Augmented Datasheet for Speech Datasets

Dataset	Paper	Link
LibriSpeech	Librispeech: An ASR corpus based on public domain audio books | IEEE Conference Publication | IEEE Xplore	openslr.org

1. Motivation

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

Librispeech; no the name does not give us any detail on the dataset, though it does allude to books.

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

Read Speech (of audio books)

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

The paper has two linguistic subpopulations: it includes (1) a split for speaker sound quality and proximity to US English and (2) balances for gender at the speaker level in terms of the amount of data available for each gender. It is unclear why the authors selected these two subpopulations.

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?

Not mentioned but could be answered by author (e.g., how much speech was discarded during the alignment process).

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

1000 hours of speech, 803 million tokens in total and 900 000 unique words after initial cleaning.

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 dataset includes a “clean” subset, which refers to speakers with higher recording quality and accents closer to US English, and an “other” dataset. They create these splits by computing the word error rate of the automatic transcript of the speech sample with the model transcript on the WSJ’s si-84 data subset. Lower WER speakers were part of the “clean” split and higher WER a part of the “other” split. No further qualitative detail is provided on what characteristics are shared across the higher WER speakers.

The authors do not provide specific details on how speaker gender was determined or defined, but allude to a custom GUI application through which gender information is produced. The gender is noted as “sex” (either M or F) in the metadata for each speaker.
The terms ‘accent’ or ‘US English’ and the methodology for which the creators determined the accent were not defined. The authors similarly allude to an automatic procedure, but the pipeline through which the automatic classification was done was not explained. It does not appear that either sets of demographic labels are self-reported.

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

In total, the distribution splits are either a 363.6 hour or 100.6 hour training set, 5.4 validation set, and 5.4 hour test set for the “clean” subset. There are 486.7 hours for training, 5.3 hours for validation, and 5.1 hours for test in the “other” subset (with high WER, i.e. with poorer sound quality and non-English).

For gender, the distribution split is approximately balanced (i.e., 500 hours for male speakers and 500 for female speakers).

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

All of the speech data has a reference text that the speaker is reading from.

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

No; the dataset contains only audio data.

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

No; the dataset is only in English and consists of reading from audiobooks. There is no encoding of accent-switching by the narrator.

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

It spans the topics across 8,000 public domain audiobooks.

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?

Unclear. The dataset contains recordings of audiobooks, so it is possible that readers will adapt different intonations via voice acting based on content of the book. However, this is not made explicit.

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

Unclear. The paper does not make any direct mention of this; however, on Librivox, they make the disclaimer that “our readers and listeners should be aware that many of them are very old, and may contain language or express notions that are antiquated at best, offending at worst.”

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.?

The speech data is derived from LibroVox project, which consists of audiobooks.

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)?

No, the audio is sourced from LibriVox which crowdsources audiobooks without any restrictions on the technical methodology for recordings, sampled at 16kHz. This is unclear in the original paper.

Is there presence of background noise?

It is not mentioned in the paper whether there is background noise.

However, considering that the LibriVox contains crowdsourced audiobook recording with no detailed stringent process for removing background noise, it is likely there is background noise present in the recordings.

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

N/A; the speech data is not collected via interviews.

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?

Gender metadata is provided.

4. Preprocessing/cleaning/labeling

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

Not mentioned in the paper. The original author should be able to answer this question.

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?

No human annotators were used to transcribe the data.

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

N/A; multiple transcription methods were not used.

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?

The transcriptions of the audiobook came from texts available on Project Gutenberg. To format the text, they were converted to upper-case, removed punctuation, and expanded common abbreviations or non-standard words.

The alignment process was two stage: the first consisted of the Smith-Waterman algorithm to help split the audio into shorter segments and the second to filter out segments that are likely to be inaccurate.

The alignments are provided with the transcripts.

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

No; transcription conventions were not disclosed along with the corpus.

Is additional coding performed, separate to transcriptions and tagging?

No

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?

N/A; redactions are not performed on the dataset (or not mentioned in the paper).

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?

No, the dataset is released and fully accessible on a project page. There is a version of the clean training set which contains 100 hours, a smaller subset than the 360 hours.

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

No; documentation does not exist.