AUDIO CONTENT REPLICATION DETECTION

Yutong Wen
University of Rochester
yutong.wen@rochester.edu

Second Author
Retain these fake authors in submission to preserve the formatting

Third Author
Affiliation3
author3@ismir.edu

ABSTRACT

Audio copy detection plays a pivotal role in preserving the legitimacy and integrity of information, especially in the context of social media platforms where manipulated and re-encoded audio clips can circulate. This paper addresses the crucial task of determining whether an audio clip is a modified version of another source audio or if two audio clips share a common origin through editing. To address this challenge, we present a comprehensive audio similarity dataset covering various real-world scenarios, including frequently employed manipulations in audio editing such as temporal, spectral, and deepfake alterations. Our dataset serves as a foundation for extrapolating algorithms to operate at scale in practical scenarios. Additionally, we propose a baseline method for audio copy detection based on contrastive learning. In summary, this paper defines the task of audio copy detection as a novel and practical challenge with broad real-world applications. The introduction of a large-scale audio similarity dataset, along with a baseline method based on contrastive learning, establishes a foundation for further research and development in this critical domain.

1. INTRODUCTION

Evaluating whether a audio clip constitutes a modified version of another source audio or determining if two audio clips have been edited from the same source audio is an important task, particularly within the context of maintaining the legitimacy and integrity of information, especially on social media platforms. [1,2]. In this context, audio copy detection is a task that aims to determine whether a part of an audio clip is copied from another audio clip through manipulations and re-encoding.

The task of audio copy detection has its own significance as it could be widely used by Internet services for violating content regulation, copyright preservation, as well as novel product features such as reverse audio search. Social media platforms could utilize audio copy detection to expedite content regulation process, particularly in real-world scenarios involving large-scale content searches where manual curation is impractical. Simultaneously, audio copy detection could be employed to identify unauthorized copies of copyrighted media. In this process, copyright holders identify the copyrighted media to be detected. Moreover, streaming services could implement audio copy detection for reverse audio search, searching similar audio clips based on a given source audio.

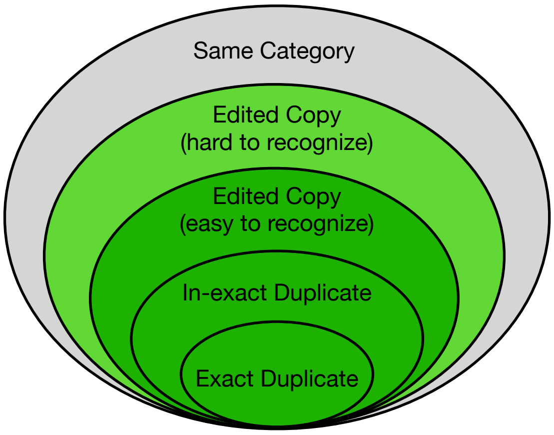

Figure 1. Audio pair similarity level at different levels of granularity. concepts at more inner means more restricted. Our dataset covers levels of granularity corresponds to all green areas.

There are existing tasks related to audio copy detection. The task of music matching aims to find a particular song or piece of music based on a short segment or snippet, audio fingerprinting is one technique to tackle this task [3–7]. There are also works do audio copyright detection using audio fingerprinting [8]. These works only focus on detecting exact duplicates in audio clips. In parallel, audio forensics involves the scientific analysis and examination of audio recordings to gather information for legal or investigative purposes. It employs various techniques and tools to enhance, authenticate, or interpret audio evidence. One of the questions that audio forensics aim to answer the question of "whether the query audio has been tampered with since its creation" [9,10]. Recently, with the advancement of deep audio generative models, the concern over the generated contents replicating training data has been also explored through copy detection [11,12]. Both works leverage self-supervised models to identify replicated contents.

Works related to audio copy detection has been exploring this area over decades. However, they focused either on exact duplicates or partial exact duplicates on very specific domain in music matching or finding the fix-length similar audio clips without providing a detailed assessment of the models being used in the copy detection for audio clips from generative models as their primary goal is to show this phenomenon. The task of audio copy detection as defined in our work is a novel and practical challenge. This paper proposes a audio similarity dataset where we present a sufficiently large and difficult dataset corresponding to different real-world scenarios to extrapolate algorithms to this operation scale. The dataset have been constructed to include frequently employed types of manipulations in audio editing, encompassing temporal, spectral, and even deepfake alterations between the query audio clip and the target ones. In addition, we propose a method to tackle this task based on contrastive learning as a baseline.

To summarize, this work introduces the task of audio copy detection which has wide real-world applications, and proposes a large-scale audio similarity dataset corresponding to different practical scenarios. We further introduces a baseline mothod based on contrastive learning as a starting point for this task.

2. TASK DESCRIPTION

Two audio clips may be considered similar according to varying criteria at different levels of granularity as shown in Figure 1.

The most restrictive form is exact duplicate, meaning two audio clips are the same sample-wisely. Closely related to this are near-exact duplicate, where two audio clips are nearly indistinguishable perceptually but differ in actual content. An example of this is loss due to file compression.

Edited copy, which is also the main focus of our dataset, corresponds to a pair of audio clips that are modified versions of each other or of a same source clip. Edited copy can be further divided into two categories: edited copy that's easy to recognize and hard to recognize. If we can identify two audio clips being edited copies of each other by their contents easily, then this pair falls into the category of edited copy that's easy to recognize. This corresponds to practical scenarios of audio content regulation. We aim to regulate an audio clip only if it retains recognizable violating content, even if it originates from a restricted audio clip. On the other hand, if a pair of audio clips are edited copies of each other but is not perceptually recognizable or very hard to recognize, this pair falls into the category of the edited copy that's hard to recognize. An example real-world scenario of this category is copyright protection. An copyright protected sound effect is allowed to be modified and included in an end product. This sound effect could be modified in very creative ways to an extend that it's very hard to recognize its origin.

There are also audio clips of the same instance and category, for instance, audio clips of a same door or a category of footsteps.

In this work, we limit the detection targets within the levels of granularity in green areas as shown in Figure 1: exact duplicates, near-exact duplicates, and edited copies of both kinds.

3. DATASET

Following the dataset construction of copy detection task in other modalities [13], our dataset is composed by four parts: the *reference set*, two *query sets*, and the *training set*. The reference set contains all the source audio clips. Each query set constitutes of edited copies or duplicates of the partial reference set and *distractor queries* which are edited copies or duplicates of source audio clips outside the reference set. The two query sets in our dataset corresponds to copy detection tasks, with and without the category of edited copies that are hard to recognize. Including this category significantly increase the level of difficulty. The *training set* is constructed similarly to the reference set.

3.1 Data Sources

For this preliminary examination of this task, we use Epidemic Sound dataset which contains 75626 audio clips of sound effects and short music pieces. For data preprocessing, we fix the length of the audio clips to four seconds and trim the silence at the beginning and the end of each audio clip in this preliminary study, as variable-length audio clips would be more challenging and more computationally expensive. In addition, we limit the sampling rate at 16kHz.

3.2 Audio Transformations

Manual transformations.

Automatic transformations are applied to source clips using common audio augmentation methods. These transformations can be classified into following categories: time domain alternations, spectral domain alternations, and overlay with other sound sources. Time domain alternations include time shift, partial inclusion, stretch or compress of audio clips. Spectral domain alternations include random spectral cropping, down-sampling, and resynthesis through reverbs. Finally, overlay with other sound sources include injecting random noise, background noise, and mix-up with other audio clips.

An audio clip may have single or multiple transformations being applied. The automatic transformations parameters are selected randomly within a range corresponding to different levels of granularity. For edited copies that are easy to recognize, we limit the range of the transformations parameters narrower so that for audio clips even with multiple transformations being added, we can still recognize their origins. On the other hand, for edited copies that are hard to recognize, we select a wider range of parameters for transformations, as we aim to reflect more creative audio manipulations employed in industries. We do not include transformations with parameters that are completely infeasible for humans to identify or make no practical sense.

3.3 Dataset Structure

For the final project, the scope of the dataset limits to a reference set, a training set, and a query set which corresponds to copies mainly including edited copies of the easy-to-recognize category, and the edited copies of the hard-to-recognize category is beyond the scope of this project.

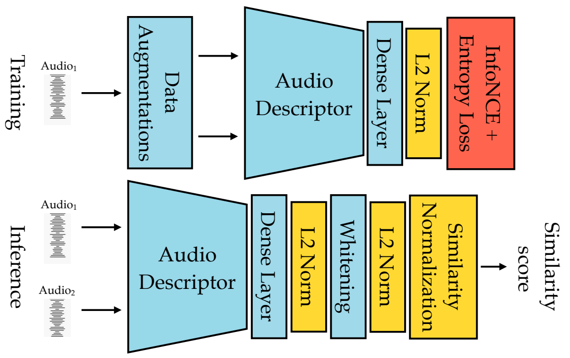

Figure 2. The model architecture for audio copy detection. The audio descriptor is adapted from PANNs.

Reference set contains 37072 four seconds mono channel audio clips at 16kHz sampling rate. The audio clips in the reference set are not processed through any transformations.

Training set contains 37072 audio clips which are collected in the same manner as the reference set. The training set not only can be used in the audio copy detection task, but can be also used in other tasks such as audio synthesis.

Query set contains 1852 audio clips in total which are in the same format as the reference set. The query set includes 1482 distractor queries and 370 true queries. The distractor queries have no matching counterpart in the reference set, and contains no overlap with the training set. All the audio clips in the query set are transformed to some extend.

3.4 Evaluation Metric

An algorithm for audio copy detection generates pairs along with confidence values, where each pair associates a query audio clip with a candidate audio clip from the *reference set*. In the case of *distractor queries*, their absence in this set is acceptable, as they do not correspond to any audio clips within the *reference set*. Indeed, any appearances of *distractor queries* should decrease the algorithm's performance. We use micro Average Precision as the metric for this task. This metric is also widely-used in image copy detection tasks and instance recognition tasks [13–15].

4. METHOD

This section introduces our proposed model for audio copy detection. The model is trained contrastively and outputs a similarity score for a given pair of fixed-length audio clips. Then we run this model multiple times if needed to compute the similarity score between two variable-length audio clips. Finally we set a threshold as the confidence level on the similarity score to make a hard decision of whether we have detected a copy.

4.1 Model Architecture

Audio descriptor.

5. EXPERIMENT

5.1 Baseline Method

We select the fine-tuned CLAP [11] as the baseline method, as this method was developed to identify audio replication as well. During inference, the audio embeddings of two audio clips are obtained using the fine-tund CLAP, and the cosine similarity between these two embeddings is then computed as the similarity score between the two audio clips. For a given query audio clip, the audio clip in the *reference set* with the highest similarity score above a given threshold is identified as a copy to the query audio clip. During training, the authors proposed to leverage the pretrained CLAP [16] and added two dense layers which contain all the trainable parameters during finetuning. Furthermore, the Triplet Loss [17] was employed to make the audio embeddings more descriptive. The authors showed that this method improves the copy detection performance. We compare our proposed algorithm with this baseline method in terms of micro Average Precision metric. We will also use this baseline method to tune the audio transformation parameters so that the dataset has a reasonable difficulty.

6. CONCLUSION AND FUTURE WORK

This paper introduced the task of audio copy detection, and proposed an audio similarity dataset to tackle and assess this task. Future directions lie in selecting appropriate range of audio transformation parameters to tailor the dataset to real-world scenarios, and at the same time has reasonable difficulty.

7. REFERENCES

[1] S. J. Nightingale and H. Farid, "Ai-synthesized faces are indistinguishable from real faces and more trustworthy," *Proceedings of the National Academy of Sciences*, vol. 119, no. 8, p. e2120481119, 2022.

[2] K. Bhagtani, A. K. S. Yadav, E. R. Bartusiak, Z. Xiang, R. Shao, S. Baireddy, and E. J. Delp, "An overview of recent work in media forensics: Methods and threats," *arXiv preprint arXiv:2204.12067*, 2022.

[3] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system." in *Ismir*, vol. 2002, 2002, pp. 107–115.

[4] A. Wang *et al.*, "An industrial strength audio search algorithm." in *Ismir*, vol. 2003. Washington, DC, 2003, pp. 7–13.

[5] T.-K. Hon, L. Wang, J. D. Reiss, and A. Cavallaro, "Audio fingerprinting for multi-device self-localization," *IEEE/ACM Transactions on Audio, Speech, and language processing*, vol. 23, no. 10, pp. 1623–1636, 2015.

[6] B. Gfeller, R. Guo, K. Kilgour, S. Kumar, J. Lyon, J. Odell, M. Ritter, D. Roblek, M. Sharifi, M. Velimirović *et al.*, "Now playing: Continuous low-power music recognition," *arXiv preprint arXiv:1711.10958*, 2017.

[7] S. Chang, D. Lee, J. Park, H. Lim, K. Lee, K. Ko, and Y. Han, "Neural audio fingerprint for high-specific audio retrieval based on contrastive learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3025–3029.

[8] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of audio fingerprinting," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 41, pp. 271–284, 2005.

[9] M. Zakariah, M. K. Khan, and H. Malik, "Digital multimedia audio forensics: past, present and future," *Multimedia tools and applications*, vol. 77, pp. 1009–1040, 2018.

[10] P. R. Bevinamarad and M. Shirldonkar, "Audio forgery detection techniques: Present and past review," in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*. IEEE, 2020, pp. 613–618.

[11] G. Zhu, Y. Wen, M.-A. Carbonneau, and Z. Duan, "Edmsound: Spectrogram based diffusion models for efficient and high-quality audio synthesis," *arXiv preprint arXiv:2311.08667*, 2023.

[12] D. Bralios, G. Wichern, F. G. Germain, Z. Pan, S. Khurana, C. Hori, and J. L. Roux, "Generation or replication: Auscultating audio latent diffusion models," *arXiv preprint arXiv:2310.10604*, 2023.

[13] M. Douze, G. Tolias, E. Pizzi, Z. Papakipos, L. Chanussot, F. Radenovic, T. Jenicek, M. Maximov, L. Leal-Taixé, I. Elezi *et al.*, "The 2021 image similarity dataset and challenge," *arXiv preprint arXiv:2106.09672*, 2021.

[14] E. Pizzi, S. D. Roy, S. N. Ravindra, P. Goyal, and M. Douze, "A self-supervised descriptor for image copy detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 532–14 542.

[15] F. Perronnin, Y. Liu, and J.-M. Renders, "A family of contextual measures of similarity between distributions with application to image retrieval," in *2009 IEEE Conference on computer vision and pattern recognition*. IEEE, 2009, pp. 2358–2365.

[16] 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 *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[17] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.