Sound event envelope estimation in polyphonic mixtures

Irene Martín-Morató, Annamaria Mesaros

Computing Sciences, Tampere University, Finland

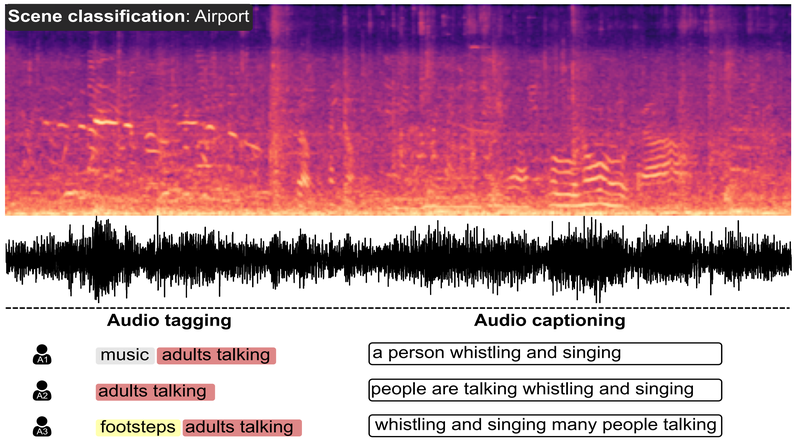
Abstract

Describing soundscapes in sentences allows better understanding of the acoustic scene than a single label indicating the acoustic scene class or a set of audio tags indicating the sound events active in the audio clip. We study how much the collection of audio captions can be guided by the instructions given in the annotation task, by analysing the possible bias introduced by auxiliary information provided in the annotation process. In this work, we address the diversity obtained when collecting descriptions of soundscapes using crowdsourcing, studying how differently annotators describe the same soundscape. We also release a new dataset of audio captions and audio tags produced by multiple annotators, publicly available, which we call MACS[1].

MACS

Multi-Annotator Captioned Soundscapes. Audio from TAU Urban Acoustic Scenes 2019 development dataset.

- Scenes: Airport, public square and park.
- Files: total of 3930 audios, each one 10 seconds long.
- Annotators: 133 students, assigned into 30 groups, each file annotated by 5 different annotators.



Other datasets

AudioCaps [4] is a collection of sentence-long descriptions for a subset of AudioSet [3]. The video was provided to be played if necessary, and the AudioSet tags were presented to the annotator as hints. The dataset contains over 46k files of 10 seconds each, and one caption per file, collected using MTurk.

Clotho [1] is collected using MTurk and contains five captions per clip, for audio clips 15 to 30 seconds long that were collected from Freesound [2]. We consider this dataset as having no bias, since the captions are based solely on the audio clip provided, and no additional information regarding the possible active sounds or clip content was available to annotators.

Dataset	Audio clips	Vocab. size	Unique sentences	Sentence length (std)
AudioCaps	57188	5218	52198	9.17 (4.27)
Clotho	5929	4373	29611	11.34 (2.78)
MACS	3930	2775	16262	9.46 (3.89)

Table 1: Statistics of the studied datasets.

Diversity, bias and similarity metrics

Lexical diversity

Measured with Type-Token ratio, often used in measuring language acquisition in infants or learners of a second language.

$$\mathrm{TTR} = \frac{\#Tokens}{Totalnumberofwords},$$

It ranges from a theoretical 0 (infinite repetition of a single word) and 1 (no repetition at all). Results considering the whole dataset (overall) and descriptions belonging to the same item (local):

S	L	AudioCaps overall	Clotho overall	Clotho local	MACS overall	MACS local
-	-	1.09%	1.30%	56.52%	1.80%	69.37%
-	✓	0.79%	0.91%	52.08%	1.38%	66.06%
✓	✓	1.27%	1.66%	60.43%	2.17%	71.02%

Table 2: Global and local lexical diversity of captions. S: removal of stopwords; L: lemmatization. AudioCaps has only a single caption per clip, thus we do not calculate local lexical diversity for it.

Most used words

- **MACS**: talk, people, adult, noise and bird.
- **AudioCaps**: man, speak, follow, talk and engine.
- **Clotho**: bird, water, background, chirp and someone.

Vocabulary bias

Defined as the proportion of hinted sounds with respect to the number of sounds mentioned in the caption. AudioSet vocabulary, a total of 722 labels, is used to identify sound events present in the captions.

	Tag bias (std)	Word bias (std)
AudioCaps	0.33 (0.35)	0.35 (0.35)
MACS	0.38 (0.36)	0.49 (0.38)

Table 3: Calculated vocabulary bias.

- MACS →hints are the 10 pre-defined tags
- AudioCaps →hints are the tags associated to the clip in AudioSet.

Similarity

To study how similar are the descriptions by different annotators from the same item, we use metrics from machine translation, calculated for every pair of captions. Basic approaches to compare similarity between sentences:

- Jaccard similarity coefficient (J)

$$J(a,b) = \frac{|S_a \cap S_b|}{|S_a \cup S_b|},$$

- Bilingual Language Understudy

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right),$$

Approaches using BERT, pre-trained model on large amounts of unlabeled data:

- BERTscore: contextual embeddings + cosine similarity
- sentence-BERT: modification of BERT using Siamese networks. Encodes the entire sentence instead of token by token.

Dataset	BLEU-4	Jaccard	sBERT	BERTscore
Clotho	0.06 (0.04)	0.22 (0.09)	0.61 (0.13)	0.88 (0.01)
MACS	0.01 (0.02)	0.16 (0.08)	0.55 (0.12)	0.87 (0.01)

Table 4: Average similarity of the captions, using multiple metrics.

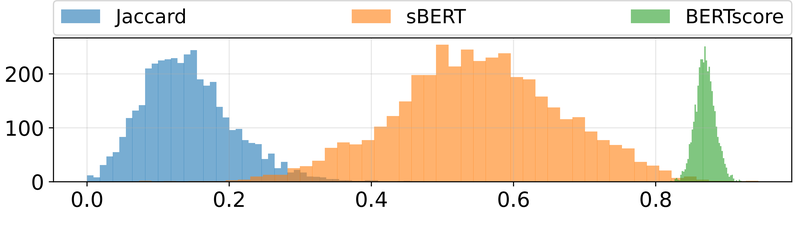


Figure 1: Similarity metrics for MACS dataset.

Conclusions

This paper presented a study of the lexical diversity, bias, and similarity of captions from three audio captioning datasets. Bias study showed how the free-text descriptions are more affected by the complexity of the soundscape. Despite the hints, the captions in the studied datasets have a high lexical diversity, being the semantic similarity between captions assigned to the same clips by different annotators high. The new captions dataset along with the tags provided by the same annotators, brings novel elements to audio captioning; for example the tag-caption pairs allow guided captioning, and the estimated annotator reliability provides a measure of trustworthiness for each caption, which can be used in the learning process.

References

[1] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: an audio captioning dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740, 2020.

[2] Frederic Font, Gerard Roma, and Xavier Serra. Freesound technical demo. In *ACM International Conference on Multimedia (MM'13)*, pages 411–412, Barcelona, Spain, Oct. 2013. ACM, ACM.

[3] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017.

[4] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the NAACL HLT, Vol. 1*, pages 119–132, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[1]MACS dataset: https://zenodo.org/record/5114771