SONG DESCRIBER: A PLATFORM FOR COLLECTING TEXTUAL DESCRIPTIONS OF MUSIC RECORDINGS

Ilaria Manco[1] **Benno Weck**[2] **Philip Tovstogan**[2]
Minz Won[2,*] **Dmitry Bogdanov**[2]
[1] School of EECS, Queen Mary University of London, London, U.K.
[2] Music Technology Group, Universitat Pompeu Fabra, Spain
Corresponding author: i.manco@qmul.ac.uk

ABSTRACT

We present Song Describer, an open-source data annotation platform for crowdsourcing textual descriptions of music recordings. Through this tool, we propose to collect annotations with the goal of creating the first public dataset of audio-caption pairs in the music domain. We believe that such a dataset will be useful in supporting the growing interest in the integration of natural language processing within music information retrieval systems. In this paper, we describe our approach to designing Song Describer, outline the data collection protocol, and illustrate the main steps involved in using the platform.

1. INTRODUCTION

The use of natural language text within music information retrieval (MIR) research is becoming increasingly common, with a growing number of publications [1–3] and research initiatives, such as the NLP4MusA workshop [4], dedicated to the topic. Automatic systems to analyse, process, and relate natural language and audio signals have many important potential applications. For example, in the music domain, they can be used to automatically generate captions to describe the content of a music recording or to retrieve a piece of music based on text inputs. However, to date, no public datasets with aligned audio-text data exist to adequately support research in this area, resulting in works so far mostly relying on private [5] or web-crawled data [6]. To fill this gap, we propose to crowdsource the first dataset of music recordings paired with textual descriptions. In this paper, we present our approach to designing and building *Song Describer*, the data annotation platform created to collect such a dataset.

Crowdsourcing has become an established practice for collecting datasets, most prominently in NLP [7–9] and occasionally in other fields [10, 11], including MIR [12–14]. Its success derives from the fact that it allows to conveniently access a large pool of annotators at a relatively low cost [15–17] and, as a result, crowdsourcing platforms such as Amazon Mechanical Turk, [1] have become part of the researcher's toolkit. However, while convenient, using such platforms comes with important ethical and practical implications. Prior papers have raised concerns about the risks crowd workers can be exposed to [18] and highlighted that misaligned incentives between researchers and workers can lead to poor data quality [19]. In an attempt to sidestep some of these issues, and access an even larger pool of data, other works have turned to web scraping as an alternative to crowdsourcing [20], especially to obtain large-scale datasets needed to train highly data-hungry models [21]. However, web scraping also carries several ethical and practical implications, such as the risk of including offensive and harmful content, inadvertently collecting biased data [22], or infringing copyright [23], a major concern in the context of music and other works of art.

For all these reasons, we argue that adopting a citizen science approach, similarly to [14], can offer a fairer and more suitable alternative for collecting music captions. We show how developing a purpose-built data collection platform allows us to exercise better control over the task design, ensure that data ownership is properly dealt with, and more actively engage participants in the research process, all while still tapping into a large pool of annotators. We make our code and supporting material available online. [2]

2. MUSIC CAPTIONS

The ultimate goal of our platform is to collect a dataset of music captions that describe a set of music recordings in a short piece of text. More specifically, we consider descriptions that do not require expert-level musical knowledge and that may be written and understood by the average music listener. This type of music caption focuses on high-level characteristics of the music, such as genre, instrumentation, era, mood, emotions evoked, and similar. The task of generating music captions shares many similarities with the more well-studied tasks of image and audio captioning, both of which have been extensively addressed

[*] Currently at ByteDance.

© I. Manco, B. Weck, P. Tovstogan, M. Won and D. Bogdanov. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** I. Manco, B. Weck, P. Tovstogan, M. Won and D. Bogdanov, "Song Describer: a Platform for Collecting Textual Descriptions of Music Recordings", in *Extended Abstracts for the Late-Breaking Demo Session of the 23rd Int. Society for Music Information Retrieval Conf.,* Bengaluru, India, 2022.

[1] https://www.mturk.com/
[2] https://github.com/ilaria-manco/song-describer

Late-Breaking / Demo Session Extended Abstract, ISMIR 2022 Conference

in machine perception research and are supported by an increasing number of datasets. While music captioning can be partly regarded as a sub-task of audio captioning, there are important differences between the two, as music descriptions are inherently more subjective and do not necessarily contain references to audio events localised in time.

3. SONG DESCRIBER

The frontend of *Song Describer* is built as a multi-page web application using Streamlit. [3] The app is intended to be easy to use and engaging, and is aimed at annotators that are both non-experts and volunteers. Participants are only required to be 18 or over, have an internet connection and have access to a supported web browser. We do not screen participants according to any other criteria, but collect answers to background questions for post-hoc quality control and analysis. To ensure user privacy, no personal data is collected and participants are never asked to provide information that may make them identifiable.

3.1 Data Collection Protocol

The overall data collection pipeline is composed of three stages: onboarding, annotation and evaluation. In the first stage, participants are asked to provide some background information, before moving to the actual tasks to complete, presented in the two later stages, each of which can be repeated an arbitrary number of times.

Task 0: Onboarding On the onboarding page, participants are asked to answer two sets of questions: the first covers basic demographic information (age, country of origin and level of comfort writing in English); the second is aimed at assessing their musical engagement. To design the questions for the second set, we take inspiration from the Goldsmiths Musical Sophistication Index (Gold-MSI) self-report inventory [24] and adapt the three most relevant questions to our task.

This onboarding stage serves two purposes: firstly, it is needed in order to create a profile of the annotator to which all their responses are associated; this also allows participants to log back onto the platform by using their unique user ID and continue contributing without losing track of their progress; secondly, this background information is needed to characterise the population from which the data is collected. This is instrumental in understanding certain aspects of the data such as how cultural differences may affect the way listeners describe music [25–28].

Task 1: Annotation On the annotation page, participants are shown an audio player and are presented with step-by-step instructions, as shown in Figure 1. In order to complete the task, participants go through three steps: listen to a two-minute extract of a music recording, type one sentence describing the track, and indicate their familiarity with the kind of music they have just listened to on a 3-point Likert scale.

[3] https://streamlit.io/

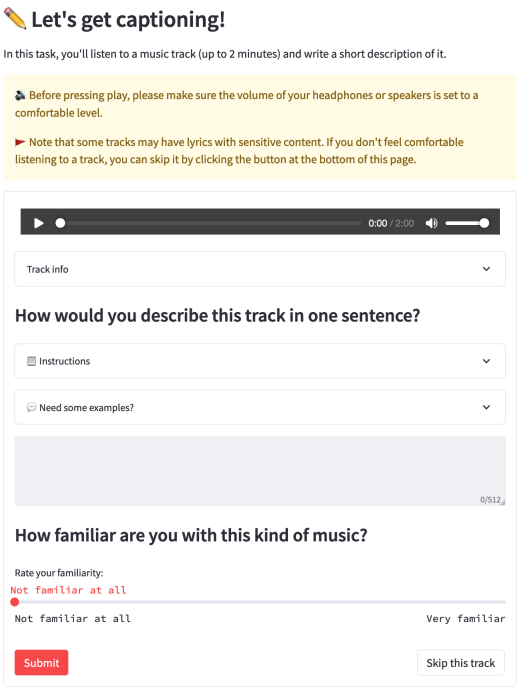

Figure 1: Annotation page of *Song Describer*.

Our goal in this task is to elicit descriptions that resemble as closely as possible how someone would concisely describe a piece of music in the real world. Therefore, we opt against giving hints to the annotators, which has been shown to influence the choice of words [29], but provide six examples of captions, as participants in our pilot study reported that this is necessary to reduce task ambiguity. As audio recordings, we use music from the MTG-Jamendo dataset [30], available under Creative Commons licenses.

Task 2: Evaluation In this task, users need to first indicate whether a given caption is valid. In case of a positive response, they are then invited to listen to a track and rate how well the caption describes it, on a 5-point Likert scale.

Asking participants to evaluate other annotations is a commonly used strategy in crowdsourcing. As in our case, this is intended both as a strategy for quality control, since annotations that are consistently rated poorly are likely to be low-quality, and as a way to measure inter-annotator agreement. The latter is useful for several reasons: alongside providing a measure of annotation noise, it is also an indicator of how difficult a data point may be [31]. This is particularly valuable when using the data for training machine learning models, since inter-rater agreement is also often positively correlated with prediction accuracy [32].

4. CONCLUSION

We have presented *Song Describer*, an open-source platform for annotating music with textual descriptions. Through this data collection initiative, we hope to create and release the first public dataset with paired music audio and natural language, with the goal of promoting audio-and-language research in the music domain.

5. REFERENCES

[1] M. Won, J. Salamon, N. J. Bryan, G. J. Mysore, and X. Serra, "Emotion Embedding Spaces for Matching Music to Stories," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[2] C. Tian, M. Michael, and H. Di, "Music autotagging as captioning," in *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*. Association for Computational Linguistics, 2020.

[3] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Learning music audio representations via weak language supervision," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

[4] "Proceedings of the 2nd Workshop on NLP for Music and Spoken Audio (NLP4MusA)," in *Proceedings of the 2nd Workshop on NLP for Music and Spoken Audio (NLP4MusA)*, S. Oramas, E. Epure, L. Espinosa-Anke, R. Jones, M. Quadrana, M. Sordo, and K. Watanabe, Eds. Association for Computational Linguistics, 2021.

[5] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Contrastive Audio-Language Learning for Music," in *23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*, 2022.

[6] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. W. Ellis, "MuLan: A Joint Embedding of Music Audio and Natural Language," in *23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*, 2022.

[7] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2013.

[8] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft COCO Captions: Data Collection and Evaluation Server," *arXiv preprint*, 2015.

[9] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics (ACL), 2016.

[10] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An Audio Captioning Dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[11] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics (ACL), 2019.

[12] M. Soleymani, M. N. Caro, E. M. Schmidt, C. Y. Sha, and Y. H. Yang, "1000 songs for emotional analysis of music," *CrowdMM 2013 - Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, 2013.

[13] E. Fonseca, J. Pons Puig, X. Favory, F. Font Corbera, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Proceedings of the 18th ISMIR Conference*. International Society for Music Information Retrieval (ISMIR), 2017.

[14] N. F. Gutiérrez Páez, J. S. Gómez-Cañón, L. Porcaro, P. Santos, D. Hernández-Leo, and E. Gómez, "Emotion Annotation of Music: A Citizen Science Approach," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2021.

[15] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, 2010.

[16] A. Drutsa, D. Ustalov, V. Fedorova, O. Megorskaya, and D. Baidakova, "Crowdsourcing Natural Language Data at Scale: A Hands-On Tutorial," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials.* Association for Computational Linguistics (ACL), 2021.

[17] A. Suhr, C. Vania, N. Nangia, M. Sap, M. Yatskar, S. R. Bowman, and Y. Artzi, "Crowdsourcing Beyond Annotation: Case Studies in Benchmark Data Collection," in *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*. Association for Computational Linguistics (ACL), 2021.

[18] B. Shmueli, J. Fell, S. Ray, and L.-W. Ku, "Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

[19] F. Daniel, P. Kucherbaev, C. Cappiello, B. Benatallah, and M. Allahbakhsh, "Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions," *ACM Computing Surveys (CSUR)*, vol. 51, no. 1, 2018.

[20] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1. Association for Computational Linguistics (ACL), 2018.

[21] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. Mccandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, 2020.

[22] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner, "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus," in *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*. Association for Computational Linguistics (ACL), 4 2021.

[23] V. Krotov and L. Silva, "Legality and Ethics of Web Scraping," in *Twenty-fourth Americas Conference on Information Systems*, 2018.

[24] D. Müllensiefen, B. Gingras, J. Musil, and L. Stewart, "The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population," *PloS one*, vol. 9, no. 2, 2 2014.

[25] C. McKay and I. Fujinaga, "Musical genre classification: Is it worth pursuing and how can it be improved?" in *ISMIR*, 2006.

[26] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music Emotion Recognition: A State of the Art Review," in *11th International Society for Music Information Retrieval Conference (ISMIR)*, 2010.

[27] J. S. Gómez-cañón, E. Cano, and S. Ug, "Joyful for You and Tender for Us: The Influence of Individual Characteristics and Language on Emotion Labeling and Classification," *Proceedings of the 21th International Symposium on Music Information Retrieval (ISMIR)*, 2020.

[28] H. Lee, F. Hoeger, M. Schoenwiesner, M. Park, and N. Jacoby, "Cross-cultural Mood Perception in Pop Songs and its Alignment with Mood Detection Algorithms," in *22nd International Society of Music Information Retrieval (ISMIR)*, 2021.

[29] I. Martín-Morató and A. Mesaros, "Diversity and bias in audio captioning datasets," in *Proceedings of the 6th Workshop on Detection and Classification of Acoustic Scenes and Events.*, 2021.

[30] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The MTG-Jamendo Dataset for Automatic Music Tagging," in *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML)*, 2019.

[31] E. Pavlick, T. Kwiatkowski, and G. Research, "Inherent Disagreements in Human Textual Inferences," *Transactions of the Association for Computational Linguistics*, vol. 7, 2019.

[32] Y. Nie, X. Zhou, and M. Bansal, "What Can We Learn from Collective Human Opinions on Natural Language Inference Data?" in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.