



Allison Koenecke

@allisonkoe

Excited to announce our new @FACCTConference paper, “Augmented Datasheets for Speech Datasets and Ethical Decision-Making”, led by @SciOrestis and @anna_sg_choi (who are presenting today) & @alicexiang! Link: dl.acm.org/doi/10.1145/35...; 📄 📌 (1/9)

Decision-making

Orestis Papakyriakopoulos*
orestis.papakyriakopoulos@sony.com
Sony AI
Zurich, Switzerland

Anna Seo Gyeong Choi*
sc239@cornell.edu
Cornell University
Ithaca, New York, USA

Jerone Andrews
jerone.andrews@sony.com
Sony AI
Tokyo, Japan

Rebecca Bourke
rebecca.bourke@sony.com
Sony AI
Tokyo, Japan

William Thong
william.thong@sony.com
Sony AI
Zurich, Switzerland

Dora Zhao
dora.zhao@sony.com
Sony AI
New York, New York, USA

Alice Xiang*
alice.xiang@sony.com
Sony AI
Seattle, Washington, USA

Allison Koenecke*
koenecke@cornell.edu
Cornell University
Ithaca, New York, USA

ABSTRACT
Speech datasets are crucial for training Speech Language Technologies (SLT); 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 language, accent, dialect, variety, and speech impairment—and the intersectionality of speech features with socioeconomic and demographic 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. To encourage standardized documentation of such speech data components, we introduce an augmented datasheet for speech datasets¹, which can be used in addition to “Datasheets for Datasets” [78]. We then exemplify the importance of each question in our augmented datasheet based on in-depth literature reviews of speech data used in domains such as machine learning, linguistics, and health. Finally, we encourage practitioners—ranging from dataset creators to researchers—to use our augmented datasheet to better define the scope, properties, and limits of speech datasets, while also encouraging consideration of data-subject protection and user community empowerment. Ethical dataset creation is not a one-size-fits-all process, but dataset creators can use our augmented datasheet to reflexively consider the social context of related SLT

Extraction of ethical considerations, properties and limitations of datasets

Dataset properties

Diversity	Inclusion	Privacy
Languages, quantity of speech/speakers, demographics, accents, dialects	Medium, source, recording environment, license, compensation, consent	Privacy considerations, method of protection

Research study considerations

Context of application, diversity, inclusion, privacy, user empowerment, crowdworker protection, data assessment, explainability

Assignment to formulate

Augment
Motivation Q:
Composition Q:
Collection Proc Q:
Processing/ ch Q:
Uses / Distribu Q:

A.3. Collection Process

- How much of the speech data have corresponding transcriptions in the dataset?
- Does the dataset contain non-speech mediums (e.g. images or video)?
- Do speakers code switch or speak multiple languages, and if so, how is this identified in the data?
- Does the speech dataset focus on a specific topic or set of topics?
- 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?
- 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.)?

A.4. Preprocessing/cleaning/labeling

- 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.?
- 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)?
- Is there presence of background noise?
- For interviewer/interviewee speech data: during the interview process, did interviewers consistently ask questions that are “fair and neutral”?
- 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?
- When generating the dataset, was any background noise deleted or adjusted to make all recording qualities similar?
- 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?
- If multiple transcription methods were used, how consistent were the annotators? How were transcripts validated?

👤 Cornell Information Science and 2 others

12:15 PM · Jun 14, 2023 · 23.5K Views

💬 1 🔄 25 ❤️ 88 📌 16 📄 📌

Reply



Allison Koenecke @allisonkoe · Jun 14, 2023

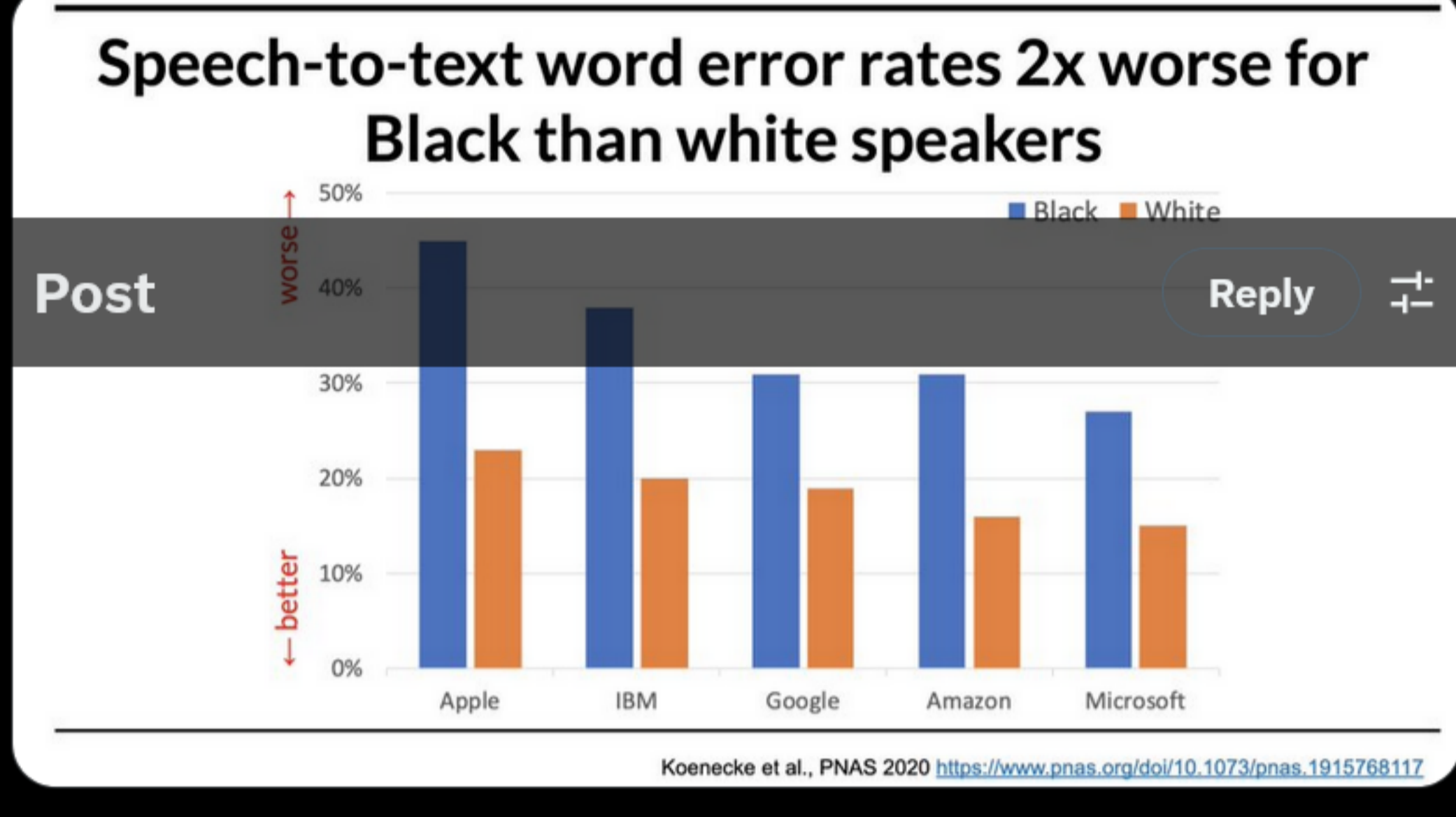
The timeless classic “Datasheets for Datasets” (by @TimnitGebru @JamieMmt @BrianaVecchione @JennWVaughan @HannaWallach @HalDaume3 @KateCrawford) proposes documentation of data provenance & uses. We extend this framework for speech data documentation. Why speech? (2/9)

💬 1 🔄 1 ❤️ 4 📌 439 📄 📌



Allison Koenecke @allisonkoe · Jun 14, 2023

Speech data are necessary for ML-based speech tech (eg. speech-to-text, used in hiring/doctors offices/courtrooms). Models not trained on enough African American English speakers → less accurate transcriptions for Black speakers → IRL disparities pnas.org/doi/10.1073/pn... (3/9)

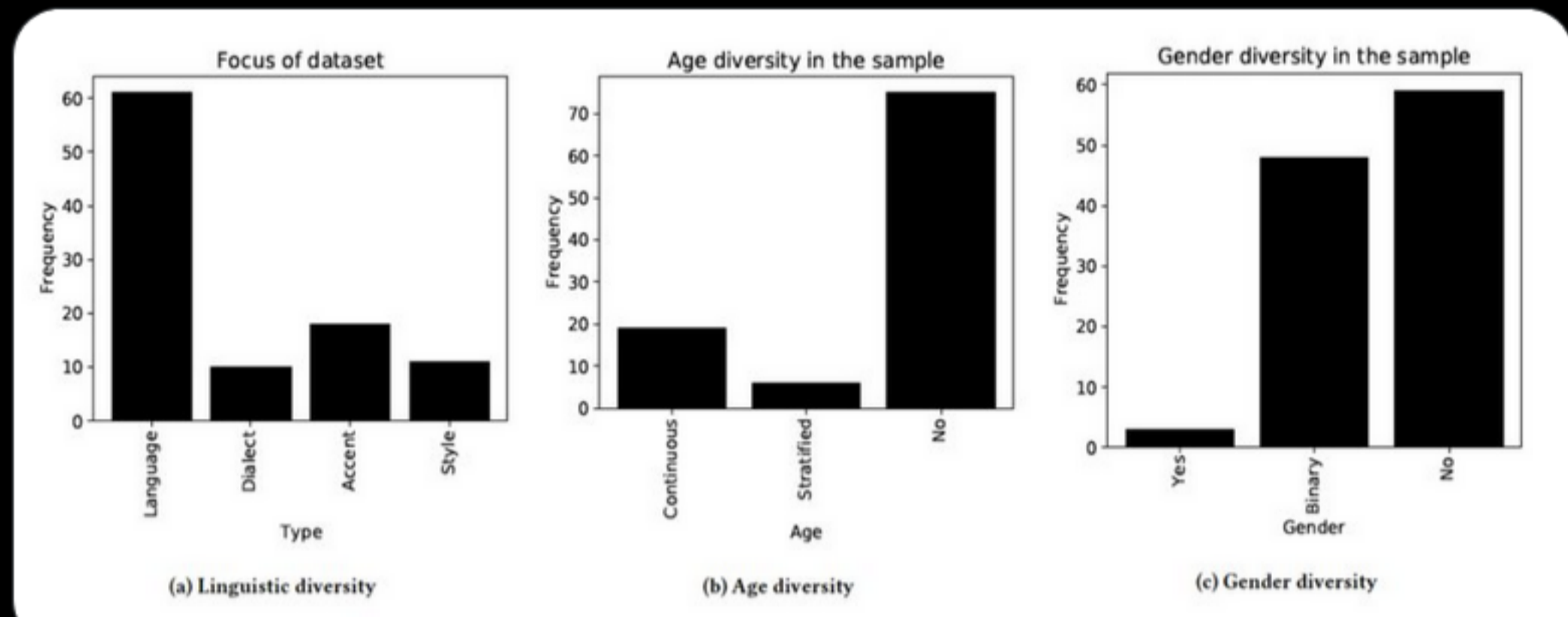


💬 1 🔄 1 ❤️ 1 📌 399 📄 📌



Allison Koenecke @allisonkoe · Jun 14, 2023

How can datasheets help? By documenting specifics for often-overlooked “linguistic subpopulations”: accents, dialects, varieties, languages, and speech arising from speech disorders or pathologies. Our lit review shows few datasets currently consider these axes of diversity (4/9)



💬 1 🔄 1 ❤️ 2 📌 309 📄 📌



Allison Koenecke @allisonkoe · Jun 14, 2023

Datasheets are useful to dataset creators: linguistic subpopulations should be considered before beginning data collection. Eg., why are you focusing on a specific language/region? Is there consensus on how you define a certain accent? Are speech types self-reported? (5/9)

💬 1 🔄 1 ❤️ 2 📌 181 📄 📌



Allison Koenecke @allisonkoe · Jun 14, 2023

Datasheets are also useful to dataset users: when combining multiple speech datasets to build models, do you have appropriate coverage across speech types? Our augmented datasheets provide a standardized way to check for linguistic make-up of speech data (6/9)

💬 1 🔄 1 ❤️ 1 📌 200 📄 📌



Allison Koenecke @allisonkoe · Jun 14, 2023

We also propose other speech-specific datasheet questions, such as background noise levels and technical recording equipment used, which are also both proxies for socioeconomic status and can yield more robust speech technology performance if included in the training data. (7/9)

💬 1 🔄 1 ❤️ 1 📌 227 📄 📌



Allison Koenecke @allisonkoe · Jun 14, 2023

We include both empty datasheet templates (.docx and .tex) AND worked examples of common speech datasets (CORAAAL, CommonVoice, LibriSpeech, VoxPopuli, WHAM) on our github: github.com/SonyResearch/p... Let us know what you think! (8/9)

SonyResearch/ project_ethics_augmente... Sony Research

Public code repo for research paper

👤 2 Contributors 0 Issues ⭐ 7 Stars 🍴 1 Fork

GitHub - SonyResearch/project_ethics_augmented_datasheets_for_speech_data...

From github.com

💬 1 🔄 1 ❤️ 1 📌 343 📄 📌



Allison Koenecke @allisonkoe · Jun 14, 2023

Many thanks to our research inspirations and members of the speech community! @alexhanna @MilagrosMiceli @blahtino @dylnbkr @vinodkpg @amironesei @cephaloponderer @Bridge2aiVoice @TweetRupal @kaaaaatok @mixedlinguist @AliciaWassink (9/9)

💬 1 🔄 1 ❤️ 📌 403 📄 📌



Allison Koenecke @allisonkoe · Jun 14, 2023

+ @CorpusAAL!!

💬 🔄 ❤️ 1 📌 287 📄 📌