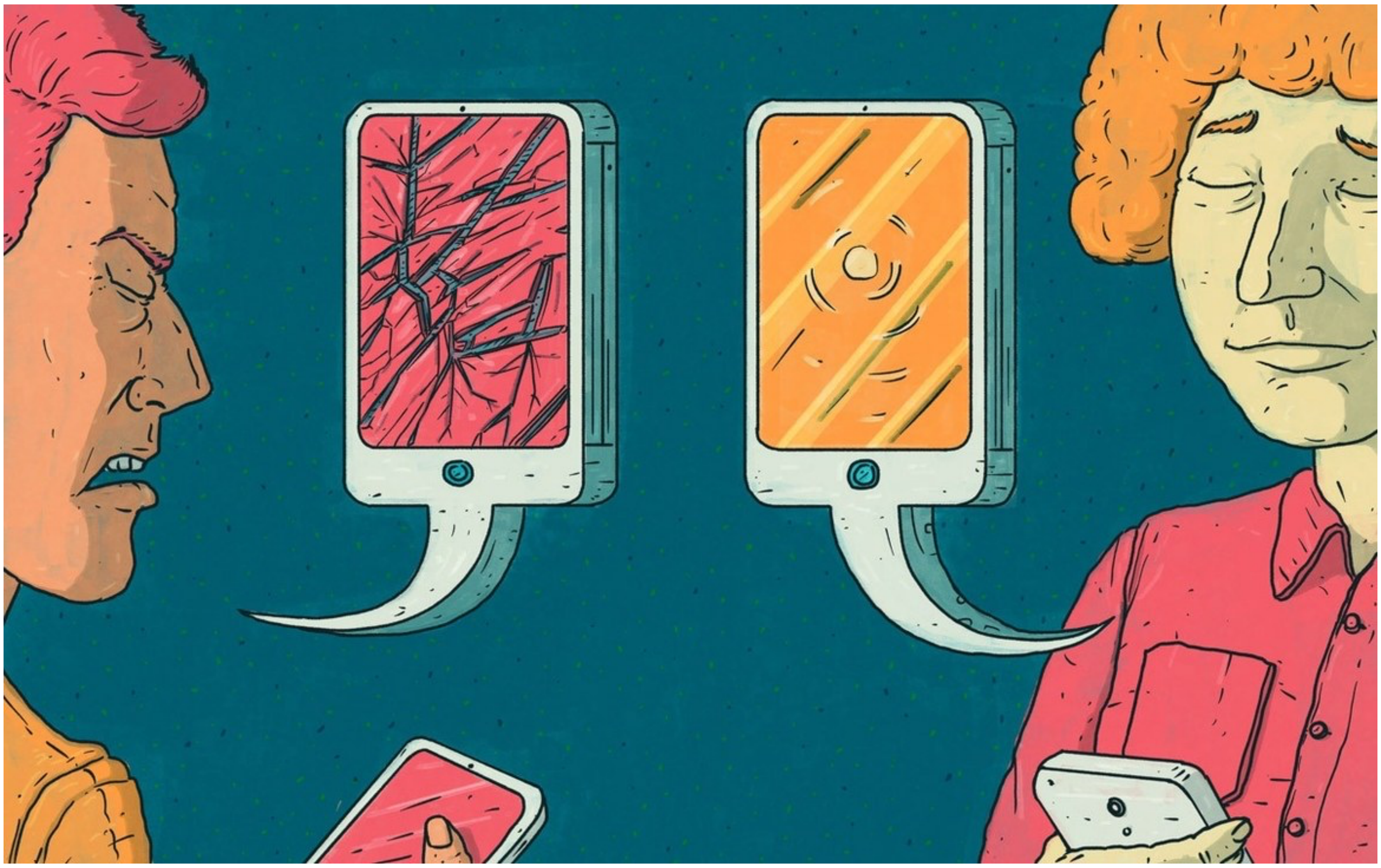
Voice Is the Next Big Platform, Unless You Have an Accent



Laurent Hrybyk

My mother waited two months for her Amazon Echo to arrive. Then, she waited again — leaving it in the box until I came to help her install it. Her forehead crinkled as I download the Alexa app on her phone. Any device that requires vocal instructions makes my mother skeptical. She has bad memories of Siri. "She could not understand me," my mom told me.

My mother was born in the Philippines, my father in India. Both of them speak English as a third language. In the nearly 50 years they've lived in the United States, they've spoken English daily — fluently, but with distinct accents and sometimes different phrasings than a native speaker. In their experience, that means Siri, Alexa, or basically any device that uses speech technology will struggle to recognize their commands.

My parents' experience is hardly exclusive or unknown. (It's even been chronicled in comedy, with this infamous trapped-in-a-voice-activated elevator sketch.) My sister-in-law told me she gave up on using Siri after it failed to recognize the "ethnic names" of her friends and family. I can vouch for the frustration: The other day, my command of "Text Zahir" morphed into "Text Zara here."

Right now, it's not much of a problem — but it's slated to become more serious, given that we are in the middle of a voice revolution. Voice-based wearables, audio, and video entertainment systems are already here. Due in part to distracted drivers, voice control systems will soon be the norm in vehicles. Google Home and Amazon's Alexa are radicalizing the idea of a "smart home" across millions of households in the US. That's why it took so long for my mother's Echo to arrive — the Echo was among Amazon's bestsellers this holiday season, with a 900 percent increase from 2016 sales. It was backordered for weeks.

Overall, researchers estimate 24.5 million voice-driven devices will be delivered to Americans' daily routines this year — evidence that underscores ComScore's prediction that by 2020, half of all our searches will be performed by voice.

But as technology shifts to respond to our vocal chords, what happens to the huge swath of people who can't be understood?

To train a machine to recognize speech, you need a lot of audio samples. First, researchers have to collect thousands of voices, speaking on a range of topics. They then manually transcribe the audio clips. This combination of data — audio clips and written transcriptions — allows machines to make associations between sound and words. The phrases that occur most frequently become a pattern for an algorithm to learn how a human speaks.

But an AI can only recognize what it's been trained to hear. Its flexibility depends on the diversity of the accents to which it’s been introduced. Governments, academics, and smaller startups rely on collections of audio and transcriptions, called speech corpora, to bypass doing labor-intensive transcriptions themselves. The University of Pennsylvania’s Linguistic Data Consortium (LDC) is a powerhouse of these data sets, making them available under licensed agreements for companies and researchers. One of its most famous corpora is Switchboard.

Texas Instruments launched Switchboard in the early 1990s to build up a repository of voice data, which was then distributed by the LDC for machine learning programs. It’s a collection of roughly 2,400 telephone conversations, amassed from 543 people from around the US — a total of about 250 hours. Researchers lured the callers by offering them long-distance calling cards. A participant would dial in and be connected with another study participant. The two strangers would then chat spontaneously about a given topic — say, childcare or sports.

For years linguists have assumed that because the LDC is located in Philadelphia, the conversations skewed towards a Northeastern accent. But when Marsal Gavaldà, the director of machine intelligence at the messaging app Yik Yak, crunched the numbers in Switchboard’s demographic history, he found that the accent pool skewed more midwestern. South and North Midland accents comprised more than 40 percent of the voice data.

Other corpora exist, but Switchboard remains a benchmark for the models used in voice recognition systems. Case in point: Both IBM and Microsoft use Switchboard to test the word error rates for their voice-based systems. “From this set of just over 500 speakers, pretty much all engines have been trained,” says Gavaldà.

But building voice technology on a 26-year-old corpus inevitably lays a foundation for misunderstanding. English is professional currency in the linguistic marketplace, but numerous speakers learn it as a second, third, or fourth language. Gavaldà likens the process to drug trials. “It may have been tried in a hundred patients, [but] for a narrow demographic," he tells me. "You try to extrapolate that to the general population, the dosage may be incorrect."

Larger companies, of course, have to think globally to stay competitive — especially because most sales of smartphones happen outside the US Technology companies like Apple, Google, and Amazon have private, in-house methods of collecting this data for the languages and accents they'd like to accommodate. And the more consumers use their products, the more their feedback will improve the products, through programs like Voice Training on the Alexa app.

But even if larger tech companies are making headway in collecting more specific data, they're motivated by the market to not share it with anyone — which is why it takes so long for the technology to trickle down. This secrecy also applied to my reporting of this piece. Amazon never replied to my request for comment, a spokesperson for Google directed me to a blog post outlining its deep learning techniques, and an Apple PR representative noted that Siri is now customized for 36 countries and supports 21 languages, language variants, and accents.

Outside the US, companies are aware of the importance of catering to accents. The Chinese search engine company Baidu, for one, says its deep learning approach to speech recognition achieves accuracy in English and Mandarin better than humans, and it's developing a "deep speech" algorithm that will recognize a range of dialects and accents. "China has a fairly deep awareness of what's happening in the English-speaking world, but the opposite is not true," Baidu chief scientist Andrew Ng told *The Atlantic*.

Yet smaller companies and individuals who can't invest in collecting data on their own are beholden to cheaper, more readily available databases that may not be as diverse as their target demographics. "[The data's] not really becoming more diverse, at least from my perspective," Arlo Faria, a speech researcher at the conference transcription startup Remeeting, tells me. Remeeting, for example, has used a corpus called Fisher that includes a group of non-native English speakers — but Fisher's accents are largely left up to chance, depending on who happened to participate in the data collection. There are some Spanish and Indian accents, for instance, but very few British accents, Faria recalls.

That's why, very often, voice recognition technology reacts to accents differently than humans, says Anne Wootton, co-founder and CEO of the Oakland-based audio search platform Pop Up Archive, "Oftentimes the software does a better job with like, Indian accents than deep Southern, like Shenandoah Valley accents," she says. "I think that's a reflection of what the training data includes or does not include."

Rachael Tatman, a PhD candidate at the University of Washington's Department of Linguistics who focuses on sociolinguistics, noted that the underrepresented groups in these data sets tend to be groups that are marginalized in general. A typical database of American voices, for example, would lack poor, uneducated, rural, non-white, non-native English voices. "The more of those categories you fall into, the worse speech recognition is for you," she says.

Still, Jeffrey Kofman, the CEO and co-founder of Trint, another automated speech-to-text software based in the UK, is confident accent recognition is something speech science will be able to eventually solve. We video chatted on the Trint platform itself, where Australian English is now available alongside British and North American English as transcription accents. Trint also offers speech-to-text in a dozen European languages, and plans to add South Asian English sometime this year, he said.

Collecting data is expensive and cumbersome, which is why certain key demographics take priority. For Kofman, that's South Asian accents, “because there are so many people from India, Pakistan, and those countries here in England, in the US and Canada, who speak very clearly but with a distinct accent,” he says. Next, he suspects, he’ll prioritize South African accents.

Obviously, it’s not just technology that discriminates against people with accents. It’s also other people. Mass media and globalization are having a huge effect on how people sound. Speech experts have documented the decline of certain regional American accents since as early as 1960, for example, in favor of a more homogenous accent fit for populations from mixed geographic areas. This effect is exacerbated when humans deal with digital assistants or operators; they tend to use a voice devoid of colloquialisms and natural cadence.

Or, in other words, a voice devoid of an identity and accent.

As voice recognition technology becomes better, using a robotic accent to communicate with a device stands to be challenged — if people feel less of a need to talk to their devices as if they are machines, they can start talking to them as naturally as they would a friend. And while some accent reduction coaches find their clients use voice assistants to practice neutralizing their thick foreign or regional accents, Lisa Wentz, a public speaking coach in San Francisco who works in accent reduction, says that she doesn’t recommend it.

That’s because, she tells me, most of her clients are aiming for other people to understand them. They don’t want to have to repeat themselves or feel like their accents prevent others from hearing them. Using devices that aren’t ready for different voices, then, only stands to make this feeling echo.

My mother and I set up her Alexa app together. She wasn’t very excited about it. I could already imagine her distrust and fear of a car purported to drive by the command of her voice. My mother would never ride in it; the risk of crashing would be too real. Still, she tried out a couple of questions on the Echo.

“Alexa, play ‘Que sera sera,’” my mother said.

“I can’t find the song ‘Kiss your ass era.’”

My mom laughed, less out of frustration and more out of amusement. She tried again, this time speaking slower, as if she were talking to a child. “Alexa, play ‘Que sera sera.’” She sang out the syllables of *sera* in a slight melody, so that the device could clearly hear “se-rah.”

Alexa understood, and found what my mom was looking for. “Here’s a sample of ‘Que sera sera,’ by Doris Day,” she said, pronouncing the *sera* a bit harsher — “se-raw.”

The 1964 hit started to play, and my mother smiled at the pleasure of recognition.