9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Augmented Datasheets for Speech Datasets and Ethical Decision-Making

Keywords: datasets, speech language technologies, datasheets, ethical decision-making, algorithmic fairness

Extended Abstract

Speech datasets are crucial for training Speech Language Technologies (SLT)—including speech-to-text systems, speaker recognition, speech synthesis, speech enhancement and denoising—now often integrated in smartphones, cameras and virtual-assistants, and applied in domains across customer service, finance, navigation, health, education, and law. However, the lack of diversity of the underlying training data can lead to serious limitations in building equitable and robust SLT products, especially along dimensions of (a) language, accent, dialect, variety, and speech impairment; (b) socioeconomic and demographic features; and (c) the intersectionality of these sets of features. Furthermore, there is often a lack of oversight on the underlying training data—commonly built on massive web-crawling and/or publicly available speech—with regard to the ethics of such data collection.

Our aim is to encourage standardized documentation of speech data components, in order to make dataset creation more transparent and accessible, while also assisting dataset users—such as SLT practitioners and researchers—to select and combine appropriately diverse datasets for their objectives. We follow in the footsteps of "Datasheets for Datasets" [2] by introducing an **augmented datasheet for *speech* datasets**. This augmented datasheet can be found at `https://anonymous.4open.science/r/augmented_datasheets-0E5F/`, which provides both datasheet templates and worked examples for common speech datasets [1].

Motivation Speech technologies—as with any machine learning applications—are prone to bias; these biases often stem from nonrepresentative or inaccurate training data. The downstream impacts of biased SLT can be severe. Individuals who do not speak "standard" varieties of a language may be disproportionately unlikely to be hired given speech-based hiring screening software. Doctors increasingly use SLT to efficiently take patient notes, which could result in serious health harms if transcribed incorrectly. Moreover, SLT technology has been developed to surveil the phone calls of incarcerated individuals; the resulting transcriptions—likely disproportionately inaccurate for Black individuals [3]—can result in differential treatment.

There is high complexity in the collection of speech data used to train SLT applications, with variability across necessary tasks: from noise in a recording environment to transcription of language and acoustic features. Additionally, as with any data collection, it is imperative to center ethical dataset creation and usage regarding the privacy, respect, and protection of data subjects, interviewers, and transcription annotators. It is hence useful for speech dataset *creators* to standardize documentation of speech data collection processes for reproducibility and transparency. Furthermore, speech dataset *users* should have access to such data collection details, as they are often the ones using the data to train SLT. Generating SLT that are robust across speakers' linguistic and demographic features necessarily involves identifying and using (or knowing not to use) different & diverse datasets.

Datasheet Questions The questions comprising our "augmented datasheet" for speech datasets are formulated by the authors' positionalities as SLT practitioners, linguists, machine learning

researchers, algorithmic fairness researchers, and lawyers; we also draw upon an in-depth literature review revealing data-centric best practices and issues in SLT, wherein we review 179 research studies and 220 speech datasets. In this review, we find that a minority of studies write about data-centric best practices with regard to diversity, and less than 10% of studies note privacy best practices. Furthermore, a minority of speech datasets focus on dialects, accents, or linguistic styles within a language; and, many speech datasets have unknown licensing—leading to concerns about data participants' speech data privacy and distribution.

Our augmented datasheet questions focus on speech datasets specifically, and are categorized in five categories: Motivation (3 questions), Composition (10 questions), Collection Process (5 questions), Preprocessing / Cleaning / Labeling (6 questions), and Uses / Distribution / Maintenance (4 questions). Full question text, ranging in topic from defining speaker accents to reporting microphone details, can be found at `https://anonymous.4open.science/r/augmented_datasheets-0E5F/`.

Calls to Action We believe our augmented datasheet template can guide researchers in ethical, robust, and inclusive speech dataset design and usage, and have released our template for immediate public use. We make a call to action—for all practitioners using speech data, whether dataset creators or users—to use speech-specific datasheets in a collaborative effort to ensure transparency regarding speech data ethics and diversity. The benefits to dataset *creators* include ensuring standardization of dataset documentation, enhancing transparency of dataset contents, clarifying the underlying motivations and process of data collection, and encouraging explicit consideration of underrepresented linguistic subpopulations and socioeconomic/demographic groups. The benefits to dataset *users* include more comprehensive understanding of dataset utility, and easier decision-making on data selection for more robust and inclusive SLT—especially for data on underrepresented groups.

Ethical dataset creation is not a one-size-fits-all process, but practitioners can use our augmented datasheet to reflexively consider the social context of related SLT applications and data sources in order to foster more inclusive SLT products downstream. We encourage SLT practitioners to view datasheets as a collaborative process, with their users, data subjects, and affected communities—especially in cases of datasets that focus on low-resource and vulnerable communities.

References

[1] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

[2] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[3] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, March 2020.

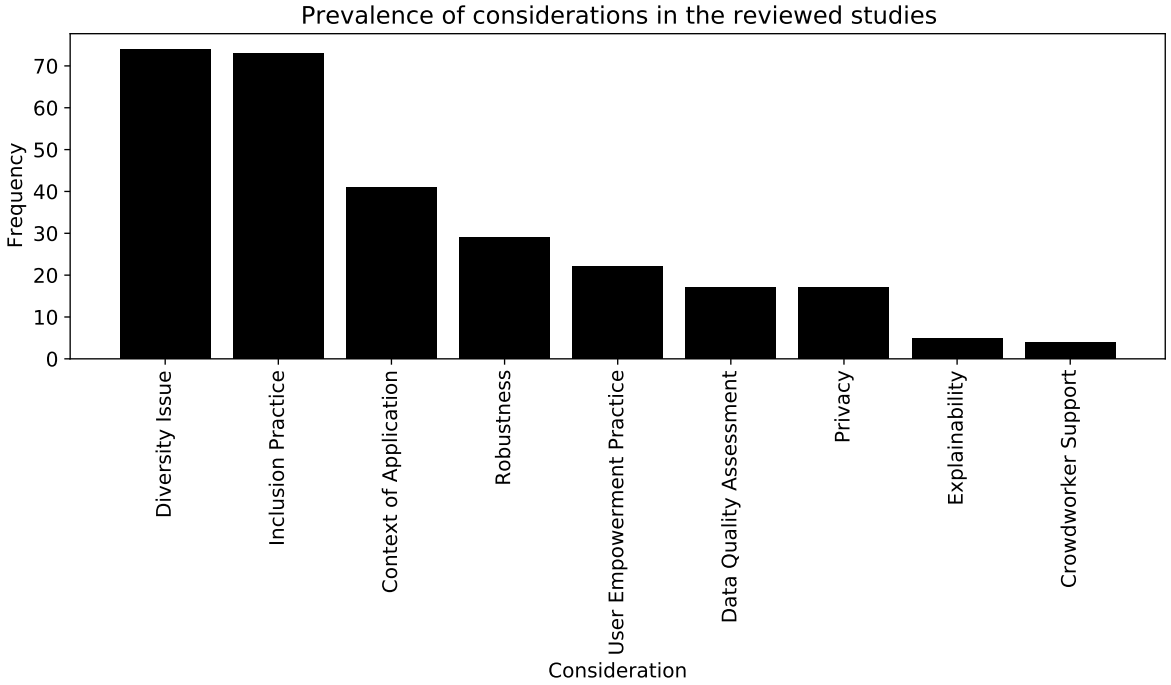
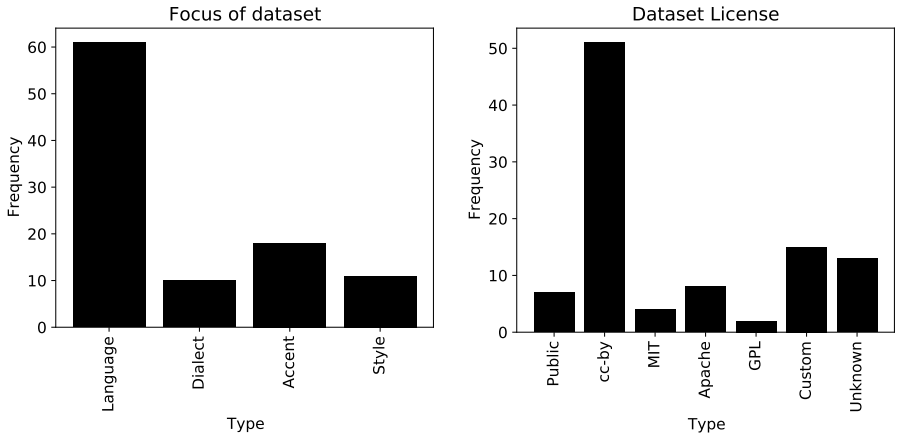


Figure 1: Distribution of ethical considerations in the reviewed sample of SLT studies (N=179). A study could contain more than one considerations.



(a) Linguistic diversity (b) License

Figure 2: Descriptive statistics in the reviewed sample of speech datasets.