Supplementary Information

Below we provide additional information on several aspects of our study. In particular, we describe how we interfaced with the automated speech recognition systems. We further describe how the human transcriptions were generated, and our coding of dialect strength. Finally, we evaluate our matching methodology, and carry out an additional regression-based robustness check of our main empirical results.

Interfacing with the speech recognition systems. All five of the speech recognition services we investigated—by Amazon, Apple, Google, IBM, and Microsoft—offer public speech-to-text interfaces, typically costing a few cents per minute of transcribed audio. Four of these five services (Amazon, Google, IBM, and Microsoft) provide similar, mostly straightforward "RESTful" APIs for transcription, and we used Python scripts to access the services in these instances.

For the fifth service provider, Apple, we built a custom iOS application to carry out the transcription. When we developed our application, Apple's Speech SDK was available only for iOS.* We wrote the iOS application in Swift to submit audio files for transcription on Apple's servers. Although we could have run the application on a physical mobile device, that would have added cost and complexity to the design. Instead, we ran the application through the Xcode device simulator, which allowed us both to read audio files and to write transcription results locally using only a laptop. Transcription via the Speech SDK was free of cost; the only constraint we found was the inability to run transcription tasks concurrently. We also note that Apple provided *streaming* transcriptions, in which results are based only on audio up until that point in the transcript. In contrast, the other four services appear to process the entirety of the audio snippet before returning any results, potentially accounting for the better overall performance of those services.

We faced periodic socket timeouts with all of the five ASR services, and thus occasionally had to restart our scripts. We further note that all five services failed to transcribe one or more snippets, even after multiple retries. Typically this was a problem for only a handful of the several thousand audio snippets we attempted to transcribe, but Google and Apple failed to return results for more than 50 snippets (from the unmatched, full dataset). In these cases, we assigned the snippet a word error rate of 1; we note, though, that given the large number of snippets we consider, these failed transcriptions did not qualitatively impact the average error rates of any of the ASRs.

In most instances, we stored the audio snippets locally and passed them to the ASR service only during the transcription call. The one exception was for Amazon's ASR system, which required that all snippets be kept in its cloud storage platform prior to transcription. Any stored snippets were deleted by all services after transcription. Some services, including Amazon and IBM, returned confidence measures for generated transcripts as well as alternative transcripts for each separate utterance. We did not use these confidence scores or alternative transcripts in our analysis.

Human transcriptions. The Voices of California corpus uses transcription guidelines from the "Automatic Alignment and Analysis of Linguistic Change — Transcription Guidelines" (February 2011).† Transcribers use standard orthography, punctuation and capitalization as well as word segmentation and word spelling. Exclamation marks are for emphatic speech and quotation marks are for direct speech or thought. When a contraction is produced by a speaker, this is transcribed in a standard contraction form (e.g., *doesn't*, *isn't*). When non-standard forms are produced, these are transcribed as *gonna*, *woulda*, *gotta*, etc., rather than the standard orthographic form. Numerals are transcribed as full words (e.g., *twenty-two*) and hyphens are used when required in compounds (e.g., *anti-capitalist*). For partial or truncated words, a single dash (-), with no preceding space, is used to mark the place where the word was cut off followed by a plus sign (+) when the transcriber has good reason to guess as to the intended word (e.g., *I thi- +think*). Restarts are indicated by a double dash with spaces (e.g., *I uh -- think*). An asterisk (*) is used for mispronounced

*At the time of writing, Speech SDK has become available in beta for macOS 10.15 (Catalina), released at the end of 2019.

†Available online at: https://www.ling.upenn.edu/~wlabov/L560/Transcription_guidelines_FAAV.pdf. Adapted from The SLX Corpus of Classic Sociolinguistic Interviews, Linguistic Data Consortium, September 30, 2003 (http://projects.ldc.upenn.edu/DASL/SLX/docs/transcription.pdf).

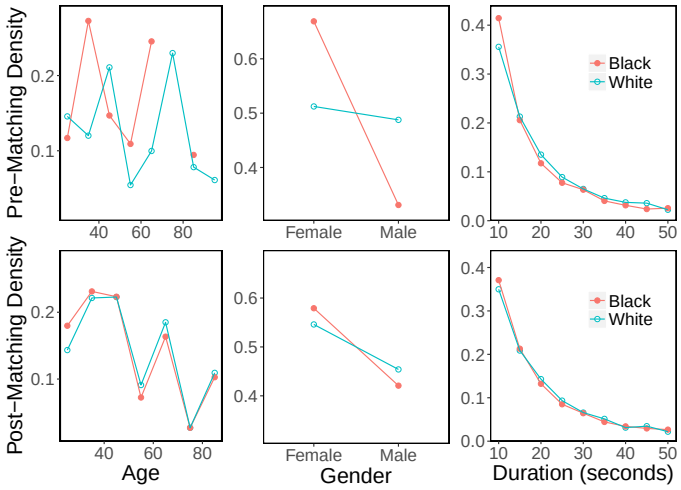

Fig. S1. Distribution of age, gender, and duration in the unmatched (top) and matched (bottom) set of audio snippets for white and black speakers. As seen in the bottom panel, our matching procedure was successful in achieving covariate balance.

words or novel words constructed spontaneously (not nonstandard or dialectal pronunciations). Unclear or unintelligible speech is marked with double parentheses (( )) left blank if the transcriber cannot ascertain what is said, and filled if the transcriber has grounds to make a reasonable guess. Interjections follow standard spellings and non-linguistic details are represented with {BR} for audible breath, {NS} for noise, {LS} for lip smack, {CG} for cough and {LG} for laughter.

The Corpus of Regional African American Language (CORAAL) uses transcription guidelines adapted from the Sociolinguistic Archive and Analysis Project (SLAAP) User Guide (June 2009) (https://slaap.chass.ncsu.edu/userguide/). In an effort to facilitate browsing and searching of interviews, time-aligned transcripts were created in Praat using TextGrid annotation objects. After the interview was transcribed, the transcription was then reviewed by two additional transcribers for accuracy and reliability. Transcribers used standard orthography and punctuation along with conventional capitalization. The hyphen was used to indicate lexical and intonation restarts and incomplete intonation. Numbers and abbreviations are completely written out (except personal titles such as *Mr.* and *Dr.*). Unintelligible or inaudible speech is notated by slashes that may or may not have the transcriber's best guess. Nonlinguistic and/or metalinguistic noises are enclosed in angled brackets. Phonological processes are not marked, but rather standard orthography is used; morphosyntactic features are transcribed as heard on the recording (e.g., possessive -s absence and possessive *they*). Due to the open access and IRB regulations of CORAAL, the interviews have been redacted; information that has been obscured includes any potential identifying information such as real names, places of work, addresses, and schools. Redacted information is replaced with a tone that has mean pitch and amplitude of the obscured speech enclosed in slashes with a redaction code (e.g., /RD-1/). We note that removing audio snippets containing such unintelligible sounds does not qualitatively change resulting WERs. For more information, refer to the full transcription conventions of CORAAL in the CORAAL user guide (http://lingtools.uoregon.edu/coraal/userguide/CORAALUserGuide_current.pdf). The detailed conventions account for nonstandard pronunciations, dialect specific items as well as local lexical items.

Matching. Prior to matching, we saw substantial differences in the age and gender distributions of our base set of 4,445 snippets of black speakers and 4,372 snippets of white speakers. In particular, as indicated in the top row of Figure S1, our sample of black speakers had a higher proportion of women. However, as shown in the bottom panel of Figure S1, we achieved near-perfect demographic alignment in our matched sample comprised of 2,141 snippets of black speakers and an equal number of snippets of white speakers. Our matching procedure approximately halved the number of snippets in our main analysis, but also allowed us to more rigorously assess racial disparities in the accuracy of speech recognition systems.

Robustness checks. In our main analysis, we assessed racial disparities by comparing differences in average error rates between white and black speakers in our matched sample of audio snippets. That approach is simple, intuitive, and appropriately adjusts for demographic differences in the raw, unmatched dataset. Here we augment that approach by reporting the results of several different regression models fit on the matched dataset. In particular, for each of the five ASR services we examined—and separately for the subset of male and female speakers—we regressed the word error rate of the machine-generated transcript on a binary variable indicating whether the speaker is black, with control variables for the age of the speaker and the natural log of the snippet duration, in seconds. In total, we thus fit 10 linear regression models on the matched data. The results are shown in Table S1 below. In all cases, we find large racial disparities, in line with our primary analysis.

With these fitted regression models, we can also estimate word error rates for hypothetical individuals of a given age on a snippet of given duration. Consider, for example, a typical 45-year-old speaking for 30 seconds. With Google's ASR—which has overall performance close to the average—we estimate an average word error rate of 0.37 for a black man and 0.23 for a black woman, compared to 0.19 for a white man and 0.14 for a white woman. Whereas error rates for white men and women are relatively similar, there is a substantial performance gap between black men and women. As discussed in the main text, this pattern is likely related to the more frequent use among black men of linguistic features characteristic of AAVE speech.

Table S1. Relationship between error rates and race, by gender

	ASR Word Error Rate Among Women				
	Apple	IBM	Google	Amazon	Microsoft
	(1)	(2)	(3)	(4)	(5)
Black Speaker	0.18^{***}	0.16^{***}	0.09^{***}	0.12^{***}	0.10^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Age	0.00	0.00^{***}	-0.00^{*}	0.00^{***}	0.00^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Log Duration	-0.05^{***}	-0.06^{***}	-0.03^{***}	-0.05^{***}	-0.04^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.34^{***}	0.28^{***}	0.26^{***}	0.23^{***}	0.19^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Observations	2,409	2,409	2,409	2,409	2,409
R^2	0.20	0.20	0.08	0.17	0.13
Adjusted R^2	0.20	0.20	0.08	0.17	0.13

	ASR Word Error Rate Among Men				
	Apple	IBM	Google	Amazon	Microsoft
	(1)	(2)	(3)	(4)	(5)
Black Speaker	0.27^{***}	0.22^{***}	0.18^{***}	0.19^{***}	0.16^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Age	0.00	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Log Duration	-0.03^{***}	-0.04^{***}	-0.03^{***}	-0.03^{***}	-0.03^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.32^{***}	0.28^{***}	0.25^{***}	0.22^{***}	0.18^{***}
	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)
Observations	1,873	1,873	1,873	1,873	1,873
R^2	0.31	0.28	0.19	0.27	0.22
Adjusted R^2	0.31	0.28	0.19	0.27	0.21
Note:					*p<0.1; **p<0.05; ***p<0.01

The results of 10 separate linear regression models, with five fit on the matched subset of female speakers (top) and five fit on the matched subset of male speakers (bottom). These models estimate word error rate (WER) as a function of a speaker's race, adjusting for the speaker's age and the snippet duration. Each column corresponds to a separate model fit on the subset of data produced by a single speech recognition service. In all cases, we see large racial disparities, with error rates for black speakers considerably higher than for white speakers.

Table S2. List of phonological / phonetic features characteristic of AAVE

Feature	Example
Final consonant cluster reduction	band → ban_
Unstressed syllable deletion (initial and medial syllables)	across → 'cross
Haplology	mississippi → misipi
Vocalization of postvocalic /l/	well → weh
Loss of /r/ after consonants, after 'th', and in unstressed syllables	throw → thow
Labialization of interdental fricatives	north → norf
Syllable initial fricative stopping	those → doze
Stopping of voiceless interdental fricatives	with → wit
Metathesis of final /s/+stop	ask → aks
Vocalization or loss of intersyllabic /r/	ordeal → ohdeal
Vocalization or stressed syllabic /r/	bird → behd
Vocalization of postvocalic /r/	for → fouh
Vocalization of unstressed syllabic /r/	never → nevuh
Glide reduction of /ai/ before voiced obstruents and finally	my → mah
Glide reduction of /ɔi/ before /l/	boil → bouh
Merger of /ɛ/ and /ɪ/ before nasals	pen → pin
Merger of tense and lax vowels before /l/	feel → fill
Fricative stopping before nasals	wasn't → wadn't
Front stressing of initial syllables	políce → pólice
Reduction of final nasal to vowel nasality	man → mah
Final consonant deletion	when → wheh
Final stop devoicing	bad → bat
Loss of /j/ after consonants	hyoosten → hoosten
Substitution of /k/ for /t/ in /str/ clusters	street → skreet
Raising of dress vowel	bet → bait
Raising of kit vowel	hit → heat or kit → keet
Deletion of initial /d/ and /g/ in certain tense-aspect auxiliaries	I don't know → ah 'on know; I'm gonna do it → ah'm 'a do it

Table S3. List of grammatical features characteristic of AAVE

Feature	Example
Copula absence	They gone
Invariant habitual *be*	She be walking (regularly)
future *be*	He be here tomorrow
Unstressed *been* for present perfect	I just been stuck
Stressed *been* to mark remote aspect	She been dead (for many years)
Completive *done*	He done did it
Resultative and future/conditional perfect *be done*	She be done had her baby
Immediate future *finna*	He finna go
Come as expression of indignation	He come walking in here like he owns the place
Simple past tense *had*	Then we had went outside
Double modals	She might can pick you up
Quasi modals: *liketa* and *poseta*	You don't poseta do it that way
Absence of third person singular present tense -s	He walk
Generalization of *is* and *was* to use with plural and second person subjects	They is some crazy folks; they was going behind the bend
Past tense as past participle	She had bit
Past participle as past tense	She bitten him
Verb stem as past tense	They come up to me
Double tense marking	She likeded the party
Absence of possessive -s	John house
Absence of plural -s	Two boy
And them to mark associate plurals	Maria and them
Appositive or pleonastic pronouns	That teacher, she yells at the kids
Y'all and *they* to mark second personal plural and third plural possessive	It's y'all ball; it's they house
Use of object pronouns after a verb as personal datives	Ahma get me a gig
Absence of relative pronoun	That's the man come here
Use of *ain't* as a preverbal negator	We ain't putting up with this
Negative concord	Not really learning nothing
Negative inversion	Can't nobody say nothing
Ain't but and *don't but* for "only"	He ain't but fourteen years old
Direct questions without inversions	Who else we had on that team?
Auxiliary verb inversion in embedded questions	I asked him could he go home with me
Existential *it* instead of *there*	It was so many movies
Existential *they got* for *there are*	They got a lot of bugs here
Here go as static locative or presentational form	Here go my own
Say to introduce quotation or verb complement	Now I say growing up with my parents...