4	9.9 $\pm$ 0.1	38.1 $\pm$ 0.4	<b>55.0<math>\pm</math>0.5</b>
Clotho	InfoNCE	28.2 $\pm$ 0.5	16.9 $\pm$ 0.5	43.3 $\pm$ 0.3	57.7 $\pm$ 0.7	15.1 $\pm$ 0.3	4.3 $\pm$ 0.2	17.0 $\pm$ 0.3	26.9 $\pm$ 0.4
	ListNet <sub>audio</sub>	<b>30.4<math>\pm</math>0.1</b>	<b>19.0<math>\pm</math>0.1</b>	<b>45.9<math>\pm</math>0.3</b>	<b>60.1<math>\pm</math>0.5</b>	16.0 $\pm$ 0.2	4.3 $\pm$ 0.1	17.8 $\pm$ 0.3	27.9 $\pm$ 0.2
	ListNet <sub>text</sub>	28.5 $\pm$ 0.4	17.3 $\pm$ 0.4	43.6 $\pm$ 0.3	58.3 $\pm$ 0.2	<b>16.5<math>\pm</math>0.3</b>	<b>4.4<math>\pm</math>0.2</b>	<b>18.7<math>\pm</math>0.2</b>	<b>28.8<math>\pm</math>0.4</b>
	ListNet <sub>audio+text</sub>	29.6 $\pm$ 0.1	18.0 $\pm$ 0.1	45.3 $\pm$ 0.2	59.4 $\pm$ 0.3	16.0 $\pm$ 0.2	<b>4.4<math>\pm</math>0.1</b>	17.9 $\pm$ 0.1	28.0 $\pm$ 0.4

Mean and sd ( $\pm$ ) are reported across five independent runs.

## V. RESULTS AND ANALYSIS

This section presents the experimental results of the proposed method on AudioCaps and Clotho.

### A. The Selection of Function $f$

We experimented with the min-max scaling and logistic functions for  $f$  in (1) and compared their results of text-based audio retrieval on Clotho. The logistic function was written as:

$$g(x_i, y_j) = \frac{1}{1 + e^{2.73 - 4.58 \cdot h(x_i, x_j)}}, \quad (5)$$

which was derived by modeling the crowdsourced audio-caption relevance ratings [14] based on the similarities of audio captions through beta regression with a logit link function. It achieved superior performance (30.4  $\pm$  0.1 in mAP@10, 19.0  $\pm$  0.1 / 45.9  $\pm$  0.3 / 60.1  $\pm$  0.5 in R@{1, 5, 10}) compared to the min-max scaling (29.0  $\pm$  0.6 in mAP@10, 17.7  $\pm$  0.5 / 44.4  $\pm$  0.6 / 58.6  $\pm$  0.3 in R@{1, 5, 10}). Thus, we used the logistic function in the following experiments.

### B. Audio and Text Retrieval

Table II summarizes retrieval performance on AudioCaps and Clotho. The model trained with computed non-binary relevances using (4) is denoted as “ListNet<sub>audio</sub>”, and the adapted version for audio-based text retrieval (Section IV-C) is denoted as “ListNet<sub>text</sub>”. We also experimented with the combination of the two, labeled “ListNet<sub>audio+text</sub>”. The model trained with binary relevances using InfoNCE [27] (the same as [4], [7]) is denoted as “InfoNCE”.

The results demonstrated that the proposed method outperformed the InfoNCE approach in both retrieval scenarios on AudioCaps and Clotho. In particular, ListNet<sub>audio</sub> delivered the best performance in text-based audio retrieval, as it was specifically optimized for ranking audio samples based on text queries. Likewise, ListNet<sub>text</sub>, designed for ranking captions given an audio sample, achieved the highest performance in audio-based text retrieval.

We assessed performance significance in text-based audio retrieval. A paired t-test was conducted on text queries by calculating mAP@10 for each across the five runs. The results showed that ListNet<sub>audio</sub> achieved significant improvements on Clotho ( $t(5224) = 11.658$ ,  $p < 0.001$ ) compared to InfoNCE. On AudioCaps, the improvement was not significant ( $t(4699) = 0.532$ ,  $p > 0.05$ ).

TABLE III  
RETRIEVAL PERFORMANCE ON AUDIOCAPS AND CLOTHO WITH PRETRAINING AND FINE-TUNING

Dataset	Method	Text-Based Audio Retrieval			
		mAP@10	R@1	R@5	R@10
AudioCaps	InfoNCE	55.5 $\pm$ 0.2	40.8 $\pm$ 0.3	76.0 $\pm$ 0.3	87.4 $\pm$ 0.2
	ListNet <sub>audio</sub>	<b>56.5<math>\pm</math>0.4</b>	<b>41.9<math>\pm</math>0.5</b>	<b>76.8<math>\pm</math>0.2</b>	<b>87.9<math>\pm</math>0.4</b>
Clotho	InfoNCE	35.2 $\pm$ 0.5	23.2 $\pm$ 0.5	51.4 $\pm$ 0.6	65.6 $\pm$ 0.3
	ListNet <sub>audio</sub>	<b>36.2<math>\pm</math>0.3</b>	<b>24.1<math>\pm</math>0.5</b>	<b>52.6<math>\pm</math>0.3</b>	<b>66.1<math>\pm</math>0.6</b>

Table II shows that ListNet<sub>audio+text</sub> utilizing computed non-binary relevances outperforms InfoNCE, which uses binary relevances. We note that both losses can be viewed as the sum of two cross-entropy terms and share the same underlying formula when  $P$  has all probability mass on one sample (Section II-B). We conclude that calculating  $P$  based on the non-binary relevances improves performance.

### C. Pretraining and Fine-Tuning

We also validated the effectiveness of the proposed method on large-scale data by pretraining the dual-encoder model on the three datasets, followed by fine-tuning on AudioCaps and Clotho, respectively. Table III summarizes the results of text-based audio retrieval on AudioCaps and Clotho. A paired t-test was conducted on text queries by calculating mAP@10 for each across the five runs. The results showed that ListNet<sub>audio</sub> achieved significant improvements in mAP@10 on both AudioCaps ( $t(4699) = 2.601$ ,  $p = 0.0093$ ) and Clotho ( $t(5224) = 2.438$ ,  $p = 0.0148$ ) compared to InfoNCE.

## VI. CONCLUSIONS

This work proposed a method for computing audio-caption relevances based on the textual similarities of audio captions, and utilized the computed relevances to train models for text-based audio retrieval. We calculated relevance scores for audio samples and captions by transforming textual similarities using a logistic function. We employed a listwise ranking objective to integrate the computed relevances into training. Additionally, we experimented on audio-based text retrieval with the proposed method. Experimental results validated the effectiveness of the proposed method in both text-based audio retrieval and audio-based text retrieval, showcasing improvements over methods that use binary relevances for training.



## REFERENCES

- [1] H. Xie, S. Lipping, and T. Virtanen, "Language-based audio retrieval task in DCASE 2022 challenge," in *Proc. 7th Detection Classification Acoustic Scenes Events 2022 Workshop*, 2022, pp. 216–220.
- [2] P. Primus and G. Widmer, "Improving natural-language-based audio retrieval with transfer learning and audio & text augmentations," in *Proc. 7th Detection Classification Acoustic Scenes Events 2022 Workshop*, 2022, pp. 166–170.
- [3] B. Weck, M. P. Fernández, H. Kirchhoff, and X. Serra, "Matching text and audio embeddings: Exploring transfer-learning strategies for language-based audio retrieval," in *Proc. 7th Detection Classification Acoustic Scenes Events 2022 Workshop*, 2022, pp. 206–210.
- [4] 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. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.
- [5] X. Mei et al., "WavCaps: A ChatGPT-Assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *IEEE/ACM Trans. Audio Speech, Lang. Process.*, vol. 32, pp. 3339–3354, 2024.
- [6] L. Sun, X. Xu, M. Wu, and W. Xie, "Auto-ACD: A large-scale dataset for audio-language representation learning," in *Proc. ACM Int. Conf. Multimedia*, 2024, pp. 5025–5034.
- [7] P. Primus, K. Koutini, and G. Widmer, "Advancing natural-language based audio retrieval with passt and large audio-caption data sets," in *Proc. 8th Detection Classification Acoustic Scenes Events 2023 Workshop*, 2023, pp. 151–155.
- [8] "Language-based audio retrieval in DCASE 2023 challenge," Accessed: Apr. 25, 2024. [Online]. Available: <https://dcase.community/challenge2023/task-language-based-audio-retrieval>
- [9] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 119–132.
- [10] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 736–740.
- [11] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-Networks," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 3982–3992.
- [12] T. Sakai, "Graded Relevance," in *Evaluating Information Retrieval and Access Tasks: NTCIR's Legacy of Research Impact*. Singapore: Springer, 2021, pp. 1–20.
- [13] K. Roitero, E. Maddalena, S. Mizzaro, and F. Scholer, "On the effect of relevance scales in crowdsourcing relevance assessments for information retrieval evaluation," *Inf. Process. Manage.*, vol. 58, no. 6, 2021, Art. no. 102688.
- [14] H. Xie, K. Khorrami, O. Räsänen, and T. Virtanen, "Crowdsourcing and evaluating text-based audio retrieval relevances," in *Proc. 8th Detection Classification Acoustic Scenes Events 2023 Workshop*, 2023, pp. 226–230.
- [15] H. Xie, K. Khorrami, O. Räsänen, and T. Virtanen, "Integrating continuous and binary relevances in audio-text relevance learning," in *Proc. 9th Detection Classification Acoust. Scenes Events Workshop*, 2024, pp. 201–205.
- [16] P. Primus, F. Schmid, and G. Widmer, "Estimated audio-caption correspondences improve language-based audio retrieval," in *Proc. 9th Detection Classification Acoust. Scenes Events Workshop*, 2024, pp. 121–125.
- [17] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 129–136.
- [18] Z. Li, C. Guo, Z. Feng, J.-N. Hwang, Y. Jin, and Y. Zhang, "Image-text retrieval with binary and continuous label supervision," 2022, *arXiv:2210.11319*.
- [19] Z. Li, C. Guo, X. Wang, H. Zhang, and Y. Wang, "Integrating listwise ranking into pairwise-based image-text retrieval," *Knowl.-Based Syst.*, vol. 287, 2024, Art. no. 111431.
- [20] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 411–412.
- [21] "BBC sound Effects," Accessed: Sep. 25, 2024. [Online]. Available: <https://sound-effects.bbcwind.co.uk>
- [22] "SoundBible," Accessed: Sep. 25, 2024. [Online]. Available: <https://soundbible.com>
- [23] S. Hershey et al., "The benefit of temporally-strong labels in audio event classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 366–370.
- [24] "Language-based audio retrieval in DCASE 2024 challenge," Accessed: Sep. 25, 2024. [Online]. Available: <https://dcase.community/challenge2024/task-language-based-audio-retrieval>
- [25] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Proc. 23rd Annu. Conf. Int. Speech Commun. Assoc.*, 2022, pp. 2753–2757.
- [26] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
- [27] A. Van Den, Y. Oord Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.
- [28] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1769–1784.