CROWDSOURCING AND EVALUATING TEXT-BASED AUDIO RETRIEVAL RELEVANCES

Huang Xie, Khazar Khorrami, Okko Räsänen, Tuomas Virtanen

Unit of Computing Sciences, Tampere University, Finland

ABSTRACT

This paper explores grading text-based audio retrieval relevances with crowdsourcing assessments. Given a free-form text (e.g., a caption) as a query, crowdworkers are asked to grade audio clips using numeric scores (between 0 and 100) to indicate their judgements of how much the sound content of an audio clip matches the text, where 0 indicates no content match at all and 100 indicates perfect content match. We integrate the crowdsourced relevances into training and evaluating text-based audio retrieval systems, and evaluate the effect of using them together with binary relevances arise from audio captioning. Conventionally, these binary relevances are defined by captioning-based audio-caption pairs, where being positive indicates that the caption describes the paired audio, and being negative applies to all other pairs. Experimental results indicate that there is no clear benefit from incorporating crowdsourced relevances alongside binary relevances when the crowdsourced relevances are binarized for contrastive learning. Conversely, the results suggest that using only binary relevances defined by captioning-based audio-caption pairs is sufficient for contrastive learning.

Index Terms— Text-based audio retrieval, non-binary relevance, crowdsourcing assessment

1. INTRODUCTION

Text-based audio retrieval, or text-to-audio retrieval, refers to searching for audio clips with free text queries, which has great potential in real-world applications, such as search engines and multimedia databases. Early works [1, 2] have mainly focused on methods of retrieving audio with carefully curated phrases (e.g., audio tags, onomatopoeic words). With the availability of large audio-caption datasets (e.g., Clotho [3] and AudioCaps [4]) in recent years, increasing attention has been drawn to developing methods for audio retrieval using free-form text [5].

Most of the literature tackles text-to-audio retrieval with cross-modal learning methods. Oncescu *et al.* [6] first established benchmarks in this topic with an adapted text-to-video retrieval model. With the recent success of large-scale pretrained audio models (e.g., PANNs [7]) and language models (e.g., BERT [8]), pretrained models are widely used for text-to-audio retrieval and fine-tuned on task-specific data to learn joint representations of audio and text [5]. Besides, several works [9, 10] explored training strategies for text-to-audio retrieval models. For example, Mei *et al.* [9] evaluated several cross-modal learning objectives (e.g., InfoNCE loss [11]) in the context of text-to-audio retrieval.

In aforementioned works [6, 9, 10], relevances of audio to be retrieved are usually assumed to be binary, i.e., either relevant or irrelevant, given a text query. For example, an audio clip is relevant to a caption if and only if the caption thoroughly describes its sound content (i.e., their content matches perfectly). Practically, due to the lack of annotated non-binary relevances in existing datasets, binary relevances defined by captioning-based audio-caption pairs are adopted for system training and evaluation. Binary relevances are positive for audio-caption pairs where the caption describes the paired audio, and negative for all other pairs. This allows producing large quantities of positive and negative examples for contrastive learning in state-of-the-art systems [5].

As a caption might partially describe the sound content of an audio clip, we explore grading audio-text relevances with non-binary numerical scores. Specifically, we crowdsource audio-text relevances graded on a scale of 0 to 100, where 0 indicates completely irrelevant (i.e., no content match at all) and 100 indicates completely relevant (i.e., perfect content match). The main contributions of this work are: 1) we crowdsource non-binary audio-text relevances for environmental audio and its existing captions; 2) we integrate the crowdsourced relevances into training and evaluating text-to-audio retrieval systems, and evaluate the effect of using them alongside binary relevances defined by captioning-based audio-caption pairs; 3) we release all the data and the crowdsourcing instructions to the research community to allow others to explore non-binary relevances [12].

2. CROWDSOURCING ASSESSMENTS

This section presents the proposed method for crowdsourcing audio-text relevances.

2.1. Crowdsourcing Pipeline

We crowdsource audio-text relevances on Amazon Mechanical Turk (MTurk). Here we first introduce the terms used in this work. A *human intelligence task* (HIT) represents a single task that a crowdworker can work on. An *assignment* is a copy of a HIT that is assigned to a crowdworker. A worker *answer* is the submitted task result when a crowdworker completes an assignment.

Fig. 1 presents an overview of the crowdsourcing pipeline. Audio clips and captions for relevance assessments are combined to form HITs, each of which consists of five audio clips and one caption. Every HIT is assigned to multiple MTurk workers. Raw answers containing graded audio-text relevances are collected and aggregated after workers complete their assignments.
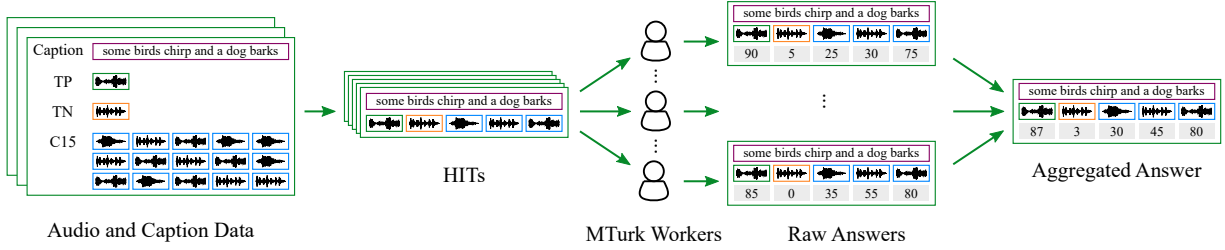
Figure 1: An overview of the pipeline for crowdsourcing relevance assessments.

Split	#Captions	#Audio		
		#TPs	#TNs	#C15s
development	200	200	200	3000
validation	200	200	200	3000
evaluation	200	200	200	3000

Table 1: Statistics of captions and audio clips selected for crowdsourcing relevance assessments.

2.2. Audio and Caption Data

We select a subset of captions and audio clips from each split in Clotho [3], with each subset containing 200 captions and 17 audio clips for each caption. Clotho crowdsources captions for each audio clip, and captions scored high by other workers results to the final captions [3]. The audio clips selected for each caption consists of one true positive clip (TP, being completely relevant to the caption), one true negative clip (TN, being completely irrelevant to the caption), and 15 relevance-unknown candidates (C15). Specifically, we select audio clips corresponding to the captions in Clotho as TPs and obtain TNs using low audio-caption similarity scores estimated by the baseline system in DCASE 2023 Challenge Task 6B[1] followed by human verification. For C15s, we choose the top five clips with high audio-caption similarity scores, together with another 10 randomly selected clips to include audio clips having diverse relevances. Table 1 summarizes the captions and audio clips for crowdsourcing relevance assessments.

2.3. Crowdsourcing Task Setting

For each caption, five HITs are created, each of which is assigned to several crowdworkers. Each HIT contains five audio clips, which are provided for crowdworkers to assess their individual relevance to the caption. The C15s of a caption are split into five batches of three C15s, with one batch per HIT. The TP and TN clips are used for quality check after crowdsourcing.

In each assignment, crowdworkers are asked to assign numeric scores (between 0 and 100) to indicate their judgements of how much the sound content in each audio clip matches the given caption in that assignment. Inspired by [13], we grade audio-text relevances on a scale of 0 to

100, where 0 indicates completely irrelevant (i.e., no content match at all) and 100 indicates completely relevant (i.e., perfect content match). An initial value of 0 is set as the default relevance score for each audio clip. Crowdworkers are required to listen to each audio clip entirely.

2.4. Quality Check

To collect high-quality answers, a quality check is conducted before and after crowdsourcing. Worker requirements are set up for selecting workers with high-quality work on MTurk (e.g., workers with a HIT approve rate greater than 98%). Besides, crowdworkers should pass a predefined qualification test by correctly answering questions about identifying the audio clip described by a given caption from three candidates before they can accept our HITs (i.e., receiving assignments).

With the fact that TPs are clearly more relevant than TNs within the same assignment, the former should receive higher scores than the latter. Let s_{tp} be the graded relevance of a TP, and s_{tn} be the one of the TN within the same assignment. Consistency verification on s_{tp} and s_{tn} is applied to check and select worker answers at the worker level.

Let $D = \{S_i \mid 1 \leq i \leq N\}$ denote the set of N answers submitted by a worker, where $S_i = \{s_{tp}^i, s_{tn}^i, s_{c1}^i, s_{c2}^i, s_{c3}^i\}$ represents the i-th answer from the worker, and s_{c1}, s_{c2}, s_{c3} are the scores of the three C15 clips within the same assignment. As mentioned above, a C15 can be either completely relevant or irrelevant, or even partially relevant to a given caption. For every $S_i \in D$, we measure two random variables for the worker: X, which represents the difference of s_{tp} and s_{tn}, and Y, which denotes the difference of every pair of s_{c1}, s_{c2}, s_{c3}. Intuitively, s_{tp} should be higher than s_{tn} by more than what is expected by chance when sampling from Y. For consistency verification on s_{tp} and s_{tn}, we therefore require that X and Y should satisfy:

$$E(X) = E(s_{tp} - s_{tn}) \geq E(Y) + \sigma(Y), \quad (1)$$

where E represents the expected value and σ represents the standard deviation across D (i.e., all answers from the worker). If the inequality is not satisfied, D will be discarded completely.

3. AUDIO-TEXT RELEVANCE SCORES

This section analyzes the crowdsourced relevances.

---

[1]https://dcase.community/challenge2023/task-language-based-audio-retrieval.

Split	#HITs	#Workers	#Answers
development	1000	109	6651
validation	1000	113	5064
evaluation	1000	118	6489

Table 2: Statistics of crowdsourced data.

3.1. Crowdsourced Raw Scores

Table 2 summarizes information about the crowdsourced data after filtering the data based on the quality check. For each HIT, answers were collected from at least five distinct crowdworkers. In total, 18204 answers were crowdsourced from 340 MTurk workers.

Fig. 2 presents the distribution of raw relevance scores of TP, TN, and C15 clips. For TPs, approximately 60% of relevance scores have a value of 100. For TNs, about 90% of relevance scores are zeros, and over 98% of these scores are less than 20. It indicates that most crowdworkers can appropriately assess the relevances of TPs and TNs to a given caption. For C15s, over 10% of relevance scores have a value of 100, which indicates that some C15s are highly relevant to a given caption. We notice that around 20% of relevance scores of TPs are zeros, which shows the necessity of further processing on the crowdsourced raw scores.

3.2. Aggregated Scores

The raw scores from different workers regarding the relevance of an audio clip to a text query are aggregated by discarding a maximum and a minimum score and then averaging the remaining to produce a statistic that is robust to outliers. Fig. 3 presents the distribution of aggregated relevance scores of TP, TN, and C15 clips. After aggregating, the distribution of relevance scores becomes more balanced, with fewer instances of extreme or polarized judgements (e.g., scores of 0 and 100). Particularly, TPs exhibit a broader spectrum of relevances (e.g., having scores spanning from 30 to 100) compared to other clips. Over 99% of TPs have a score above 10, while about 99% of TNs have a score below 10. Around 30% of C15s have a score above 10, and roughly 10% have a score above the mean score of TPs (i.e., $s > 72$).

4. EXPERIMENTS

This section reports experimental results of using the crowdsourced relevances for text-to-audio retrieval. Due to the lack of established methods for using non-binary relevances for training and evaluation, we binarize the crowdsourced relevances.

4.1. Audio-Caption Pairs

Similar to previous studies [5], we tackle text-to-audio retrieval with contrastive learning. To obtain positive and negative examples for contrastive learning, we binarize the crowdsourced relevances (see 3.2) using the mean score of TPs as an arbitrary threshold. Specifically, we obtain positive audio-caption pairs by combining: 1) a caption with its high--graded C15 clips, which have a score above the threshold;

Split	BiCrRel	BiRel	BiCrRel+BiRel
development	3890	2370	6260
validation	2560	1580	4140
evaluation	2440	1390	3830

Table 3: Number of positive audio-caption pairs in BiCrRel, BiRel, and their combination ("BiCrRel+BiRel").

2) the TP clip of a caption with the captions corresponding to its high-graded C15 clips in Clotho; 3) the siblings (i.e., captions describing the same TP clip in Clotho) of a caption with its high-graded C15 clips. All other audio-caption combinations are treated as negative pairs. The resulting positive and negative pairs are referred to as content-matching pairs with "Binarized Crowdsourced Relevances" (BiCrRel).

As a baseline, we created a subset of Clotho by selecting those audio-caption pairs of which the audio or the caption were part of BiCrRel, i.e., using the captioning-based clip-specific audio-caption pairs from Clotho [3]. The selected pairs are referred to as captioning-based pairs with "Binary Relevances" (BiRel), which include the same audio clips and captions as in BiCrRel. Table 3 summarizes information about BiCrRel, BiRel, and their combination ("BiCrRel+BiRel"). The development / validation / evaluation splits are used for training / validation / evaluation, respectively.

4.2. Retrieval System

We experiment with the retrieval system proposed as the baseline in DCASE 2023 Challenge Task 6B[2], where a pretrained CNN14 [7] is employed as the audio encoder and the Sentence-BERT (i.e., "all-mpnet-base-v2") [14] is used as the text encoder. This system is trained by optimizing the InfoNCE loss [11] such that embeddings of the paired audio and text are pulled together while those of the unpaired are pushed far away.

Audio Encoder. The CNN14 [7], which is pretrained on AudioSet [15], is employed as the audio encoder, with its last linear layer discarded. An extra linear layer is added on the top to generate 300-dimensional audio embeddings. The audio encoder is fine-tuned during training.

Text Encoder. The Sentence-BERT [14], which is derived from BERT [8] for the purpose of generating robust sentence embeddings, is used as the text encoder. An extra linear layer is also added on the top to generate 300-dimensional text embeddings. The Sentence-BERT is frozen during training.

InfoNCE Loss. The InfoNCE loss [11] is a symmetric cross-entropy loss, taking the form of

$$\mathcal{L} = -\frac{1}{M} \sum_{i=1}^{M} [\log \frac{\exp(z_{ii}/\tau)}{\sum_{j=1}^{M} \exp(z_{ij}/\tau)} + \log \frac{\exp(z_{ii}/\tau)}{\sum_{j=1}^{M} \exp(z_{ji}/\tau)}],$$
$$(2)$$

---

[2]https://dcase.community/challenge2023/task-language-based-audio-retrieval.
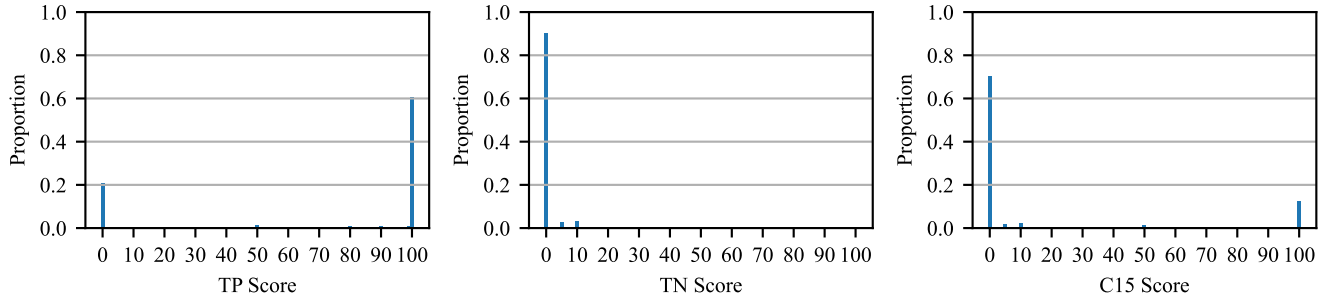
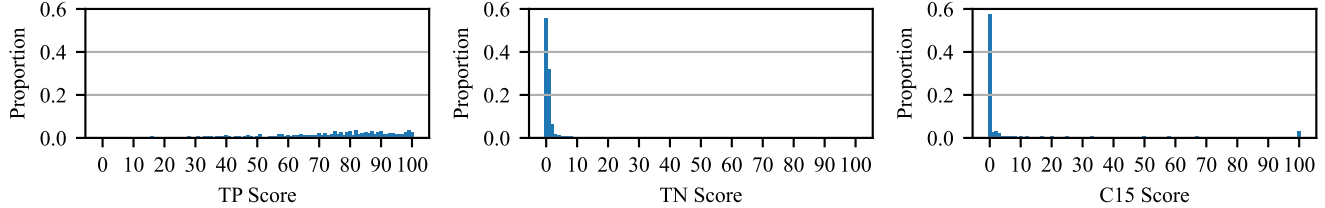Figure 2: Distribution of raw relevance scores of TP, TN, and C15 clips.

Figure 3: Distribution of aggregated relevance scores of TP, TN, and C15 clips.

where τ represents the temperature hyper-parameter, M denotes the number of audio-text pairs, and z_{ij} represents the cosine similarity of the i-th audio embedding and the j-th text embedding. It has been widely used to train cross-modal retrieval systems [5].

Training Setup. The retrieval system is trained with mini-batches consisting of 32 audio-text pairs from a development split. An Adam optimizer with an initial learning rate of 0.001 is adopted to optimize training. Learning rate is reduced by a factor of ten once the validation loss does not improve for five epochs. Training is terminated by early stopping with a patience of ten epochs.

4.3. Evaluation Metrics

Retrieval performance is measured with recall at 10 (R@10) on different evaluation splits. The R@10 is defined as the proportion of relevant items among the top 10 results to all the relevant items contained in the data and is averaged over queries [5]. The more relevant items are within top 10 results, the higher R@10 it is.

4.4. Results

Table 4 shows that training the system on BiRel leads to high R@10 (e.g., 0.566 on BiRel), whereas training the system on BiCrRel yields low R@10 (e.g., 0.478 on BiRel) and on BiCrRel+BiRel results in intermediate R@10 (e.g., 0.509 on BiRel). We conclude that the crowdsourced relevances do not improve the performance from captioning-based audio-caption pairs when they are reduced to binary relevances. A possible explanation is that captions in Clotho are crowdsourced specifically to describe an exact audio clip, while the crowdsourced relevances are graded based on their matching content with a given caption (i.e., different underlying purposes and criteria for generating captions and assessing relevances). Besides, with the fact that each caption in BiRel

Training Data	Evaluation Data		
	BiCrRel	**BiRel**	**BiCrRel+BiRel**
BiCrRel	0.357	0.478	0.407
BiRel	0.412	0.566	0.479
BiCrRel+BiRel	0.363	0.509	0.426

Table 4: Evaluation R@10 of text-to-audio retrieval with the retrieval system trained on different pairs.

is annotated with one relevant audio clip while a caption in BiCrRel can have several relevant audio clips, it makes text-to-audio retrieval on BiCrRel more difficult and leads to a decrease in R@10.

5. CONCLUSIONS

We explore grading audio-text relevance for text-based audio retrieval via crowdsourcing assessments. We crowdsource audio-text relevances graded on a scale of 0 to 100, where 0 indicates completely irrelevant and 100 indicates completely relevant. We integrate crowdsourced relevances into training and evaluating text-to-audio retrieval systems, and evaluate the effect of using them alongside binary relevances defined by captioning-based audio-caption pairs. Experimental results show that the crowdsourced relevances do not positively contribute to the performance when they are reduced to binary relevances, and using only binary relevances defined by captioning-based audio-caption pairs is sufficient for contrastive learning.

6. ACKNOWLEDGMENT

7. REFERENCES

[1] M. Slaney, "Semantic-audio retrieval," in *Proc. Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2002, pp. IV–4108–IV–4111.

[2] S. Ikawa and K. Kashino, "Acoustic event search with an onomatopoeic query: Measuring distance between onomatopoeic words and sounds," in *Proc. Detect. Classif. Acoust. Scenes Events Work. (DCASE)*, 2018, pp. 59–63.

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

[4] C. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. (HLT-NAACL)*, 2019, pp. 119–132.

[5] H. Xie, S. Lipping, and T. Virtanen, "Language-based audio retrieval task in dcase 2022 challenge," in *Proc. Detect. Classif. Acoust. Scenes Events Work. (DCASE)*, 2022, pp. 216–220.

[6] A. Oncescu, A. Koepke, J. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2021, pp. 2411–2415.

[7] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, pp. 2880–2894, 2020.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. (NAACL-HLT)*, 2019, pp. 4171–4186.

[9] X. Mei, X. Liu, J. Sun, M. Plumbley, and W. Wang, "On metric learning for audio-text cross-modal retrieval," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2022, pp. 4142–4146.

[10] H. Xie, O. Räsänen, and T. Virtanen, "On negative sampling for contrastive audio-text retrieval," in *Proc. Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2023, pp. 1–5.

[11] A. Van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018. [Online]. Available: https://arxiv.org/abs/1807.03748

[12] TAU Audio-Text Graded Relevances 2023 Dataset, https://github.com/xieh97/retrieval-relevance-crowdsourcing.

[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. Manag.*, p. 102688, 2021.

[14] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Empherical Methods Nat. Lang. Process. (EMNLP)*, 2019, pp. 3982–3992.

[15] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2017, pp. 776–780.