CLICK THIS LINK FOR ACCESS THAT INCLUDES GRAPHICS

https://www.consumerreports.org/disability-rights/auto-captions-often-fall-short-on-zoom-facebook-and-others-a9742392879/

Lost in Transcription: Auto-Captions Often Fall Short on Zoom, Facebook, Google Meet, and YouTube

New research found errors that pose hurdles for users who are deaf or hard of hearing, or whose first language isn't English

ILLUSTRATION: KLAWE RZECZY

August 24, 2022

By Kaveh Waddell

Data visualizations by Andy Bergmann

- 82 shares of the article

- shares of the article

When William Albright watches tutorial videos about coding on YouTube, sometimes the presentations stop making sense. That's because Albright, a software developer in San Francisco who is deaf, uses automatic captions to follow what people are saying in videos—and those captions are often awash in errors.

"Sometimes it's just gibberish," he says.
"I can deal with 'SQL' spelled as 'sequel,' " says Albright, referring to a widely used programming language. "But if the whole sentence is littered with mistakes like that, and the material itself is supposed to be challenging, it's hard to stay focused."

Mistakes in auto-captions plague many leading videoconferencing and social media apps, according to a new study by researchers at Northeastern University and Pomona College, who worked with Consumer Reports to test auto-captions in seven popular products.

All of the programs made some mistakes, with some getting about 1 in 10 words wrong. And the results were even worse when English wasn't the speaker's first language—even if they were fluent.

Transcription Error Example

20 transcription errors per 100 words

In this excerpt from a TED Talk transcribed by Zoom, the Polish-born French–American mathematician Benoit Mandelbrot discusses fractals found in nature.

These flaws can cause problems for people who are deaf or hard of hearing and often use auto-captions to participate in work meetings and one-on-one calls or to just enjoy online entertainment. They also matter for people who aren't completely comfortable with English, who might use captions to keep up with a recorded college lecture or watch how-to videos on YouTube or Facebook.

An estimated 11.4 million Americans ages 5 and older have a hearing disability, and over 25 million Americans ages 5 and older don't speak English very well, according to census data.

"If captions are wildly inaccurate, you can't figure out what the whole conversation is about," says Christian Vogler, PhD, director of the Technology Access Project at Gallaudet University. Vogler, who is deaf, advocates for technology that works better for people with hearing disabilities. "If there are too many errors, you might not even be able to figure out the gist," he says. "And it also imposes a cognitive load trying to compensate for errors and puzzling out the meaning."

For Albright, an error-ridden transcript can be a deal-breaker. He often has to give up on a video that's poorly captioned and go looking for a more accessible alternative.

It's about "equal access," he says. "Hearing folks wouldn't accept unintelligible audio. Same principle."

Tracking Auto-Caption Errors

It's hard for software to get transcriptions exactly right. Any given recording might have patchy audio quality or people speaking over one another. Plus, even just among English speakers, there are countless accents, dialects, and regional quirks to account for.

To simplify things for our testing, CR and our partners at Northeastern and Pomona settled on audio from the popular TED lecture series. (CR funded the study, which is currently under peer review ahead of publication.)

TED Talks are recorded professionally, and most feature one speaker at a time who is trying to communicate clearly, making the audio a relatively easy challenge for automated captioning systems. (We didn't use any clips that included multiple speakers, music, or video playback.) Because they are consistent, TED Talks let us compare the talks against one another to see if the software had a harder time transcribing people from a particular racial group, age, or gender.

We used seven popular software platforms to generate transcripts for nearly 850 TED Talks, delivered by a broad range of speakers. Then we compared the computer-generated captions with official transcripts prepared manually by trained TED volunteers. To see how the auto-captions did, we tallied the number of times the computer-generated transcripts differed from the official human-written ones.

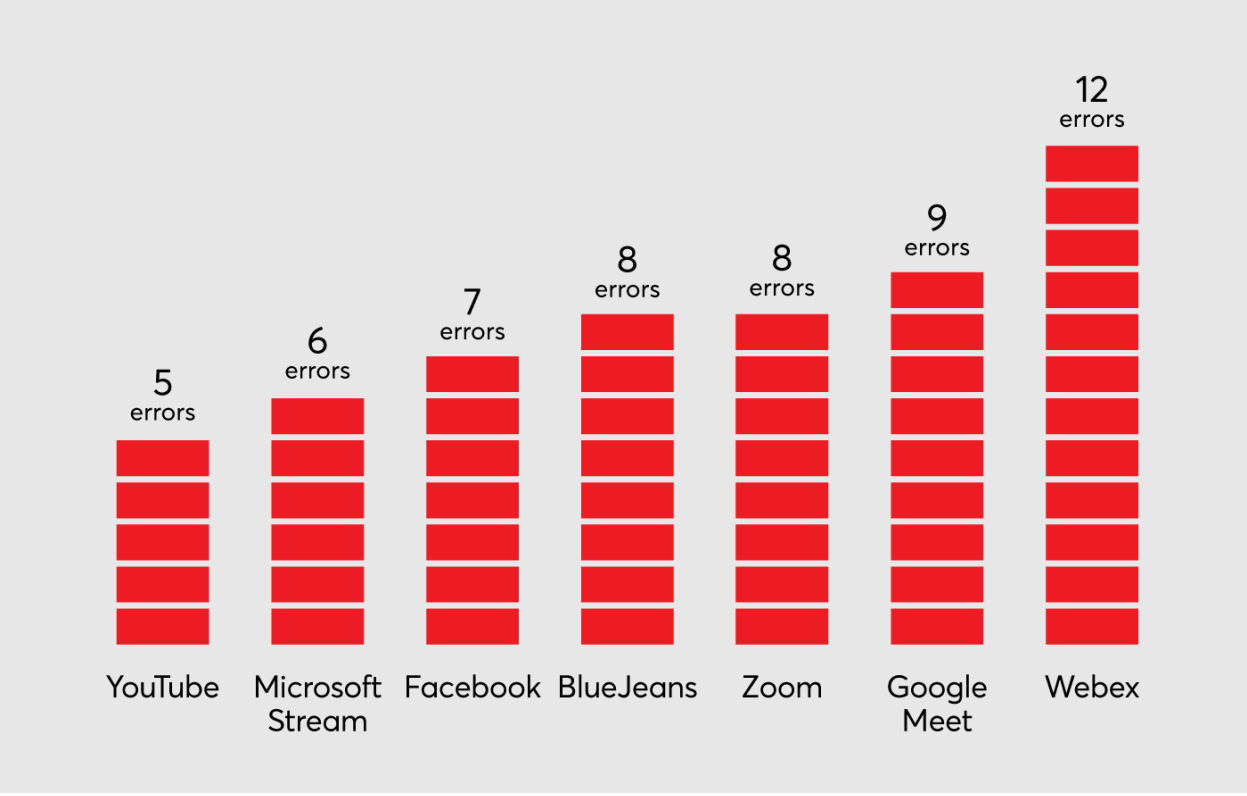
Our study included four videoconferencing platforms: BlueJeans, Cisco Webex, Google Meet, and Zoom. We also looked at Microsoft Stream, the company's video streaming service, because it uses the same voice technology as Skype and Microsoft Teams. And we tested Facebook and YouTube. (CR posts videos to YouTube and our own website. We did not test the captioning software we use on our site because it's not a consumer product.)

Overall, the platforms made a lot of mistakes, even though TED Talks were a best-case scenario. "We threw them a softball and they're still not knocking it out of the park," says David Choffnes, PhD, an associate professor of computer science at Northeastern who was one of the researchers.

But some products performed better than others. Webex had more mistakes than Google Meet, for instance, and YouTube did better than Facebook. (See the graph below.) Even within each platform, however, there were big differences. For Zoom, for example, the very best transcription had just two errors per 100 words, while at its worst the software mistranscribed nearly every third word.

System Transcription Errors

Errors Per 100 Words



Note: The differences among BlueJeans, Zoom, and Google Meet are not statistically significant.

Next, we ran statistical models to determine whether any characteristics of the various speakers explained the variation in the error rates. We controlled for the speaker's age, gender, race and ethnicity, first language, and speech rate. As it turned out, only gender and first language status independently affected the variation in transcription mistakes.

Though the accuracy differences we found between groups of speakers may not seem large, they can have a real impact on comprehension.

Take Zoom's average accuracy gap between a native and non-native speaker: It's 3.6 percent, which looks like a small number. But imagine if you misunderstood three or four extra words out of every hundred—on top of the roughly 8 percent of words that the auto-captions already bungled, according to our study. English is often spoken at about 150 words per minute, so those mistakes can pile up fast.

That means that an auto-caption user is less likely to be able to understand people whose native language isn't English, whether that's a colleague, a teacher, or a YouTube personality.

"Some groups are at a disadvantage in situations where they need captions: less authenticity, more potential for miscommunication, and by extension more potential for major screw-ups," says Gallaudet's Vogler.

What the Companies Said

We asked representatives from all seven products to comment on our findings.

A spokesperson for YouTube said that our results matched up with the company's "expectations for performance," and that the company is working on improving YouTube "so it works better for everyone," including by working with linguists.

Microsoft said our findings are roughly in line with its internal testing, which also reveal lower accuracy when transcribing men and second-language English speakers. The spokesperson said Microsoft regularly assesses caption accuracy by gender, age, language variety, and regional and foreign language accents.

A Zoom spokesperson said, "We're continuously enhancing our transcription feature to improve accuracy toward a variety of factors, including English dialects and accents." And Google said through a spokesperson that it's working to "improve the accuracy of live captions and translations so even more users can participate and stay engaged using Google Meet."

A spokesperson for Verizon, which owns BlueJeans, said that the company uses third-party software for its captions, but it declined to name the software provider. The spokesperson said, "We have the ability to train our models to improve accuracy and are always evaluating ways to do so."

Cisco, which owns Webex, says its own auto-caption testing puts Webex ahead of two "best-in-class speech recognition engines," but wouldn't say which products those were. That's in contrast to our study, where Webex had the highest error rate. A Cisco spokesperson said the discrepancy may be explained by the fact that Webex's captions are fine-tuned for videoconferencing rather than other scenarios like TED-style lectures.

Meta, Facebook's parent company, declined to comment on CR's findings, but a spokesperson pointed to recent work from its AI research team that found accuracy gaps by speaker gender and skin tone. (The research was not performed on Facebook's own auto-caption technology.)

More on Video and Speech Technology

Apple's New Voices Resonate With Some Black iPhone Users

When People Think Siri Sounds Black, Biases Often Surface

Guide to Videoconference Services

It's Not Just Zoom. Google Meet, Microsoft Teams, and Webex Have Privacy Issues, Too.

How Speech Technology Can Go Wrong

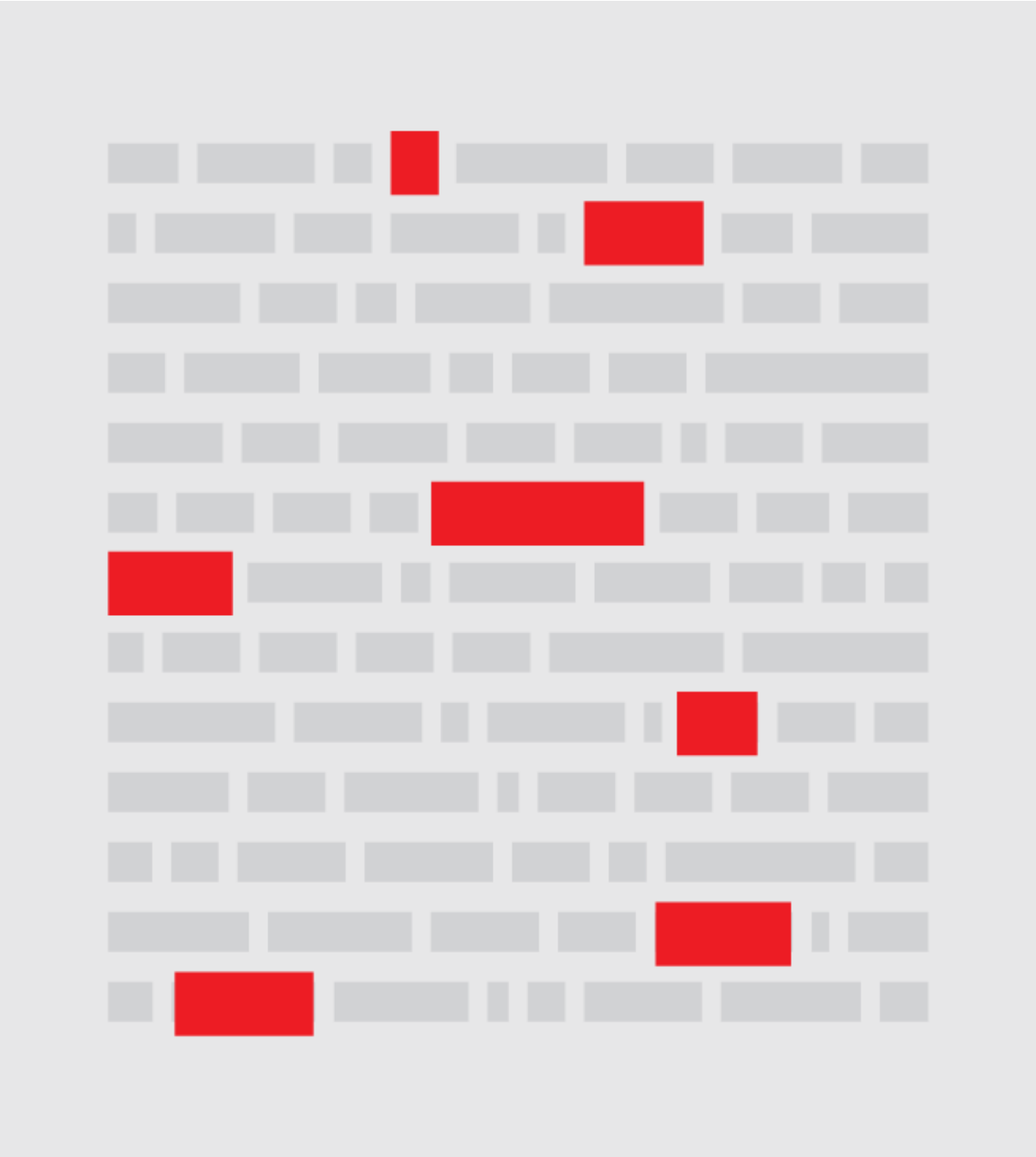
Experts say accuracy gaps like the ones we found usually arise when a system hasn't been taught to understand a broad variety of English speakers.

Speech recognition systems learn to interpret spoken language by training on enormous datasets of recorded speech and transcripts. That allows the systems to match patterns in real-world speech with patterns it learned from the training data, and produce transcripts without human intervention.

But if a system was trained on data that mostly featured a certain type of speaker—people who grew up speaking English, say, or white American English speakers—it may end up lagging when it tries transcribing other types of English speakers in the real world.

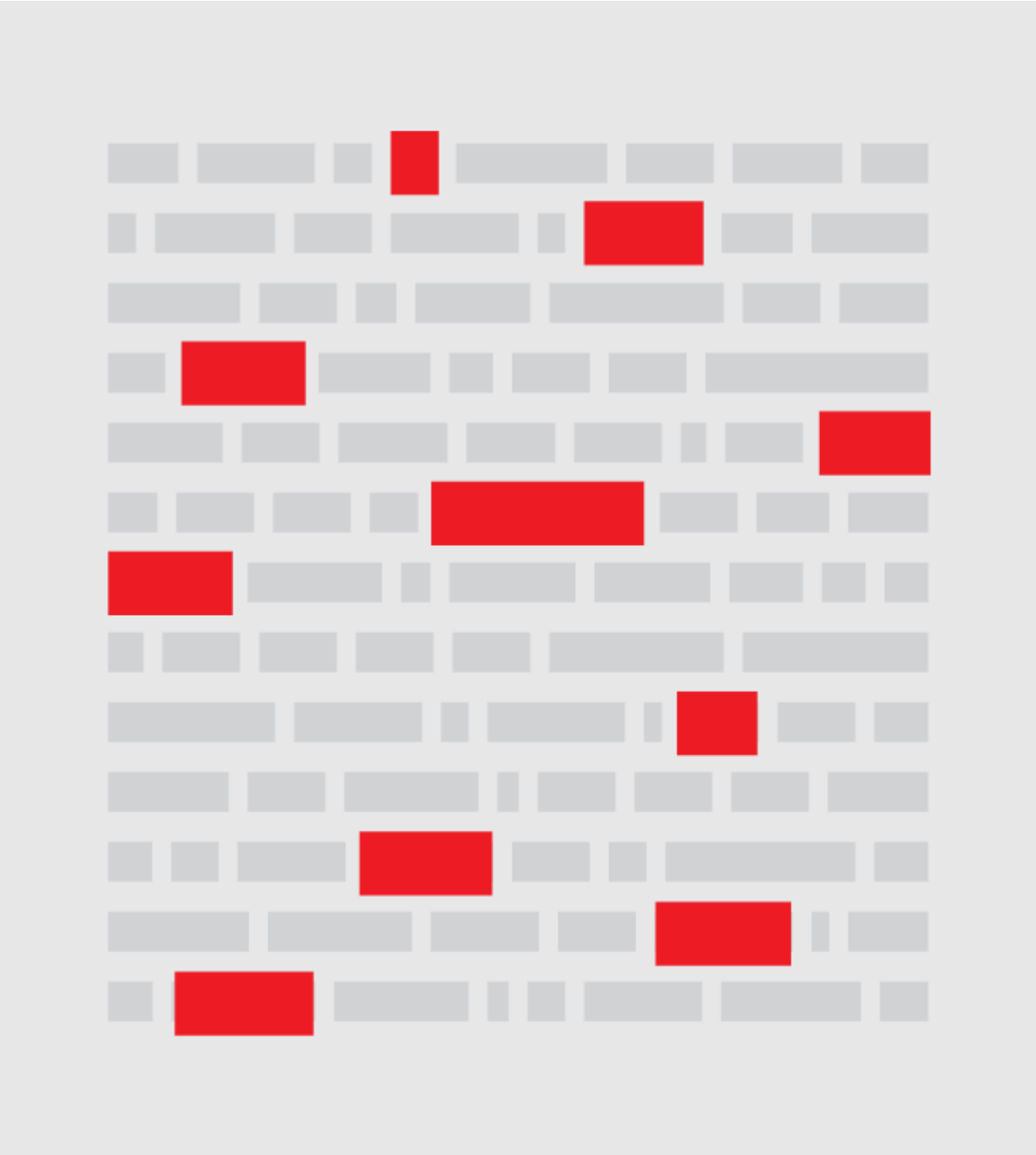
Native English Speakers

7 transcription errors per 100 words



Non-Native English Speakers

10 transcription errors per 100 words



This technology powers much more than auto-captions. It's also behind digital assistants like Siri and Alexa, dictation services that let you speak a text message rather than typing it, and call-center software that tries to understand spoken commands from consumers. Companies are looking for more opportunities to let people control their gadgets through voice instead of typing and swiping. And as that happens, any shortcomings with speech-to-text software will become even more important.

In our research, only gender and first language affected auto-caption accuracy, but other academic studies have found differences correlated with race and ethnicity.

In a 2020 study, for example, Stanford researchers found that major speech recognition products misunderstood Black users at nearly twice the rate that they misunderstood white users. The almost twofold disparity existed in products from all five companies they studied: Apple, Amazon, Google, IBM, and Microsoft.

And in a pair of papers published in 2017, Rachael Tatman, who has a PhD in linguistics from the University of Washington, found some significant differences in YouTube's auto-caption accuracy based on a speaker's regional dialect.

It's likely these differences didn't appear in our research because we used TED Talks to test caption quality, says Nicole Holliday, PhD, an assistant professor of linguistics at Pomona College and co-author of the CR-associated study. Much of the variation that exists in people's informal speech disappears when they're speaking from a stage.

"The speech style used by TED talkers is very formal and very rehearsed, and it's recorded with very high audio quality," Holliday says. "This means that in the real world, the effects that we found are likely to be amplified. It also means that there exist other biases that we didn't find, related to the speech style."

The hyperformal setting also helps explain the gender differences in accuracy, Holliday and Tatman say. Linguistics research has shown that generally women tend to use more "standardized" language. That means they may be less likely to use informal or socially stigmatized speech patterns that often cause problems for speech recognition systems.

Building Better Auto-Captioning

Research like CR's should prod tech companies to improve their speech-recognition systems, experts say.

"In some tech spheres, there's an incorrect belief that automatic speech recognition is a 'solved problem,' " Holliday says. "But in reality, it works much better for a very specific type of idealized speaker who is speaking in a very formal style. The vast majority of the time when people are speaking, they're not fitting the mold that companies assume."

That doesn't mean it's an easy fix. Auto-caption systems have become steadily better in recent years, according to experts and people who use them regularly. Improvements are increasingly costly, Tatman says, and will require better training data as well as better signal processing for filtering out background noise or zeroing in on a speaker's voice.

Some newcomers to the field are trying to chip away at the problem. Vogler, the Gallaudet assistive technology expert, is helping to build an auto-captioning system called GoVoBo, which is designed for computer users who are deaf or hard of hearing.

If you want to help make speech recognition work better for a broader group of people, you can donate a recording to Common Voice, a project at the nonprofit Mozilla Foundation to create a large, openly accessible database of speech samples in dozens of languages.

You don't have to stick to your first language, either. Common Voice encourages volunteers to speak in second languages to help create a training set that's less likely to be tilted against second-language speakers.

"One of the leading causes of performance bias in [automated speech recognition] and speech technology more widely is unbalanced, unrepresentative training data," says EM Lewis-Jong, product lead for the Mozilla project. "Common Voice gives people around the world a real, practical chance to mobilize their voice to address these bias issues, by contributing to a dataset and mitigating bias at the source."

Editor's note: *Our work on privacy, security, and data issues is made possible by the vision and support of the Ford Foundation, Omidyar Network, Craig Newmark Philanthropies, and the Alfred P. Sloan Foundation.*

Kaveh Waddell

Kaveh Waddell has been an investigative journalist at Consumer Reports since 2020, focusing on digital rights and environmental justice. Previously, he reported on emerging technology at Axios, covered digital privacy and surveillance at the Atlantic, and freelanced from Beirut. Follow Kaveh on Twitter @kavehwaddell.

v