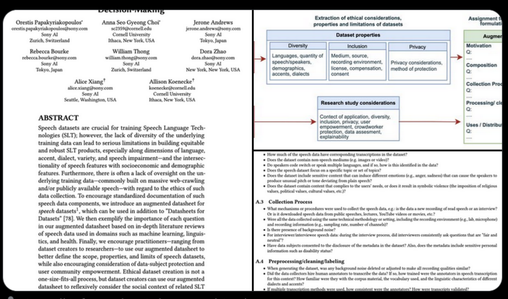
Allison Koenecke
@allisonkoe

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)



Cornell Information Science and 2 others

12:15 PM · Jun 14, 2023 · **23.5K** Views

1 · 25 · 88 · 16

Reply

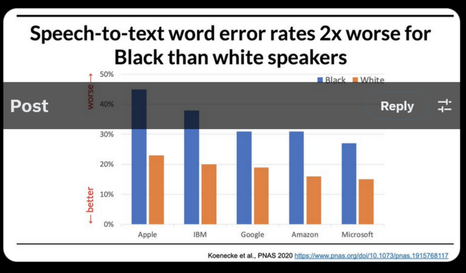
Allison Koenecke @allisonkoe · Jun 14, 2023

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

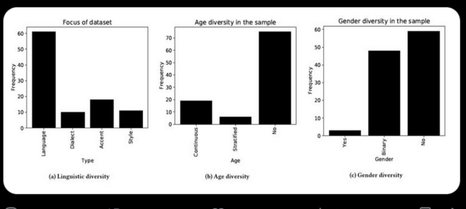
Speech data are necessary for ML-based speech tech (eg. speech-to-text, used in hiring/doctors offices/courtrooms). Models not trained on enough African American English speakers ➡️ less accurate transcriptions for Black speakers ➡️ IRL disparities pnas.org/doi/10.1073/pn... (3/9)



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

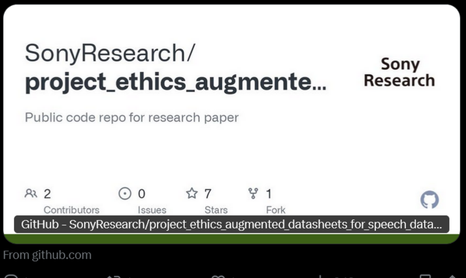
Allison Koenecke @allisonkoe · Jun 14, 2023

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 (CORAAL, CommonVoice, LibriSpeech, VoxPopuli, WHAM) on our github: github.com/SonyResearch/p... Let us know what you think! (8/9)



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