Racial disparities in automated speech recognition

Allison Koenecke[a], Andrew Nam[b], Emily Lake[c], Joe Nudell[d], Minnie Quartey[e], Zion Mengesha[c], Connor Toups[c], John R. Rickford[c], Dan Jurafsky[c,f], and Sharad Goel[d,1]

[a]Institute for Computational & Mathematical Engineering, Stanford University, Stanford, CA 94305; [b]Department of Psychology, Stanford University, Stanford, CA 94305; [c]Department of Linguistics, Stanford University, Stanford, CA 94305; [d]Department of Management Science & Engineering, Stanford University, Stanford, CA 94305; [e]Department of Linguistics, Georgetown University, Washington, DC 20057; and [f]Department of Computer Science, Stanford University, Stanford, CA 94305

Edited by Judith T. Irvine, University of Michigan, Ann Arbor, MI, and approved February 12, 2020 (received for review October 5, 2019)

Automated speech recognition (ASR) systems, which use sophisticated machine-learning algorithms to convert spoken language to text, have become increasingly widespread, powering popular virtual assistants, facilitating automated closed captioning, and enabling digital dictation platforms for health care. Over the last several years, the quality of these systems has dramatically improved, due both to advances in deep learning and to the collection of large-scale datasets used to train the systems. There is concern, however, that these tools do not work equally well for all subgroups of the population. Here, we examine the ability of five state-of-the-art ASR systems—developed by Amazon, Apple, Google, IBM, and Microsoft—to transcribe structured interviews conducted with 42 white speakers and 73 black speakers. In total, this corpus spans five US cities and consists of 19.8 h of audio matched on the age and gender of the speaker. We found that all five ASR systems exhibited substantial racial disparities, with an average word error rate (WER) of 0.35 for black speakers compared with 0.19 for white speakers. We trace these disparities to the underlying acoustic models used by the ASR systems as the race gap was equally large on a subset of identical phrases spoken by black and white individuals in our corpus. We conclude by proposing strategies—such as using more diverse training datasets that include African American Vernacular English—to reduce these performance differences and ensure speech recognition technology is inclusive.

fair machine learning | natural language processing | speech-to-text

The surge in speech-related research and, in particular, advances in deep learning for speech and natural language processing, have substantially improved the accuracy of automated speech recognition (ASR) systems. This technology is now employed in myriad applications used by millions of people worldwide. Some examples include virtual assistants built into mobile devices, home appliances, and in-car systems; digital dictation for completing medical records; automatic translation; automated subtitling for video content; and hands-free computing. These last two applications are particularly useful for individuals with hearing loss and motor impairments and point to the value of ASR systems to increase accessibility.

There is worry, however, that speech recognition systems suffer from racial bias (1–4), a problem that has recently come to light in several other advancing applications of machine learning, such as face recognition (5, 6), natural language processing (7–11), online advertising (12, 13), and risk prediction in criminal justice (14–17), healthcare (18, 19), and child services (20, 21). Here, we assess racial disparities in five commercial speech-to-text tools—developed by Amazon, Apple, Google, IBM, and Microsoft—that power some of the most popular applications of voice recognition technology.

Our analysis is based on two recently collected corpora of conversational speech. The first is the Corpus of Regional African American Language (CORAAL) (22), a collection of sociolinguistic interviews with dozens of black individuals who speak African American Vernacular English (AAVE) (23–25) to varying degrees. These interviews were conducted at three US sites: Princeville, a rural, nearly exclusively African American community in eastern North Carolina; Rochester, a moderate-sized city in Western New York; and the District of Columbia. The second dataset we use is Voices of California (VOC) (26), an ongoing compilation of interviews recorded across the state in both rural and urban areas. We focus our analysis on two California sites: Sacramento, the state capitol; and Humboldt County, a predominately white rural community in Northern California.

In both datasets, the interviews were transcribed by human experts, which we use as the ground truth when evaluating the performance of machine transcriptions. The original recorded interviews contain audio from both the interviewer and the interviewee. Our study is based on a subset of audio snippets that exclusively contain the interviewee and are 5 to 50 s long. We match these snippets across the two datasets based on the age and gender of the speaker and the duration of the snippet. After matching, we are left with 2,141 snippets from each dataset, with an average length of 17 s per snippet, amounting to 19.8 total hours of audio. In the matched dataset, 44% of snippets were of male speakers, and the average age of speakers was 45 y.

We assess the performance of the ASR systems in terms of the word error rate (WER) (27), a standard measure of discrepancy between machine and human transcriptions. Formally, WER is defined as:

$$\mathrm{WER} = \frac{S + D + I}{N}, \qquad [1]$$

where S, D, and I denote the number of word substitutions, deletions, and insertions between the machine and ground-truth

Significance

Automated speech recognition (ASR) systems are now used in a variety of applications to convert spoken language to text, from virtual assistants, to closed captioning, to hands-free computing. By analyzing a large corpus of sociolinguistic interviews with white and African American speakers, we demonstrate large racial disparities in the performance of five popular commercial ASR systems. Our results point to hurdles faced by African Americans in using increasingly widespread tools driven by speech recognition technology. More generally, our work illustrates the need to audit emerging machine-learning systems to ensure they are broadly inclusive.

Author contributions: A.K., A.N., E.L., J.N., M.Q., Z.M., C.T., J.R.R., D.J., and S.G. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Data deposition: Data and code to reproduce the findings discussed in this paper are available on GitHub (https://github.com/stanford-policylab/asr-disparities).

[1]To whom correspondence may be addressed. Email: scgoel@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1915768117/-/DCSupplemental.

First published March 23, 2020.

transcriptions, respectively, and N is the total number of words in the ground truth. A higher WER thus indicates a greater difference between the two transcriptions and hence worse ASR performance in our setting.

Results

We start by computing the average word error rates for machine transcriptions across our matched audio snippets of white and black speakers. For each of the five commercial ASR systems we examine, Fig. 1 shows that the average WER for black speakers is substantially larger than the average WER for white speakers. For example, for Microsoft's ASR, which has the best overall performance, the WER for black speakers is 0.27 (SE: 0.004) compared with 0.15 (SE: 0.003) for white speakers. Furthermore, for Apple, whose ASR has the worst overall performance, the WERs for black and white speakers are 0.45 (SE: 0.005) and 0.23 (SE: 0.003), respectively.* Despite variation in transcription quality across systems, the error rates for black speakers are nearly twice as large in every case. Averaging error rates across ASR services yields an aggregate WER of 0.35 (SE: 0.004) for black speakers versus 0.19 (SE: 0.003) for white speakers.

The error rates are particularly large for black men in our sample. Averaging across the five ASR systems, the error rate for black men is 0.41 (SE: 0.006) compared with 0.30 (SE: 0.005) for black women. In comparison, the average error rates for white men and women are more similar at 0.21 (SE: 0.004) and 0.17 (SE: 0.003), respectively.† Past work has also found that ASRs perform somewhat worse on conversational speech from male speakers than female speakers, likely due to male speakers using more informal style with shorter, more reduced pronunciations and more disfluencies (28, 29). This decreased performance on male speakers is more pronounced for the black speakers in our sample—a point we return to below.

To add more detail to the average error rates discussed above, we next consider the full distribution of error rates across our populations of white and black speakers. To do so, for each snippet, we first compute the average WER across the five ASRs we consider. Fig. 2 plots the distribution of this average WER across snippets, disaggregated by race. In particular, Fig. 2 shows the complementary cumulative distribution function (CCDF): for each value of WER on the horizontal axis, it shows the proportion of snippets having an error rate at least that large. For example, more than 20% of snippets of black speakers have an error rate of at least 0.5; in contrast, fewer than 2% of snippets of white speakers are above that threshold. Thus, if one considers a WER of 0.5 to be the bar for a useful transcription, more than 10 times as many snippets of black speakers fail to meet that standard. In this sense, the racial disparities we find are even larger than indicated by the average differences in WER alone.

We next examine variation in error rate by location. The black speakers in our matched sample were interviewed in Princeville ($n = 21$); Washington, DC ($n = 39$); and Rochester ($n = 13$); the white speakers were interviewed in Sacramento ($n = 17$) and Humboldt County ($n = 25$). As above, we first compute the average WER for each snippet across the five ASRs. Fig. 3 summarizes the distribution of these average error rates for each location as a boxplot, with the center lines of each box indicating the median error rate and the endpoints indicating the interquartile range. The median error rates in Princeville (0.38) and Washington, DC (0.31), are considerably larger than those in Sacramento and Humboldt (0.18 and 0.15, respectively). However, the error rate in the third AAVE site, Rochester (0.20), is comparable to the error rates in the two California locations with white speakers.

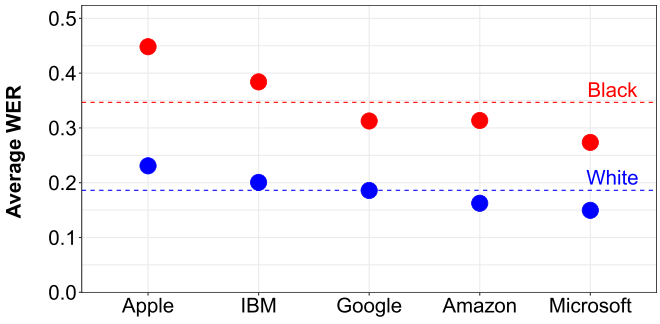

Fig. 1. The average WER across ASR services is 0.35 for audio snippets of black speakers, as opposed to 0.19 for snippets of white speakers. The maximum SE among the 10 WER values displayed (across black and white speakers and across ASR services) is 0.005. For each ASR service, the average WER is calculated across a matched sample of 2,141 black and 2,141 white audio snippets, totaling 19.8 h of interviewee audio. Nearest-neighbor matching between speaker race was performed based on the speaker's age, gender, and audio snippet duration.

To better understand the geographical patterns described above—particularly the anomalous results in Rochester—we hand-coded a random sample of 150 snippets of black speakers for usage of AAVE linguistic features, with 50 snippets coded from each of the three AAVE interview sites. Specifically, for each snippet, we counted the number of phonological and grammatical features characteristic of AAVE speech and then normalized this count by the number of words in the snippet, yielding a dialect density measure (DDM).

We find that average DDM is lowest in Rochester (0.047)—and also relatively small on an absolute scale—followed by Washington, DC (0.088), and Princeville (0.19), mirroring the ordering of word error rates by location seen in Fig. 3. The pairwise differences in DDM by location are statistically significant, with $P < 0.05$ in all cases. In Fig. 4, we directly examine the relationship between DDM (on the horizontal axis) and WER (on the vertical axis), which illustrates the positive correlation between DDM and error rates. Although there are many factors that affect error rates, these results suggest that the location-specific patterns we see are, at least in part, driven by differences in the degree of AAVE usage among speakers in our sample. Given the relatively small number of speakers in each location, we cannot determine whether these patterns are representative of more general geographic differences in dialect or are simply idiosyncratic trends in our particular sample of speakers.

This coding of dialect density also reveals gender differences. Aggregated across the three AAVE sites, the DDM for male speakers is 0.13 ($n = 52$; SE: 0.02), compared with 0.096 for female speakers ($n = 98$; SE: 0.01). As with location, this pattern is in line with the higher ASR error rate for male speakers discussed above.

We conclude by investigating two possible mechanisms that could account for the racial disparities we see: 1) a performance gap in the "language models" (models of lexicon and grammar) underlying modern ASR systems; and 2) a performance gap in

*The relatively poor quality of Apple's ASR may be due to the fact that it produces streaming transcriptions, in which results are generated in real time as the audio is processed. In contrast, it appears that the other ASRs consider the entirety of an audio snippet before producing the final transcript.

†Our matching procedure only ensures that our samples of white and black speakers are directly comparable. In particular, within each race group, the subset of women is not explicitly matched to the subset of men. We address this issue in *SI Appendix* via a set of linear regression models that estimate error rates as a function of race, age, gender, and snippet duration. That approach again indicates the gender gap in performance is substantially larger for black speakers than for white speakers, corroborating the results discussed above.

the acoustic models underlying these systems. As we discuss next, we find evidence of a gap in the acoustic models but not in the language models.

Speech recognition systems typically have a fixed—although potentially quite large—vocabulary that forms the basis of transcriptions. In theory, it is possible that the black speakers in our sample more often use words that are simply not included in the vocabulary of the ASR systems we investigate, which, if true, could explain the racial disparities we observe. To examine this hypothesis, we first approximately reconstruct the lexicon of each of the five ASR systems by aggregating all unique words that appear in each ASR's transcriptions, combining the transcriptions for black and white speakers. These approximate lexicons are a subset of the true list, as the ASR systems may have in their vocabularies words that were never spoken by our speakers (or were never correctly recognized). For example, we find 8,852 distinct words that appear at least once in the transcripts produced by Google's ASR.

Now, we compute the proportion of words in the ground-truth human transcripts—including repeated instances—that are present in the reconstructed machine vocabularies. For both white and black speakers, and across the five ASR systems, 98 to 99% of the words spoken are in the reconstructed vocabularies. For example, of the 104,486 words uttered by black speakers in our sample, Google's ASR had at least 103,142 (98.7%) of them in its vocabulary; in comparison, of the 98,653 words spoken by white individuals in our sample, at least 97,260 (98.6%) were in the vocabulary. These modest lexical differences do not appear large enough to explain the substantial gap in overall error rates we find—and, indeed, a slightly greater fraction of words spoken by black sample members are in the machine vocabulary than that of white sample members.

We next investigate potential racial disparities in the full computational model of language used by ASR systems. At a high level, language models predict the next word in a sequence given the previous words in that sequence. For example, given the incomplete phrase "the dog jumped over the ___," a language model might estimate that there is a 5% chance the next word is "fence."

The standard performance metric for language models is perplexity, which roughly can be viewed as the number of reasonable continuations of a phrase under the model. Accordingly, better language models have lower perplexity. Formally, given a language model M and a sequence of words $x_1, \ldots, x_N$ (corresponding, in our case, to a ground-truth human transcription of an audio snippet), perplexity is:

$$\exp\left(-\frac{1}{N-1}\sum_{i=2}^{N}\log \mathbb{P}_M(x_i \mid x_{i-1}, \ldots, x_1)\right), \qquad [2]$$

where $\mathbb{P}_M(x_i \mid x_{i-1}, \ldots, x_1)$ is the conditional probability assigned by the model to the word at index i.

The exact language models underlying commercial ASR systems are not readily available. However, it is likely that these systems use language models that have similar statistical properties to state-of-the-art models that are publicly available, like Transformer-XL (30), GPT (31), and GPT-2 (32). We thus examine potential racial disparities in these three models, using the publicly available versions that have been pretrained on large corpora of text data.[‡]

Under all three language models, we find the average perplexity of snippets by black speakers is lower—meaning better performance—than the average perplexity of snippets by white speakers in our sample. In particular, Transformer-XL has perplexity of 115 for black speakers compared with 153 for white speakers; GPT has perplexity of 52 and 68 for black and white speakers, respectively; and GPT-2 has perplexity of 45 and 55, respectively. These three language models—and, by extension, likely the language models used in commercial ASR systems—are, on average, better able to predict the sequences of words spoken by black individuals in our sample than those spoken by the white individuals.

To investigate this result, we consider a sample of phrases spoken by black speakers in our dataset that exhibit a common grammatical feature of AAVE: copula absence, or omission of the verb "be." For example, one black speaker in our corpus said, "he a pastor," rather than using the Standard English phrasing, "he's a pastor." In Table 1, we list a representative selection of five such AAVE phrases drawn from the set of snippets coded for dialect density (discussed above). We compute the perplexity of both the original phrase and a modified version in which

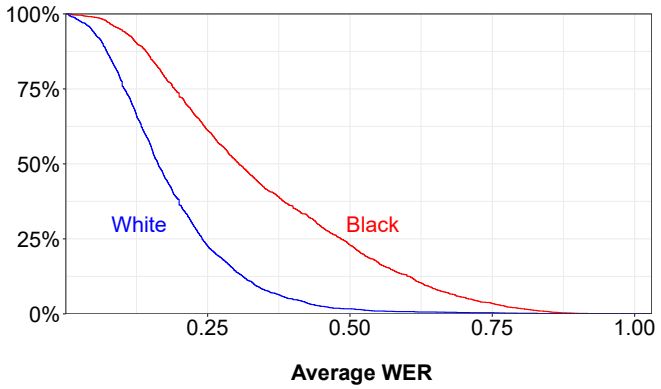

Fig. 2. The CCDF denotes the share of audio snippets having a WER greater than the value specified along the horizontal axis. The two CCDFs shown for audio snippets by white speakers (blue) versus those by black speakers (red) use the average WER across the five ASR services tested. If we assume that a WER >0.5 implies a transcript is unusable, then 23% of audio snippets of black speakers result in unusable transcripts, whereas only 1.6% of audio snippets of white speakers result in unusable transcripts.

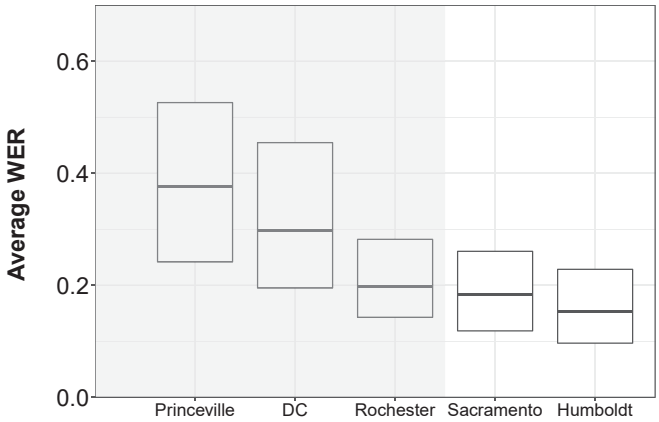

Fig. 3. For each audio snippet, we first computed the average error rate across the five ASR services we consider: Amazon, Apple, Google, IBM, and Microsoft. These average WERs were then grouped by interview location, with the distributions summarized in the boxplots above. In the three AAVE sites, denoted by a gray background (Princeville, NC; Washington, DC; and Rochester, NY), the error rates are typically higher than in the two white sites (Sacramento, CA, and Humboldt, CA), although error rates in Rochester are comparable to those in Sacramento.

[‡]We use the language models available at https://huggingface.co/transformers/.

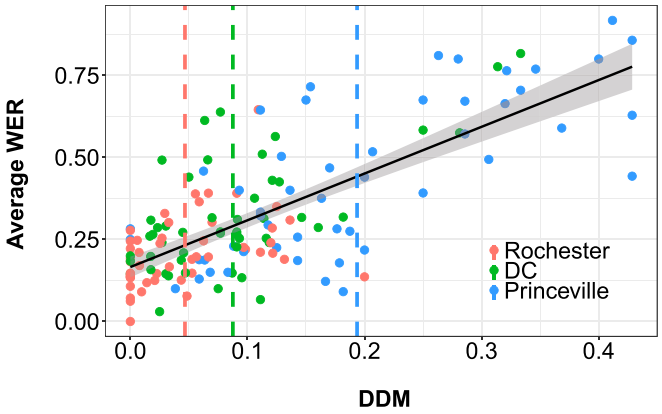

Fig. 4. The relationship between a measure of dialect density (DDM, on the horizontal axis) and average ASR error rate (WER, on the vertical axis) for a random sample of 50 snippets in each of the three AAVE sites we consider. The dashed vertical lines indicate the average DDM in each location. The solid black line shows a linear regression fit to the data and indicates that speakers who exhibit more linguistic features characteristic of AAVE tend to have higher WER.

the copula is inserted to comport with Standard English convention. For simplicity, perplexity is computed under the GPT-2 language model, although results are qualitatively similar under GPT-1 and Transformer-XL.

For all five of the listed phrases, the perplexity of the original AAVE phrasing is considerably greater than the perplexity of the Standard English phrasing. For example, "he a pastor" has perplexity of 305, compared with 67 for "he's a pastor." The language models we consider thus appear to exhibit a statistical preference for the Standard English inclusion of the copula over the AAVE copula absence.

Given this behavior, the overall lower average perplexity for snippets of black speakers seems even more surprising. We believe this difference is at least partially due to the relative number of unique words spoken by black and white sample members. Although the total duration and number of words spoken by black and white speakers in our sample were similar, black speakers uttered fewer unique words (5,651) than white speakers (6,280). All else being equal, a smaller vocabulary generally yields lower model perplexity, as it is easier to predict the next word in a sequence.[§]

Our investigation thus indicates that the lexical and grammatical properties of ASR systems do not account for the large overall racial disparities in WERs. If anything, since these snippets from black speakers have fewer unique words and lower perplexity, they should be easier for the ASRs to transcribe.

These results suggest that the problem may instead lie with the acoustic models underlying ASRs. To examine this possibility, we compare error rates on a set of short phrases uttered by black and white speakers in our sample that have identical ground-truth human transcripts. We specifically limit to phrases comprised of at least five words that were spoken by white and black individuals of the same gender and of approximately the same age. This process yielded 206 matched utterances of 5 to 8 words (e.g., "and then a lot of the" and "and my mother was a").

[§]We similarly find that effective vocabulary size is smaller for black speakers in our sample (386) than for white speakers (452). Effective vocabulary size is the perplexity of a null language model M_0 that sets the probability of a word in a sequence to be the overall empirical frequency of that word, irrespective of context. Specifically, given a corpus of text $\mathcal{C}$, $\mathbb{P}_{M_0}(x_i \mid x_{i-1}, \ldots, x_1) = n_{x_i}/n$, where n_{x_i} is the number of occurrences of x_i in $\mathcal{C}$, and n is the total size of $\mathcal{C}$.

Error rates for this set of matched phrases are presented in Table 2. For each of the five ASR systems we consider, WERs are about twice as large when the phrases were spoken by black individuals rather than whites. For example, with Microsoft's ASR—which has the best overall performance—the WER for black speakers is 0.13 (SE: 0.01) compared with 0.07 (SE: 0.01) for white speakers. Given that the phrases themselves have identical text, these results suggest that racial disparities in ASR performance are related to differences in pronunciation and prosody—including rhythm, pitch, syllable accenting, vowel duration, and lenition—between white and black speakers.

Discussion

As noted above, modern automated speech recognition systems generally include a language model trained on text data and an acoustic model trained on audio data. Our findings indicate that the racial disparities we see arise primarily from a performance gap in the acoustic models, suggesting that the systems are confused by the phonological, phonetic, or prosodic characteristics of African American Vernacular English rather than the grammatical or lexical characteristics. The likely cause of this shortcoming is insufficient audio data from black speakers when training the models.

The performance gaps we have documented suggest it is considerably harder for African Americans to benefit from the increasingly widespread use of speech recognition technology, from virtual assistants on mobile phones to hands-free computing for the physically impaired. These disparities may also actively harm African American communities when, for example, speech recognition software is used by employers to automatically evaluate candidate interviews or by criminal justice agencies to automatically transcribe courtroom proceedings.

One limitation of our study is that the audio samples of white and black speakers came from different geographical areas of the country, with the former collected in California and the latter in the Eastern United States. As such, it is possible that at least some of the differences we see are a product of regional—rather than ethnic—linguistic variation. We note, however, two reasons to believe that AAVE speech itself is driving our results. First, word error rate is strongly associated with AAVE dialect density, as seen in Fig. 4. Second, the two California sites of white speakers that we consider, Sacramento and Humboldt, exhibit similar error rates despite diversity in regional speech patterns across the state and differences in the sociogeographical contexts of these two locations—for example, Humboldt is a rural community, whereas Sacramento is the state capitol. Nevertheless, we hope that future work examines error rates among white and black speakers from the same region.

Our findings highlight the need for the speech recognition community—including makers of speech recognition systems,

Table 1. Perplexity for Standard English phrasing, with the copula in bold, and AAVE phrasing, without the bolded copula

	AAVE	Standard English
	perplexity	perplexity
He**'s** a pastor.	305	67
We**'re** going to the arc.	190	88
We**'re** able to fight for the cause.	54	51
Where **are** they from?	570	20
Have you decided what you**'re** going to sing?	106	25

Perplexity is computed under the GPT-2 language model. In all of these examples, the AAVE phrasing has higher perplexity (and is hence more surprising to the model) than the Standard English phrasing.

Table 2. Error rates on a matched subset of identical short phrases spoken by white and black individuals in our sample

	Average WER for black speakers	Average WER for white speakers
Apple	0.28	0.12
IBM	0.21	0.10
Google	0.17	0.11
Amazon	0.18	0.08
Microsoft	0.13	0.07

academic speech recognition researchers, and government sponsors of speech research—to invest resources into ensuring that systems are broadly inclusive. Such an effort, we believe, should entail not only better collection of data on AAVE speech but also better collection of data on other nonstandard varieties of English, whose speakers may similarly be burdened by poor ASR performance—including those with regional and nonnative-English accents. We also believe developers of speech recognition tools in industry and academia should regularly assess and publicly report their progress along this dimension. With adoption of speech recognition systems likely to grow over time, we hope technology firms and other participants in this field foreground the equitable development of these important tools.

Materials and Methods

We briefly describe our data filtering, standardization, and matching procedures below, as well as our process for measuring dialect density. Further details are provided in *SI Appendix*.

Data. Our audio snippets come from the full set of 108 CORAAL interviews and 109 VOC interviews in the five geographic sites we consider. The CORAAL interviews conducted in Washington, DC, Rochester, and Princeville were recorded in 2016, 2016, and 2004, respectively; and the VOC interviews conducted in Sacramento and Humboldt were recorded in 2014 and 2017, respectively. The majority of our data come from 2014 to 2017—a span that does not represent a significant time gap for sociolinguistic analysis—but the Princeville data were collected a decade earlier, in 2004. Relatedly, the Princeville data were recorded on cassette tape and then later digitized, whereas interviews in the other sites were all recorded using digital devices. Given the obstacles in assembling data from a large number of speakers across multiple field sites, it is not uncommon in dialectology studies to combine audio collected across different years and recorded with different equipment. While it is important to recognize these limitations of our study design, we believe they are unlikely to impact our main results.

We restricted our analysis to interviews of adults (older than 18 y) that had generally good audio quality (e.g., without significant background noise). In the VOC data, we additionally restricted to non-Hispanic white speakers. In this restricted set of interviews, we extracted the longest continuous, full-phrase interviewee segments that were between 5 and 50 s long. In particular, we removed audio segments containing interruptions or overlapping utterances from the interviewer (or other noninterviewees, if any). We also ensured that audio snippets began and ended at natural pauses, such as the completion of a sentence. We limited our analysis to segments of at most 50 s, as some of the ASR systems we examined could not transcribe longer audio files. This process resulted in 4,449 audio snippets of black speakers and 4,397 audio snippets of white speakers.

Next, we cleaned the ground-truth human transcripts to ensure consistency across the two datasets. More specifically, we modified nonstandard spellings: for example, we changed occurrences of the word "aks" to "ask," since no ASRs spell this utterance using the AAVE pronunciation. Flags for unintelligible audio content (e.g., an "/unintelligible/" string occurring in the ground-truth human transcript) occur in 16% of CORAAL snippets and 11% of VOC snippets. Typically, the ASR systems simply ignored these unintelligible segments of the audio snippet, and so we accordingly removed the flags from the human transcripts. We likewise removed flags for redacted words and nonlinguistic markers (e.g., for breath and laughter), as these were not transcribed by the ASR systems. We confirmed that our results were nearly identical if, instead of performing the above operations, snippets with questionable content were removed entirely. Some location-specific words uttered in CORAAL and VOC were particularly hard for the ASR systems to spell (e.g., "Tarboro" and "Yurok"); the ASRs regularly misspelled Humboldt as "humble" or "humbled." We compared our results with those where all snippets containing a list of hard-to-spell city names uttered in the audio snippets were removed. Again, our results did not meaningfully change, as such problematic words were relatively rare.

We additionally standardized all of the human and machine transcripts using the following rules to facilitate error rate calculations. Single spacing was enforced between words; Arabic numerals were converted to numeric strings; flags indicating hesitation were removed from the transcripts; the "$" sign was replaced with the "dollar" string; all other special characters and punctuation were removed; cardinal direction abbreviations (e.g., "NW") were replaced with full words (e.g., "Northwest"); full state names were replaced with their two-letter abbreviations; and all words were converted to lowercase. Also, certain spellings were standardized: for example, "cuz," "ok," "o," "till," "imma," "mister," "yup," "gonna," and "tryna" were, respectively, replaced with "cause," "okay," "oh," "til," "ima," "mr," "yep," "going to," and "trying to"). Finally, we removed both filler words ("um," "uh," "mm," "hm," "ooh," "woo," "mhm," "huh," "ha") and expletives because the ASR systems handle these words differently from each other (e.g., removing them from the transcription outputs), similar to how different human transcribers might also treat them subjectively.

Lastly, we restricted our analysis to snippets with a cleaned ground-truth word count of at least five words. This entire filtering and cleaning process yielded a set of 4,445 audio snippets by 73 black speakers and 4,372 audio snippets by 51 white speakers, totaling 39.8 h of audio. On this restricted set of snippets, we calculated the WERs generated by each ASR. Specifically, the WER was calculated between the cleaned version of the original snippet transcription (from CORAAL or VOC) and the cleaned version of each ASR-generated transcript. Our main statistical analysis was based on a subset of matched snippets, as described next.

Matching. We used propensity-score matching to select a subset of audio snippets of white and black speakers with similar distributions of age, gender, and snippet duration. This restriction allowed us to focus on racial disparities, as age and gender are also known to impact the performance of ASR systems (28, 29). Matching was done with the R package MatchIt (33), with propensity scores estimated via a logistic regression model on the combined data from black and white speakers. Specifically, in our propensity score model, we regressed an indicator for race on the following covariates: indicator variables for 10-y-wide age bins for ages 25 through 94 y (e.g., 25 to 34 y and 35 to 44 y); integer age; an indicator variable for gender; and natural log of the snippet length, measured in seconds.

Nearest-neighbor matching without replacement was performed on the propensity scores, with a caliper size of 0.001. The final set of matched audio snippets is comprised of 2,141 snippets by 73 black speakers and an equal number of snippets by 42 white speakers, corresponding to 19.8 total hours of audio. As shown in *SI Appendix*, Fig. S1, the matched samples of black and white snippets—in the bottom row, as opposed to the prematched samples in the top row—have closely aligned distributions on our three target covariates: speaker age, speaker gender, and duration.

Measuring Dialect Density. We utilized a DDM to determine the relative number of AAVE features employed in a given audio snippet, dividing the total number of dialect features by the number of words in the snippet. Most previous studies using DDMs have focused on the syntactic complexity of AAVE (34, 35). For this study, however, we modified that approach to account for both AAVE grammar and phonology, with both grammatical and phonological features given equal weight. DDMs do not capture a speaker's entire linguistic system (36, 37), but, in our setting, the measure we use provides insight into drivers of the ASR performance gaps we see.

In our primary analysis, a subset of 150 snippets was annotated by a linguist familiar with AAVE. The annotator listened to a snippet and recorded each AAVE phonological feature and grammatical feature. For example, in the Princeville snippet "Well at that time it was Carolina Enterprise, but it done changed name," there are five AAVE features (three phonological and two grammatical): 1) final consonant deletion in "at"; 2) syllable initial fricative stopping in "that"; 3) vocalization of postvocalic/r/ in "enterprise"; 4) absence of plural -s in "name"; and 5) and completive "done" in "it done changed." Because the snippet is 13 words long, the DDM is $5/13 = 0.38$. The complete list of AAVE features that we tagged is based on past work

(23–25) and is shown in *SI Appendix*, Tables S2 and S3. Across the full set of 150 coded snippets, the average length was 47 words, with 3.5 phonological features and 0.5 grammatical features, on average; the average DDM was 0.11.

To gauge interrater reliability, we compared DDM scores of the primary coder with those of two other trained sociolinguists on a test set of 20 snippets—10 snippets for each of the two secondary coders. The Pearson correlation between the primary coder and the two secondary coders was 0.92 and 0.74, respectively, indicating high agreement.

ACKNOWLEDGMENTS. We thank Tyler Kendall and Sharese King for their advice on working with the CORAAL dataset, Penelope Eckert and Rob Podesva for their assistance with the VOC data, and speech teams at the ASR firms for their help.

1. R. Tatman, "Gender and dialect bias in YouTube's automatic captions" in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, D. Hovy *et al.*, Eds. (Association for Computational Linguistics, 2017), pp. 53–59.
2. R. Tatman, C. Kasten, "Effects of talker dialect, gender & race on accuracy of Bing speech and YouTube automatic captions" in *INTERSPEECH*, F. Lacerda *et al.*, Eds. (International Speech Communication Association, 2017), pp. 934–938.
3. D. Harwell, B. Mayes, M. Walls, S. Hashemi, The accent gap. *The Washington Post*, 19 July 2018. https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/. Accessed 28 February 2020.
4. F. Kitashov, E. Svitanko, D. Dutta, Foreign English accent adjustment by learning phonetic patterns. arXiv:1807.03625 (9 July 2018).
5. J. Buolamwini, T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification" in *Proceedings of the Conference on Fairness, Accountability and Transparency*, S. A. Friedler, C. Wilson, Eds. (Association for Computing Machinery, New York, NY, 2018), pp. 77–91.
6. I. D. Raji, J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products" in *AAAI/ACM Conference on AI Ethics and Society* (Association for Computing Machinery, 2019), Vol. 1, pp. 429–435.
7. S. L. Blodgett, L. Green, B. O'Connor, "Demographic dialectal variation in social media: A case study of African-American English" in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, X. Carreras, Eds. (Association for Computational Linguistics, 2016), pp. 1119–1130.
8. S. L. Blodgett, B. O'Connor, Racial disparity in natural language processing: A case study of social media African-American English. arXiv:1707.00061 (30 June 2017).
9. A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
10. M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, "The risk of racial bias in hate speech detection" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, L. Màrquez, Eds. (Association for Computational Linguistics, 2019), pp. 1668–1678.
11. M. De-Arteaga *et al.*, "Bias in bios: A case study of semantic representation bias in a high-stakes setting" in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (ACM, 2019), pp. 120–128.
12. M. Ali *et al.*, Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. arXiv:1904.02095 (12 September 2019).
13. A. Datta, A. Datta, J. Makagon, D. K. Mulligan, M. C. Tschantz, "Discrimination in online advertising: A multidisciplinary inquiry" in *Proceedings of the Conference on Fairness, Accountability and Transparency*, S. A. Friedler, C. Wilson, Eds. (Association for Computing Machinery, New York, NY, 2018), pp. 20–34.
14. S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, "Algorithmic decision making and the cost of fairness" in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2017), pp. 797–806.
15. S. Corbett-Davies, S. Goel, The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv:1808.00023 (14 August 2018).
16. J. Kleinberg, S. Mullainathan, M. Raghavan, "Inherent trade-offs in the fair determination of risk scores" in *Proceedings of Innovations in Theoretical Computer Science*, C. H. Papadimitriou, Ed. (ITCS, 2017).
17. A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* **5**, 153–163 (2017).
18. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
19. S. N. Goodman, S. Goel, M. R. Cullen, Machine learning, health disparities, and causal reasoning. *Ann. Intern. Med.* **169**, 883 (2018).
20. A. Chouldechova, D. Benavides-Prado, O. Fialko, R. Vaithianathan, "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions" in *Proceedings of the Conference on Fairness, Accountability and Transparency*, S. A. Friedler, C. Wilson, Eds. (Association for Computing Machinery, New York, NY, 2018), pp. 134–148.
21. R. Shroff, Predictive analytics for city agencies: Lessons from children's services. *Big Data* **5**, 189–196 (2017).
22. T. Kendall, C. Farrington, The corpus of regional African American language (2018). https://oraal.uoregon.edu/coraal. Accessed 28 February 2020.
23. J. R. Rickford, *Phonological and Grammatical Features of African American Vernacular (AAVE)* (Blackwell Publishers, 1999).
24. E. R. Thomas, Phonological and phonetic characteristics of African American Vernacular English. *Lang. Linguist. Compass* **1**, 450–475 (2007).
25. G. Bailey, E. Thomas, "Some aspects of African-American Vernacular English phonology" in *African-American English: Structure, History and Use*, S. Mufwene, J. R. Rickford, G. Bailey, J. Baugh, Eds. (Routledge, New York, NY, 1998), pp. 85–109.
26. Stanford Linguistics, Voices of California. http://web.stanford.edu/dept/linguistics/VoCal/. Accessed 28 February 2020.
27. D. Klakow, J. Peters, Testing the correlation of word error rate and perplexity. *Speech Commun.* **38**, 19–28 (2002).
28. S. Goldwater, D. Jurafsky, C. D. Manning, Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Commun.* **52**, 181–200 (2010).
29. M. Adda-Decker, L. Lamel, "Do speech recognizers prefer female speakers?" in *Proceedings of INTERSPEECH-2005* (International Speech Communication Association, 2005), pp. 2205–2208.
30. Z. Dai *et al.*, "Transformer-XL: Attentive language models beyond a fixed-length context" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, L. Màrquez, Eds. (Association for Computational Linguistics, 2019), pp. 2978–2988.
31. A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf. Accessed 28 February 2020.
32. A. Radford *et al.*, Language models are unsupervised multitask learners. https://openai.com/blog/better-language-models/. Accessed 28 February 2020.
33. D. E. Ho, K. Imai, G. King, E. A. Stuart, MatchIt: Nonparametric preprocessing for parametric causal inference. *J. Stat. Software* **42**, 1–28 (2011).
34. H. K. Craig, J. A. Washington, An assessment battery for identifying language impairments in African American children. *J. Speech Lang. Hear. Res.* **43**, 366–379 (2000).
35. J. B. Oetting, J. L. McDonald, Methods for characterizing participants' nonmainstream dialect use in child language research. *J. Speech Lang. Hear. Res.* (2002).
36. A. H. C. Hudley, "Language and racialization" in *The Oxford Handbook of Language and Society*, O. García, N. Flores, M. Spotti, Eds. (Oxford University Press, 2016), pp. 381–402.
37. L. Green, "Beyond lists of differences to accurate descriptions" in *Data Collection in Sociolinguistics: Methods and Applications*, C. Mallinson, B. Childs, G. Van Herk, Eds. (Routledge, 2017), pp. 281–285.