# **Augmented Datasheets for Speech Datasets**

**Dataset name:**

**Dataset link:**

## **Motivation**

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

|  |
| --- |
|  |

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

|  |
| --- |
|  |

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

|  |
| --- |
|  |

1. Any other comments?

|  |
| --- |
|  |

## 

## **Composition**

1. 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? E.g., if the speech data are from an interview and the dataset contains only the interviewee's responses, how many hours of speech were collected in interviews from both interviewer and interviewee?

|  |
| --- |
|  |

1. How many hours of speech and number of speakers & words are in the dataset (by each type, if appropriate)?

|  |
| --- |
|  |

1. 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?

|  |
| --- |
|  |

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

|  |
| --- |
|  |

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

|  |
| --- |
|  |

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

|  |
| --- |
|  |

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

|  |
| --- |
|  |

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

|  |
| --- |
|  |

1. 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?

|  |
| --- |
|  |

1. 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.)?

|  |
| --- |
|  |

1. Any other comments?

|  |
| --- |
|  |

## 

## **Collection Process**

1. 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.?

|  |
| --- |
|  |

1. 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)?

|  |
| --- |
|  |

1. Is there presence of background noise?

|  |
| --- |
|  |

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

|  |
| --- |
|  |

1. 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?

|  |
| --- |
|  |

1. Any other comments?

|  |
| --- |
|  |

## **Preprocessing/ cleaning/ labeling**

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

|  |
| --- |
|  |

1. 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?

|  |
| --- |
|  |

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

|  |
| --- |
|  |

1. 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?

|  |
| --- |
|  |

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

|  |
| --- |
|  |

1. Is additional coding performed, separate to transcriptions and tagging?

|  |
| --- |
|  |

1. Any other comments?

|  |
| --- |
|  |

## **Uses/ Distribution/ Maintenance**

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

|  |
| --- |
|  |

1. 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.

|  |
| --- |
|  |

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

|  |
| --- |
|  |

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

|  |
| --- |
|  |

1. 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?

|  |
| --- |
|  |

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

|  |
| --- |
|  |

1. Any other comments?

|  |
| --- |
|  |

## 

## **Uses/ Distribution/ Maintenance**

1. 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?

|  |
| --- |
|  |

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

|  |
| --- |
|  |

1. 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?

|  |
| --- |
|  |

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

|  |
| --- |
|  |

1. Any other comments?

|  |
| --- |
|  |