Augmented Datasheet for Speech Datasets

Dataset	Paper	Link
Common Voice	https://arxiv.org/pdf/1912.06670.pdf	Common Voice (mozilla.org)

1. Motivation

What is the speech dataset name, and does the name accurately describe the contents of the dataset?

The full title “Common Voice: A Massively-Multilingual Speech Corpus” suggests that the focus is on multiple languages, but the short title “Common Voice” doesn’t show this meaning. The term “common” has the connotation that this includes crowdsourced speech from everyday speakers.

Can the dataset be used to draw conclusions on read speech, spontaneous speech, or both?

Read speech only.

Describe the process used to determine which linguistic subpopulations are the focus of the dataset.

The major linguistic subpopulations in Common Voice are 29 different languages. This dataset was designed in response to the current state of affairs in speech technology, in which training data is unavailable for most languages, when it should be decentralized. The aim of this project is to collect many different languages, both major and minor.

2. Composition

How many hours of speech were collected in total (of each type, if appropriate), including speech that is not in the dataset? If there was a difference between collected and included, why?

2,500 hours were collected and 2,000 hours were included in the dataset, after the audio-transcription validation process.

How many hours of speech and number of tokens are in the dataset (by each type, if appropriate)?

The dataset is constantly growing. As of May 03, 2023, there are 17,690 hours of speech from 108 languages. The table containing the number of unique clips is included in the paper.

Are there standardized definitions of linguistic subpopulations that are used to categorize the speech data? How are these linguistic subpopulations identified in the dataset and described in the metadata?

The terms ‘language’ or ‘accent’ are not explicitly defined, though regional languages are included (e.g. Welsh, Chinese from Taiwan, or Breton). Additional metadata includes optional self-reported speaker attributes of age, gender, and accent.

For any linguistic subpopulations identified in the dataset, please provide a description of their respective distributions within the dataset.

The table containing specific numbers of hours (both collected and validated) for each language is included in the paper. The creators also mention that the range of distributions per language is very wide and poses a challenge.

How much of the speech data have corresponding transcriptions in the dataset?

All speech data have a corresponding transcription.

Does the dataset contain non-speech mediums (e.g. images or video)?

No.

Do speakers code switch or speak multiple languages, and if so, how is this identified in the data?

The given read speech prompts are supposed to be monolingual with no code-switching. The speakers may be multilingual, as there is no barrier to contribution.

Does the speech dataset focus on a specific topic or set of topics?

This is not mentioned. Sentences are extracted from Wikipedia.

Does the dataset include sensitive content that can induce different emotions (e.g., anger, sadness) that can cause the speakers to produce unusual pitch or tone deviating from plain speech?

This is not mentioned. Sentences are extracted from Wikipedia; the nature of the sentences could induce different emotions in different speakers.

Does the dataset contain content that complies to the users' needs, or does it result in symbolic violence (the imposition of religious, political, cultural values etc.)?

This is not mentioned. Sentences are extracted from Wikipedia; the nature of the sentences could induce different reactions in different lingual groups.

3. Collection Process

What mechanisms or procedures were used to collect the speech data, e.g.: is the data a new recording of read speech or an interview? Or is it downloaded speech data from public speeches, lectures, YouTube videos or movies, etc.?

Participants need to record themselves reading a sentence through a custom recording interface.

Were all the data collected using the same technical methodology or setting, including the recording environment (e.g., lab, microphone) and recording information (e.g., sampling rate, number of channels)?

Data has been crowdsourced, which means a variety of devices and microphones are used for recording. Audio clips are standardized when released (i.e., mono-channel, 48kHz sampling, 16bit).

Is there presence of background noise?

Not mentioned, assumed to have background noise since it is crowdsourced data.

For interviewer/interviewee speech data: during the interview process, did interviewers consistently ask questions that are ``fair and neutral''?

Not applicable. This is a read speech dataset.

Have data subjects consented to the disclosure of the metadata in the dataset? Also, does the metadata include sensitive personal information such as disability status?

Only age, gender and accent are self-reported, and these are optional. There is a mention of metadata being disclosed, after being anonymized.

4. Preprocessing/cleaning/labeling

When generating the dataset, was any background noise deleted or adjusted to make all recording qualities similar?

Not mentioned.

Did the data collectors hire human annotators to transcribe the data? If so, how trained were the annotators in speech transcription for this context? How familiar were they with the corpus material, the vocabulary used, and the linguistic characteristics of different dialects and accents?

Actual transcription is not necessary as the speakers are given prompts. Rather, crowdsourced contributors use a specifically devised application to vote on audio-transcription pairs for validation. They were not trained nor did not need to be native speakers of the language. Because the dataset is crowdsourced, there is no barrier to contributing to the audio-transcription validation process.

If multiple transcription methods were used, how consistent were the annotators? How were transcripts validated?

Not applicable because this is a read speech dataset. However, for validation of audio-transcription pairs via crowdsourced contributors: three listeners validate the speech, and two of them need to approve it to be validated.

If the speech data include transcriptions, what software was used to generate the transcriptions (including, e.g., software used by human transcribers)? Are timestamps included in transcriptions? Are the alignments provided with the transcripts?

Actual transcription is not necessary as the speakers are given prompts; other than the validation described above, no other tools are used. Timestamps and alignments are not provided.

Were transcription conventions (such as tagging scheme, treatment of hate speech or swear words, etc.) disclosed along with the corpus?

Not mentioned. Sentences are taken from Wikipedia, and could potentially include hate speech or swear words; however, this is not verified.

Is additional coding performed, separate to transcriptions and tagging?

No additional coding is performed.

5. Uses / Distribution / Maintenance

How are redactions performed on the dataset? Are personally identifiable information or sensitive information removed from only transcripts, audio censored from the speech data, or both?

Not applicable. This is a read speech dataset. That said, given that sentences are from Wikipedia, it can be assumed that sentences do not contain private information.

Is there any part of this dataset that is privately held but can be requested for research purposes?

No.

Is there a sample dataset distributed? If so, how well does the sample represent the actual dataset? Do they include all forms of speech included in the dataset? How big is the sample?

There is not a separate sample dataset. The interface for recording and listening provides samples in all supported languages.

Aside from this datasheet, is other documentation available about the data collection process (e.g., agreements signed with data subjects and research methodology)?

Yes. There is a research paper and a website with more information, as well as an online forum (https://discourse.mozilla.org/c/voice/239) for support.