Who Decides if AI is Fair? The Labels Problem in Algorithmic Auditing

Abhilash Mishra[*†]
The University of Chicago & EquiTech Futures
abhilash@equitechfutures.com

Yash Gorana[*]
EquiTech Futures
yash@equitechfutures.com

Abstract

Labelled "ground truth" datasets are routinely used to evaluate and audit AI algorithms applied in high-stakes settings. However, there do not exist widely accepted benchmarks for the quality of labels in these datasets. We provide empirical evidence that quality of labels can significantly distort the results of algorithmic audits in real-world settings. Using data annotators typically hired by AI firms in India, we show that fidelity of the ground truth data can lead to spurious differences in performance of ASRs between urban and rural populations. After a rigorous, albeit expensive, label cleaning process, these disparities between groups disappear. Our findings highlight how trade-offs between label quality and data annotation costs can complicate algorithmic audits in practice. They also emphasize the need for development of consensus-driven, widely accepted benchmarks for label quality.

1 Introduction

There is now a compelling body of evidence that AI systems deployed in high-stakes domains like healthcare, criminal justice, and education can be biased against socioeconomically disadvantaged communities. Examples include bias in facial and object recognition (Buolamwini and Gebru, 2018; de Vries et al., 2019), speech (Koenecke et al., 2020), and healthcare systems (Obermeyer et al., 2019). These concerns have been heightened with the growing dominance of so-called foundation models (Bommasani et al., 2021) which can amplify biases across a wide range of applications through single points of failure.

Fairness and ethical concerns around AI have given rise to calls for greater transparency in the training data used to build the algorithms. These include proposals such as "datasheets" for training datasets (Gebru et al., 2018), model cards (Mitchell et al., 2019), and robust algorithmic auditing both by citizens and governments (Raji and Buolamwini, 2019).

But who audits the auditor? In this paper we show that poorly defined standards of label quality can lead to spurious results in algorithmic audits. The challenge of quality control in machine learning datasets is widely known (Paullada et al., 2020). Our work highlights that even in "small data" settings like algorithmic audits, label quality can distort results.

These pose new policy challenges for the deployment, auditing, and governance of AI systems, particularly in high-stakes settings. There are no universal benchmarks for label quality. In the context of algorithmic auditing, label quality is both a normative and economic question. Whose labels are better? What is an acceptable label quality? How much should we spend on creating high-quality labels for benchmarks? These are all highly practical and relevant questions about auditing AI algorithms in the real world.

---

[*]Both authors contributed equally.
[†]Corresponding author

In this paper we use a new voice dataset collected in India to demonstrate that fidelity of "ground truth data" can have serious implications on the results of algorithmic audits. Our dataset is collected from low-resourced settings in India. The Indian context is interesting for three reasons: technology adoption has grown exponentially in the last decade with over 740 million internet users and this number is projected to grow . The adoption of AI enabled applications in healthcare and education is particularly promising in many low-resource settings in India. Furthermore, India possesses a staggering diversity of languages, socioeconomic backgrounds which makes the design and governance of AI enabled services particularly challenging.

2 Related Work

Algorithmic audits have become a popular strategy to highlight scenarios where AI applications can be biased against specific sub-populations (Guszcza et al., 2018). For example, Koenecke et al. (2020) conducted an audit of state-of-the-art, "off the shelf" automated speech recognition (ASR) systems developed by Amazon, Apple, Google, IBM, and Microsoft. They found that all five ASR systems exhibit significantly lower accuracy for Black English speakers (word error rate = 0.35) than for White English speakers (word error rate = 0.19). There is promise that such audits can influence AI development and governance. For example, Raji and Buolamwini (2019) show that an algorithmic audit of gender and skin type performance in commercially available facial recognition models led to developers significantly improving the accuracy of these models within a year.

Our work is closely related to Koenecke et al. (2020). We set out to conduct an audit of commercially available ASR systems from three primary cloud vendors in India (Amazon, Google, and Microsoft). However we ask a different, more "upstream" data quality question: how does the quality of *human* transcriptions affect the audit of ASR systems? This is a critical question for AI governance because widely accepted benchmarks on label quality do not exist, especially for audit of AI systems in developing countries. As far as we know, this has not been highlighted in the literature.

While the focus of this paper is the need for clear label quality benchmarks needed for algorithmic audits, similar concerns about label quality have been recently discussed in the context of training data for ML algorithms. Tsipras et al. (2020) use human studies to show that specific design choices in image annotation pipelines can lead to significant misalignment between a benchmark (like ImageNet) and the "ground truth" it serves as a proxy for.

In this paper we empirically demonstrate how the issue of label quality can pose serious challenges in not just the development of new AI systems but auditing commercially available AI systems. This is a first order priority in AI governance that the data-centric AI movement needs to address.

3 Materials and Methods

3.1 Data

We collected a novel dataset of conversational speech in India from 112 high school and college students. Our sample includes students from rural and urban areas and all students speak English as a second language. Thus this data is ideal to evaluate if widely deployed ASR systems in India perform poorly compared to global benchmarks and if there is significant regional disparities.

We hired freelance transcribers (with an educational background similar to data annotators hired by AI labelling firms in India). These transcribers had prior experience in voice-to-text transcription. The transcriptions produced underwent three rounds of reviews ("iterations"):

Iteration 1: This consists of the raw transcriptions we received from the hired transcribers without any edits.

Iteration 2: This version consists of the research team at EquiTech Futures correcting instances where transcribers forced grammatical corrections (like paraphrasing and contracting sentences). The transcript after this iteration roughly matched the audio clips.

Iteration 3: This version is where all missing minute details were captured by a second review by the EquiTech Futures team. We added any missing filler words (eg: "so", "umm", "err"). We
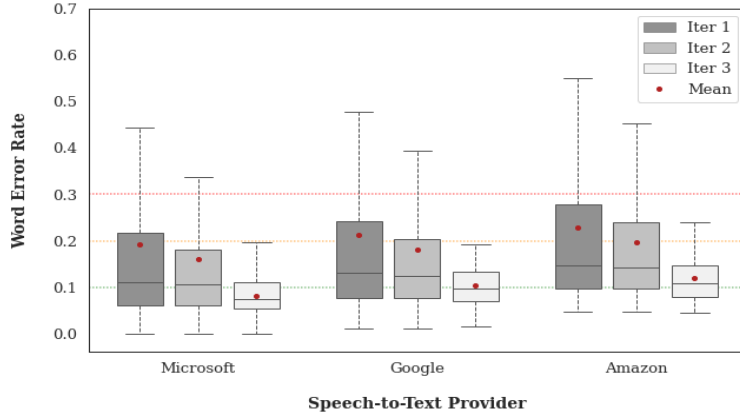
Figure 1: Word Error Rate (WER) for three ASR providers (Microsoft, Google, and Amazon). We show the WER for the three iterations of data cleaning described in Section 3.1

captured all the incomplete uttered words i.e. when the utterance was interrupted in the middle to form a related word (for eg: "the equation required simply... simplification").

3.2 Results

We begin by assessing the performance of off-the-shelf ASR systems by primary cloud service providers (Amazon, Google, Microsoft) using the word error rate (WER) defined as:

$$\text{WER} = \frac{S + D + I}{N} \tag{1}$$

where S/N, D/N, and I/N are the substitution, deletion, and insertion error rates respectively between the machine and human transcription. N is the total number of words in the human transcription.

Figure 1 shows that for all three providers the mean WER decreases by $\sim 10\%$ between the raw transcript and the final iteration. In fact, after rigorous cleaning of the transcripts the WER for all providers is close to the industry recommended benchmark of WER$= 0.1$.

In Figure 2 we plot the disaggregated insertion, substitution, and deletion ratios between different iterations. Evidently, the human transcriptions do not significantly remove words in the original audio. The WER is dominated by transcribers making substitutions and inserting new words during the transcription process.

Finally, in Figure 3 we investigate a hypothetical audit for ASR systems for rural vs urban populations for different iterations of the transcriptions. Transcriptions that have not been through detailed cross-checks would have revealed a stark but spurious difference in performance of ASR systems between rural and urban technology users. These differences disappear after correcting for the quality of ground-truth data.

An interesting feature of our hypothetical audit is the improvement of ASR performance only for rural populations after subsequent data cleaning. This is because the rural English speakers in our sample have more grammatically incorrect sentences (eg: misplaced articles, incorrect prepositions, incomplete sentences). The human transcribers (who are urban) automatically correct these grammatical errors (even after being instructed not to) but the ASR transcribes the text faithfully. This leads to a spurious reduction in ASR performance driven by human annotation errors.

4 Discussion

The initial goal of this project was to replicate results by Koenecke et al. (2020) for the South Asian context by creating a similar audit for widely deployed ASR systems in India. We were especially interested in looking at disparities between rural and urban English speakers and low and high-income
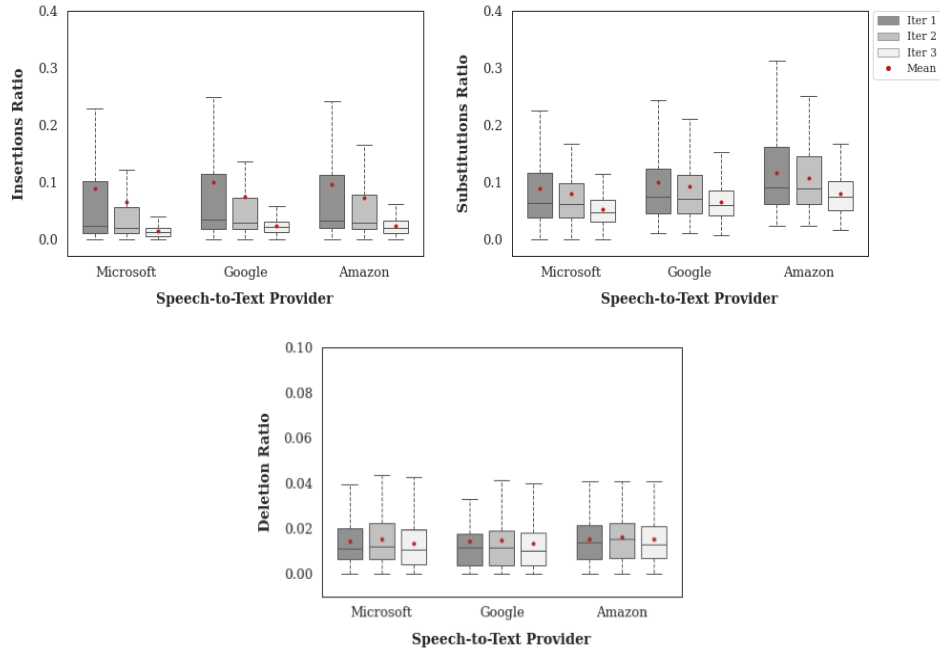
Figure 2: Insertion ratio, substitution ratio, and deletion ratio for the three ASR providers disaggregated by three iterations of data cleaning.
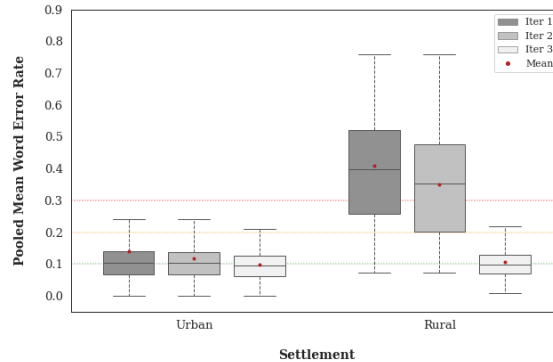

Figure 3: Comparison of mean WER between urban and rural sub-samples in our dataset after different iterations of data cleaning described in Section 3.1.

English speakers (since accents and fluency can vary significantly between these groups). During this process we discovered significant quality challenges with the transcribed "ground truth" data by human data labelers representative of data annotators typically employed in India. Our goal in writing this paper was to demonstrate empirically how algorithmic audits can suffer from challenges of data quality similar to the challenge confronted by algorithm developers. We hope our work highlights the need for widely accepted, consensus-driven standardized protocols for label quality. Such protocols are critical for AI governance globally.

Acknowledgments and Disclosure of Funding

References

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018.

Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019.

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, 2020.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019.

Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, 2019.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv preprint arXiv:2012.05345*, 2020.

James. Guszcza, Iyad Rahwan, Will Bible, Manuel Cebrian, and Vic Katyal. Why we need to audit algorithm, 2018. URL `https://hbr.org/2018/11/why-we-need-to-audit-algorithms`.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning*, pages 9625–9635. PMLR, 2020.