AUDIOCITE.NET DATASHEET

From the template provided by the paper, “Datasheets for Datasets”*

Motivation

1. For what purpose was the dataset created? *Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

The dataset was created for the Lebenchmark project (http://lebenchmark.com/) a collaboration between University Grenoble Alpes, University of Avignon and Université Paris Dauphine. It was used in this project to train self-supervised Wav2vec 2.0 models for French speech. The dataset was added to other resources available in French. Multiple downstream tasks can be achieved with this type of pre-trained model such as automatic speech recognition, automatic machine translation, spoken language understanding, and automatic emotion recognition.

The objective was to make available a large publicly available French speech dataset for building large language processing models for French. Some other smaller datasets exist but they were already included in the LeBenchmark project and we needed to get an amount of data equivalent to other languages to enable comparison between studies across languages. The audiocite.net data was particularly relevant in this respect since most large datasets available in other languages are composed of read speech. Furthermore, the audiocite.net records have been made to be open and freely shareable.

2. Who created the dataset (e.g. which team, research group) and on behalf of which entity (e.g. company, institution, organization)?

The dataset was created by the GETALP team (*https://lig-getalp.imag.fr*) of the LIG (Laboratoire d'Informatique de Grenoble) which is part of the University Grenoble Alpes.

3. Who funded the creation of the dataset? *If there is an associated grant, please provide the name of the grantor and the grant name and number.*

This work was partially supported by MIAI@Grenoble-Alpes (ANR-19-P3IA-0003).

4. Any other comments?

None.

Composition

5. What do the instances that comprise the dataset represent (e.g. documents, photos, people, countries)? *Are there multiple types of instances (e.g. movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

The data in this corpus are audio (in mp3 format). Metadata only describe the audio files and those contain only read speech by one or multiple speakers. The information contained in the description files comes from information provided by the speakers downloaded from the website. Some of the information has been completed by our team.

6. How many instances are there in total (of each type, if appropriate)?

There are 28485 audio files containing 6682 hours of audio recording. There are a total of 130 speakers composed of 51 males, 70 females and 9 unknown speaker gender.

7. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g. geographic coverage)? If so, please describe how this* *representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g. to cover a more diverse range of instances, because instances were withheld or unavailable).*

This corpus is a large subpart of instances that were freely available on the audiocite.net website at the time of the download. We used the list of records that was advertised by the website. Only a handful of records were not downloaded or used due to a problem in the downloading or because of file corruption. The dataset is thus representative of what can be found on the original website.

8. What data does each instance consist of? *“Raw” data (e.g. unprocessed text or images)or features? In either case, please provide a description.*

Each instance consists of an audio file stored in a folder with the name of the corresponding audiobook. Each audio file has a corresponding entry in a json file. For each instance, the json file provides the following information about the mp3: the name of the audio file, the ID of the speaker(s), the duration of the recording, the path to the audio file, and the gender of the speaker[1].

In a separate csv file, further information is provided such as the title of the book, the link to the webpage of the audiobook on audiocite.net, the author of the book, the category of the book, the license that protects the work, the identifier of the speaker, the link to the audio file or archive and the relative path where the file(s) is located in the dataset.

9. Is there a label or target associated with each instance? *If so, please provide a description.*

Multiple labels exist to describe every mp3 file as the category of the book read. The categories available are : contes (tales), planete-actuelle (world news), nouvelles (short stories), poesies (poetry), charme (erotic story), documents, science-fiction, romans (novels), animaux (animals), audiocite-juniors, religions, philosophies, histoire (history) and theatre.

For each instance, the original author and the speaker is indicated. For each speaker, one of the classes (Male, Female, Unknown) has been manually associated.

10. Is any information missing from individual instances? *If so, please provide a description, explaining why this information is missing (e.g. because it was unavailable). This does not include intentionally removed information, but might include, e.g. redacted text.*

No.

11. Are relationships between individual instances made explicit (e.g. users' movie ratings, social network links)? *If so, please describe how these relationships are made explicit.*

No.

12. Are there recommended data splits (e.g. training, development/validation, testing)? *If so, please provide a description of these splits, explaining the rationale behind them.*

The mp3 files are distributed according to the json files named « dev_files.json », « test_files.json » and « train_files.json ». This partitioning has been performed from « all_files.json » which describes all the instances of the dataset. « dev_files.json », « test_files.json » and « train_files.json » represent 10%, 10%, and 80% of « all_files.json » respectively (no overlap).

During the partitioning, the « dev_files.json » and the « test_files.json » be composed of audio files that do not contain content that may be considered as sensitive (categories "charmes" (erotic), "planete-actuelle" (geopolitics) and "religions"). Furthermore, they do not contain files whose speaker gender was labeled as unknown. The quota of male and female speech was also equally distributed in the dev and test recommended partition.

13. Are there any errors, sources of noise, or redundancies in the dataset? *If so, please provide a description.*

The recordings provided on the website have undergone minimal processing, as per the website's recommendations (see https://www.audiocite.net/lecteurs.html#4). As a result, some recordings may contain background music, noise, or unexpected speakers. Additionally, some of the recordings are sourced from LibriVoice and this is indicated at the beginning of the audio files. With the exception of a few cases, the original mp3 encoding has been preserved, resulting in some recordings being mono and others being stereo, as well as different bit rates. Gender information may be missing for some recordings, but no duplicates or redundancies have been found in the audio or metadata files (cf. Question 36). It should also be noted that some speakers have read the same book. The majority of the recordings are book readings, but there are also articles or podcasts included.

14. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g. websites, tweets, other datasets)? *If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e. including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g. licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

Yes, the dataset is self-contained.

15. Does the dataset contain data that might be considered confidential (e.g. data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals non-public communications)? *If so, please provide a description.*

Knowing that the information related to the books and the readers are given freely by themselves, there is no confidential data in our dataset. However, we did not check whether the information in the metadata of the mp3 files contains sensitive information.

16. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? *If so, please describe why.*

It is possible that there is offensive content in the books read (cf. categories "charmes" (erotic), "planete-actuelle" (geopolitics) and "religions"), in which case it will show up in the audio, but we do not expect this to be the norm (cf. Question 12).

17. Does the dataset identify any subpopulations (e.g. by age, gender)? *If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

We have deduced a part of the speaker gender based on their name and their voice as explained above. We could find that our corpus contains more male speech (4131 hours) than female speech (2279 hours), not forgetting the cases where the gender of the speaker remains unknown (287 hours). No other subpopulation has been identified clearly.

18. Is it possible to identify individuals (i.e. one or more natural persons), either directly or indirectly (i.e. in combination with other data) from the dataset? *If so, please describe how.*

Yes, some people used their explicit full name some other pseudonyms. Indirect identification might be possible since each reader has a fair amount of speech in the dataset. Given that all authors distributed their record in a CC license, we respect such license.

19. Does the dataset contain data that might be considered sensitive in any way (e.g. data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? *If so, please provide a description.*

Apart from names there is no other direct sensitive information in the dataset metadata. Sensitive information might be found in the recordings as well as the mp3 metadata, but we did not find such instances in the records we have manually analyzed.

20. Any other comments?

None.

Collection Process

21. How was the data associated with each instance acquired? *Was the data directly observable (e.g. raw text, movie ratings), reported by subjects (e.g. survey responses), or indirectly inferred/derived from other data (e.g. part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

The data associated with each instance was collected by audiocite.net except for the gender information. For this, it was inferred from their identifier and verified through listening to their voice. However, this information is not complete and not certified as it isn't based on the speaker's self-identification.

22. What mechanisms or procedures were used to collect the data (e.g. hardware apparatuses or sensors, manual human curation, software programs, software APIs)? *How were these mechanisms or procedures validated?*

The data were collected by a script developed by the GETALP team through requests made on the audiocite.net website.

23. If the dataset is a sample from a larger set, what was the sampling strategy (e.g. deterministic, probabilistic with specific sampling probabilities)?

N/A.

24. Who was involved in the data collection process (e.g. students, crowd workers, contractors) and how were they compensated (e.g. how much were crowding workers paid)?

Concerning the extraction of the audio and the information for the creation of the dataset, permanent staff, Ph.D. students, and contractual worked on it at different levels. All contributed to the task as salaried people of the University Grenoble Alpes.

25. Over what timeframe was the data collected? *Does this timeframe match the creation timeframe of the data associated with the instances (e.g. recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

The audio and information related were downloaded to form this dataset in November 2021.

However, we have not collected the dates of the deposit of each audio individually.

26. Were any ethical review processes conducted (e.g. by an institutional review board)? *If so, please provide a description of these review processes, including the outcomes,as well as a link or other access point to any supporting documentation.*

No.

27. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g. websites)?

The data was collected through the audiocite.net website, where it was shared freely with an associated Creative Common license.

28. Were the individuals in question notified about the data collection? *If so, please describe (or show with screenshots or other* *information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

We have contacted the author of the website who has given us permission to publish this data.

29. Did the individuals in question consent to the collection and use of their data? *If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

By choosing a Creative Common license, the speakers have consented to the use and distribution of the data.

30. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? *If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

No, but the corpus maintenance team is committed to deleting the data for any justified request (as indicated on the audiocite.net website).

31. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g. a data protection impact analysis) been conducted? *If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

No.

32. Any other comments?

None.

Preprocessing/cleaning/labeling

33. Was any preprocessing/cleaning/labeling of the data done (e.g. discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? *If so, please provide a description. If not, you may skip the remaining questions in this section.*

The audio files are provided as they were downloaded. We have developed scripts to create metadata files from the downloaded data. Moreover, a completion of the gender information was carried out thanks to the information of the identifiers as well as the acoustic information.

34. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g. to support unanticipated future uses)? *If so, please provide a link or other access point to the “raw” data.*

Yes.

35. Is the software that was used to preprocess/clean/label the data available? *If so, please provide a link or other access point.*

Yes, the scripts used to download, make the json file and the partitionning are provided in the dataset archive.

36. Any other comments?

In the creation of json files, we faced a problem of file entry deletion when several audios had the same name but not the same path (the name of the files being the key/entry of the json). So we chose to rename the files by their name + "_number", for example "STE_015.mp3_3711".

Uses

37. Has the dataset been used for any tasks already? *If so, please provide a description.*

The corpus was used for the LeBenchmark project to train a self-supervised Wav2vec 2.0 models in French, considering its only speech without labels or transcripts.

38. Is there a repository that links to any or all papers or systems that use the dataset? *If so, please provide a link or other access point.*

See LeBenchmark website : http://lebenchmark.com/.

39. What (other) tasks could the dataset be used for?

This dataset was only made for pretraining machine learning model, but multiple downstream tasks can be considered.

40. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? *For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g. stereotyping, quality of service issues) or other risks or harms (e.g. legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

The corpus is provided as is, so any issues that would come from the content of the books are to be considered.

41. Are there tasks for which the dataset should not be used? *If so, please provide a description.*

No.

42. Any other comments?

None.

Distribution

43. Will the dataset be distributed to third parties outside of the entity (e.g. company, institution, organization) on behalf of which the dataset was created? *If so, please provide a description.*

The audios of this dataset are already available but it is meant to be uploaded on a corpus repository.

44. How will the dataset will be distributed (e.g. tarball on website, API, GitHub)? *Does the dataset have a digital object identifier (DOI)?*

We are planning to make this dataset available on OpenSLR as a set of zip archives. At the time of writing, it does not have a DOI.

45. When will the dataset be distributed?

During the year 2023.

46. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? *If so, please describe this license and/or ToU, and provide a link or* *other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

Each audio is subject to its own copyright, indicated in the metadata. It is necessary to respect this one because it protects the author’s right of the book read. The corpus does not have a unique copyright.

47. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these*

restrictions.

The restrictions that have been imposed are the CC licenses that differ from speaker to speaker (Public domain, Share alike, No modification, No commercial use or a combination of the above). However, all the books read are in the public domain so it is only the license chosen by the speaker on his voice that will restrict the use of the audiobook.

48. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

No.

49. Any other comments?

None.

Maintenance

50. Who will be supporting/hosting/maintaining the dataset?

This corpus is going to be hosted by OpenSLR.

51. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The curator of the dataset, François Portet (Professor of computing science at University Grenoble Alpes), can be contacted at francois.portet@imag.fr.

52. Is there an erratum? If so, please provide a link or other access point.

No.

53. Will the dataset be updated (e.g. to correct labeling errors, add new instances, delete instances)? *If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?*

The dataset can be updated for correction. For instance, if the speaker wishes to have his/her personal audio and data deleted, the corpus maintenance team is committed to take care of it, according to the moral right of the Code of the intellectual property, "the author will always be able to be opposed to the exploitations attacking his honor or his reputation or to the modifications denaturing his work" and because the CC are not framed in the time "the author can thus withdraw his license at any time".

54. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? *If so, please describe these limits and explain how they will be enforced.*

N/A.

55. Will older versions of the dataset continue to be supported/hosted/maintained? *If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

No.

56. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? *If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

There is no specific mechanism to do so. If others want to extend significantly the corpus, they are free to do so and to find their own hosting. Our team is, of course, open to discussion about how to do this in the most relevant way. There is no mechanism to inform potential users of the dataset. If the dataset is going to be superseded by another one, we will inform of such event on the original download platform so that potential users can be redirected to the most recent dataset.

57. Any other comments?

None.

* Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92.

[1] The gender information was inferred from the names of the authors, with verification upon listening in case of ambiguity. There is no guarantee of accuracy concerning that information.