People's Speech NeurIPS 2021 Datasets and Benchmarks Track Submission Supplementary Materials

Table of Contents

The People's Speech Dataset Datasheet

The original questions are in **bold**. The subtext to each question is in *italics.* The answers are in plain text with no formatting. The questions were copied from https://arxiv.org/pdf/1803.09010.pdf

Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.

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 People's Speech Dataset is a supervised speech recognition dataset created with three goals in mind:

1. Be at a scale rivaling the dataset sizes used internally at companies to train productions models. This means the dataset should be in the tens of thousands of hours of audio.
2. Legally permit commercial usage. Concretely, this means that we download only known-public-domain works, CC-BY-licensed, and CC-BY-SA-licensed works.
3. Begin to answer questions about the diversity of human speech generally available on the Internet under licenses permitting commercial use.

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

The dataset was funded wholly by MLCommons.

Any other comments?
No

Composition

Most of these questions are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for specific tasks. The answers to some of these questions reveal information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

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 instances represent audio or video documents with transcripts uploaded to The Internet Archive (archive.org).

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

A total of 76,503 audio or video documents were used., for a total of 52,500 hours.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
It is a sample from the larger set. Ideally, we would download all appropriately licensed audio data with transcripts on the Internet, but this is hard to do. Instead, we download data from services that allow searching by the license of uploaded data. In particular, we download from archive.org.

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).*

The larger set is all audio of the Internet with transcripts that permits commercial usage. It is had to know whether or not the sample is representative

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

The raw data consists audio waveforms that you can listen to, along with transcripts that you can read.

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

Yes, the transcript of a particular audio file acts as the "label".

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.

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.

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.

We do not recommend creating a test or validation dataset from the dataset at this time for two reasons:

- We don't do data deduplication, so the same instance may appear in both the train and test or validation set.
- We don't have a way to detect whether the same speaker appears in two separate audio files. It is considered bad practice in speech recognition research for the same speaker to appear in both the train and test or dev set.

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

We do not know the origin of the transcripts for our data. It is well known that untrained transcriptionists do not transcribe . In addition, it is possible to some of our transcripts may have come from existing speech recognition systems.
There is acoustic noise in our dataset, relative to other datasets. This is considered a good thing, as we view the acoustic environment as one axis of diversity for speech data. Please see section 3.3 of our paper.
There is almost certainly redundancy in our dataset. There is nothing preventing two users from uploading the same document twice to archive.org. We note that there are ways to detect similar audio documents (e.g., for recognizing copyrighted music in videos), so this is an area for improvement.

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 future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. All data from external sources is downloaded kept within MLCommons's archives.

There are license restrictions on our data. We intend to release the public domain, CC-BY, and CC-BY-SA data under a single CC-BY-SA licensed dataset.

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.

We do not audit the dataset for potentially confidential information. However, it is unlikely that someone would upload their own data under a CC-BY or CC-BY-SA license (or otherwise declare a work as public domain) without understanding the implications of their doing so.

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

Yes, for example, this instance in our dataset discusses whether the Quran condones ISIS's practice of chopping off the hands and feet of thieves. While the video is not threatening, it is

understandable that the subject matter may cause anxiety. Since one of our goals was to quantify the diversity of audio on the Internet today, we erred on the side of including potentially offensive content.

Does the dataset relate to people?
If not, you may skip the remaining questions in this section.

Yes. All of our data comes from the voices of real people.

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 do not do this.

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. Several of our sources are legal and government proceedings, spoken histories, speeches, and so on. Given that these were intended as public documents and licensed as such, it is natural that the involved individuals are aware of this.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial 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.

Yes. We cannot control what people say in these audio recordings. Discussions about removing potentially sensitive content led to concerns about removing under-represented accents associated with particular racial or ethnic groups. We note that accented speech is under-represented in speech recognition datasets today. Finally, we note again that, by declaring their data as public domain, CC-BY-licensed, or CC-BY-SA-licensed, the creators are aware of how their data can be used.

Any other comments?

No

Collection process

[T]he answers to questions here may provide information that allow others to reconstruct the dataset without access to it.

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 data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Data was downloaded via the archive.org API. No data inference was done.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?
How were these mechanisms or procedures validated?

We wrote software to use archive.org's public APIs to retrieve our data.

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

We simply downloaded all the data available from archive.org, with the assumption that it would be representative of audio data available on the rest of the Internet. However, it may be biased towards the sorts of audio that people who wish to archive data choose to upload.

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

No people were involved in the data collection process. The authors simply downloaded data that already existed.

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 data was collected around the week of March 7th, 2021. This timeframe does not match the creation timeframe of the data. Data has been archived by archive.org since its creation in 1996. Even then, much of the data is older. For example, audio works that are in the public domain because of copyright expiration are from 1925 at the earliest.

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

Does the dataset relate to people?
If not, you may skip the remainder of the questions in this section.

Yes

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

Via third parties (the aforementioned websites).

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.

For the most part, yes. There are three cases to consider:

- Government works are public domain. Government employees and participants in such works are assumed to be aware of this.
- CC-BY and CC-BY-SA licenses provide implicit consent by the copyright owner for anyone else to “share” and “adapt” their work.
- The creators of works that are public domain because their copyright has expired may not have provided consent to the use of their data. However, it is legal precedent in the USA that free use of works with expired copyright is an overall public good.

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).

Yes, we provide a way for copyright owners to request that their data be removed from the dataset. In particular, it is possible that, when they chose to license their work in a particular way, they had not considered that that would allow their work to

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

Any other comments?

No

Preprocessing/cleaning/labeling

The questions in this section are intended to provide dataset consumers with the information they need to determine whether the “raw” data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a “bag-of-words” is not suitable for tasks involving word order.

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 remainder of the questions in this section.

“Forced-alignment” of the data to create smaller chunks of audio for training was done. This “cleans” the data as well by removing chunks of audio with poor transcripts. Please refer to section 4.2 of our paper.

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. Those who are interested may request access to the google cloud storage drive containing the raw data: gs://the-peoples-speech-west-europe/archive_org/Mar_7_2021

Is the software used to preprocess/clean/label the instances available?
If so, please provide a link or other access point.

Yes. The code is open source https://github.com/mlcommons/peoples-speech Please see our paper for a description of the exact scripts run to preprocess the dataset. In particular, the forced alignment system is run via this script: https://github.com/mlcommons/peoples-speech/blob/main/galvasr2/align/spark/align_cuda_decoder.py

Any other comments?

No

Uses

These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

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

Yes, for training speech recognition models.

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.

No.

What (other) tasks could the dataset be used for?
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 future user 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 undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

It could be used for speech synthesis. However, this requires careful cleaning of the dataset, as background noise is not tolerable for speech synthesis.

The dataset could be used for keyword spotting tasks as well. In particular, this is good use case for the non-English audio in the dataset.

Our sincere hope is that the large breadth of sources our dataset incorporates reduces existing quality of service issues today, like speech recognition system's poor understanding of non-native English accents. We cannot think of any unfair treatment that come from using this dataset at this time.

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

Please do not use this dataset for cloning particular people's voices without their permission. Also, please do not train speech recognition or speaker identification models on the dataset for the purpose of surveillance.

Any other comments?

No

Distribution

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.

Yes. The dataset will be publicly available under CC-BY-SA license.

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

Please read: https://github.com/mlcommons/peoples-speech/blob/main/docs/data_access.md
This may change in the future. There is no DOI at this time.

When will the dataset be distributed?
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.

The dataset will be distributed before NeurIPS 2021. It will be under a CC-BY-SA license.

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.

No

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

Any other comments?

No

Maintenance

These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.

Who is supporting/hosting/maintaining the dataset?
How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
Is there an erratum?
If so, please provide a link or other access point.

MLCommons's dataset working group handles hosting and maintenance. Please contact greg@mlcommons.org with questions. Instead of an "erratum", we plan to publish updates to the emails that people use to request the dataset.

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 users (e.g., mailing list, GitHub)?

Yes. It will be updated on an as-needed basis, with updates sent to all emails provided by users who request data access.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were 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.

No

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 users.

No. In the case that some data needs to be removed for legal or ethical reasons, we do not want to keep maintaining that data.

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 other users? If so, please provide a description.

Absolutely. Our entire pipeline to create the dataset is open source: https://github.com/mlcommons/peoples-speech
If someone would like to contribute directly back to The People's Speech, we recommend joining the working group meetings (instructions here).

Any other comments?

No

Data Access

Please email greg@mlcommons.org for access to the dataset. It is a manual process to approve people to download it right now.

We fully expect to make the dataset publicly accessible before Neurips 2021. Public access will be hosted accessible from https://mlcommons.org/.

Documentation on how to access the data and the data's format is available on github: https://github.com/mlcommons/peoples-speech/blob/main/docs/data_access.md

Statement of Responsibility

The authors hereby declare that they bear all responsibility for violations of rights and that this dataset is CC-BY-SA-licensed.

Hosting and Maintenance Plan

MLCommons will host the dataset and handle maintenance concerns like data removal in case of misuse of data. It is a well-funded non-profit with no concerns about its ability to deliver on these requirements. Currently, the data can be downloaded from a Google Cloud Storage bucket for which MLCommons pays for download bandwidth costs on behalf of downloaders.

Persistent Dereferenceable Identifier

MLCommons has applied for an Identifiers.org prefix for MLCommons datasets and other artifacts. Application is pending, but should resolve prior to publication.