Augmented Datasheets for Speech Datasets and Ethical Decision-Making

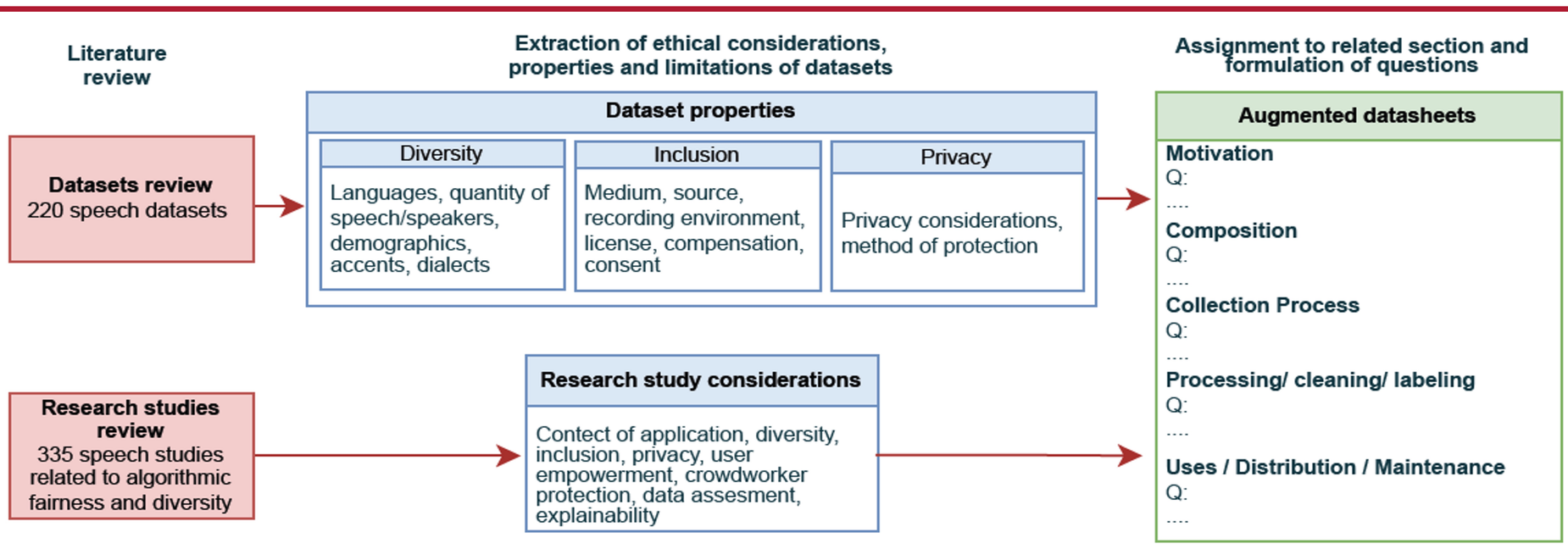
Orestis Papakyriakopoulos[1], Anna Seo Gyeong Choi[2], Alice Xiang[1], Allison Koenecke[2]
1: Sony AI, 2: Cornell University

Sony AI

Research Motivation

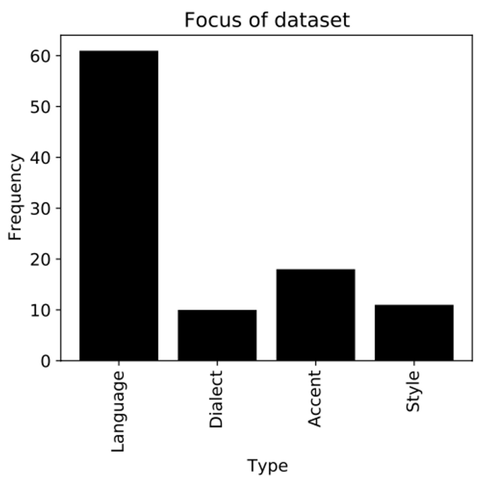
- How can we better document and diversify speech data that is used to train speech language technologies?
- We augment Gebru et al.'s "Datasheets for Datasets" to (a) advocate for standardized documentation of speech-specific features and (b) make linguistic diversity more transparent

Methodology

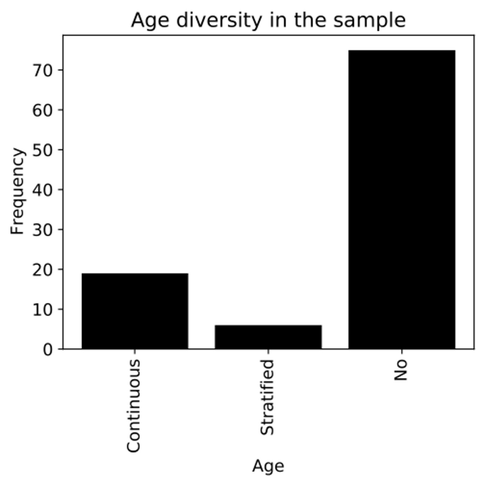


Diversity Representation

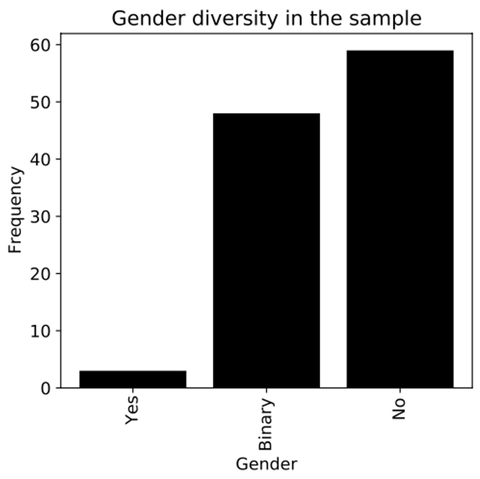
- Descriptive statistics of a variety of diversity properties reviewed from the speech datasets.



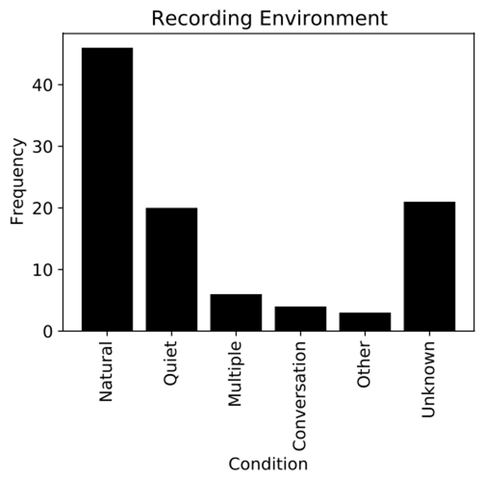
(a) Linguistic diversity



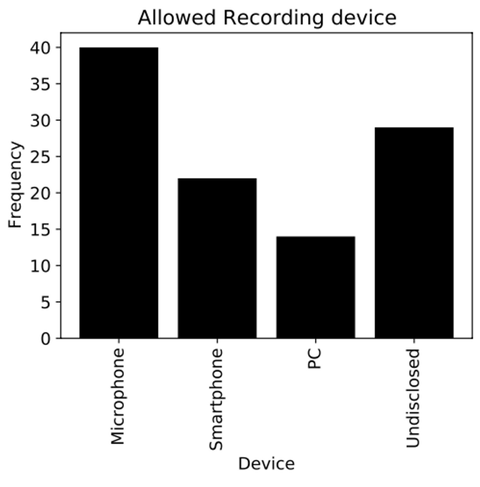
(b) Age diversity



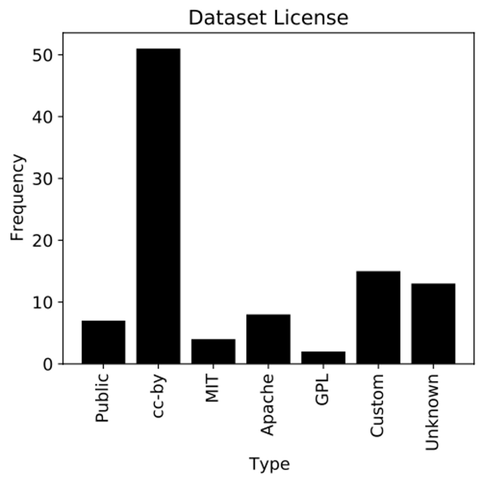
(c) Gender diversity



(d) Recording environment



(e) Recording device



(f) License

Sample Augmented Questions

- **Motivation**
  - Describe the process used to determine which linguistic subpopulations are the focus of the dataset.
- **Composition**
  - How many hours of speech, number of speakers & words are in the dataset?
  - Are there standardized definitions of linguistic subpopulations that are used to categorize the speech data? How are these linguistic subpopulations identified int eh dataset and described in the metadata?
  - How much of the speech data have corresponding transcriptions in the dataset?
  - Does the dataset contain non-speech mediums?
  - Do speakers code switch or speak multiple languages?
- **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 from public speeches, lectures, YouTube videos or movies, etc.?
  - Is there presence of background noise?
- **Preprocessing/Cleaning/Labeling**
  - If multiple transcription methods were used, how consistent were the annotators? How were transcripts validated?
  - Is additional coding performed, separate to transcriptions and tagging?
- **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?
  - Aside from this datasheet, is there other documentation available about the data collection process?