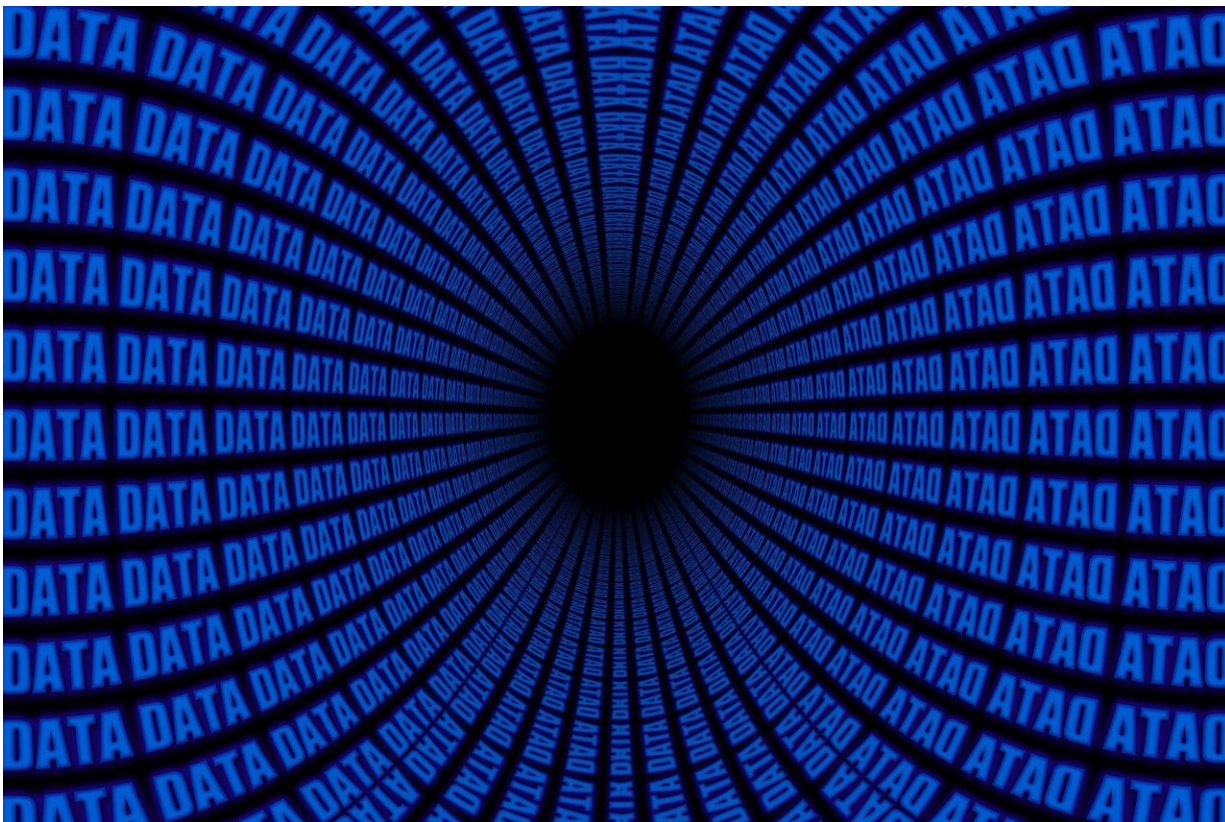# Voice technology for the rest of the world

December 17 2021, by Leah Burrows

Credit: CC0 Public Domain

Voice-enabled technologies like Siri have gone from a novelty to a routine way to interact with technology in the past decade. In the coming years, our devices will only get chattier as the market for voice-enabled apps, technologies and services continues to expand.

But the growth of voice-enabled technology is not universal. For much of the world, technology remains frustratingly silent.

"Speech is a natural way for people to interact with devices, but we haven't realized the full potential of that yet because so much of the world is shut out from these technologies," said Mark Mazumder, a Ph.D. student at the Harvard John A. Paulson School of Engineering and Applied Sciences (SEAS) and the Graduate School of Arts and Sciences.

The challenge is data. Voice assistants like Apple's Siri or Amazon's Alexa need thousands to millions of unique examples to recognize individual keywords like "light" or "off". Building those enormous datasets is incredibly expensive and time-consuming, prohibiting all but the biggest companies from developing voice recognition interfaces.

Even companies like Apple and Google only train their models on a handful of languages, shutting out hundreds of millions of people from interacting with their devices via voice. Want to build a voice-enabled app for the nearly 50 million Hausa speakers across West Africa? Forget it. Neither Siri, Alexa nor Google Home currently support a single African language.

But Mazumder and a team of SEAS researchers, in collaboration with researchers from the University of Michigan, Intel, NVIDIA, Landing AI, Google, MLCommons and Coqui, are building a solution to bring voice technology to the rest of the world.

At the Neural Information Processing Systems conference last week, the team presented a diverse, multilingual speech dataset that spans languages spoken by over 5 billion people. Dubbed the Multilingual Spoken Words Corpus, the dataset has more than 340,000 keywords in 50 languages with upwards of 23.4 million audio examples so far.

"We have built a dataset automation pipeline that can automatically identify and extract keywords and synthesize them into a dataset," said Vijay Janapa Reddi, Associate Professor of Electrical Engineering at SEAS and senior author of the study. "The Multilingual Spoken Words Corpus advances the research and development of voice-enabled applications for a broad global audience."

"Speech technology can empower billions of people across the planet, but there's a real need for large, open, and diverse datasets to catalyze innovation," said David Kanter, MLCommons co-founder and executive director and co-author of the study. "The Multilingual Spoken Words Corpus offers a tremendous breadth of languages. I'm excited for these datasets to improve everyday experiences like voice-enabled consumer devices and speech recognition."

To build the dataset, the team used recordings from Mozilla Common Voice, a massive global project that collects donated voice recordings in a wide variety of spoken languages, including languages with a smaller population of speakers. Through the Common Voice website, volunteer speakers are given a sentence to read aloud in their chosen language. Another group of volunteers listens to the recorded sentences and verifies its accuracy.

The researchers applied a machine learning algorithm that can recognize and pull keywords from recorded sentences in Common Voice.

For example, one sentence prompt from Common Voice reads: "He played college football at Texas and Rice."

First, the algorithm uses a common machine learning technique called forced alignment—specifically a tool called the Montreal Forced Aligner—to match the spoken words with text. Then the algorithm filters and extracts words with three or more characters (or two

characters in Chinese). From the above sentence, the algorithm would pull "played" "college" "football" "Texas" "and" and "Rice." To add the word to the dataset, the algorithm needs to find at least five examples of the word, which ensures all words have multiple pronunciation examples.

The algorithm also optimizes for gender balance and minimal speaker overlap between the samples used for training and evaluating keyword spotting models.

"Our goal was to create a large corpus of very common words," said Mazumder, who is the first author of the study. "So, if you want to train a model for smart lights in Tamil, for example, you would probably use our dataset to pull the keywords "light", "on", "off" and "dim" and be able to find enough examples to train the model."

"We want to build the voice equivalent of Google search for text and images," said Reddi. "A dataset search engine that can go and find what you want, when you want it on the fly, rather than rely on static datasets that are costly and tedious to create."

When the researchers compared the accuracy of models trained on their dataset against models trained on a Google dataset that was manually constructed by carefully sourcing individual and specific words, the team found only a small accuracy gap between the two.

For most of the 50 languages, the Multilingual Spoken Words Corpus is the first available keyword dataset that is free for commercial use. For several languages, such as Mongolian, Sakha, and Hakha Chin, it is the first keyword spotting dataset in the language.

"This is just the beginning," said Reddi. "Our goal is to build a database with 1,000 words in 1,000 different languages."

"Whether it's on Common Voice or YouTube, Wikicommons, archive.org, or any other creative commons site, there is so much more data out there that we can scrape to build this dataset and expand the diversity of the languages for voice-based interfaces," said Mazumder. "Voice interfaces can make technology more accessible for users with visual or physical impairments, or for lower literacy users. We hope free datasets like ours will help assistive technology developers to meet these needs."

The corpus is available on MLCommons, a not-for-profit, open engineering consortium dedicated to improving machine learning for everyone. Reddi is Vice President and a board member of MLCommons.

  More information: Presentation: datasets-benchmarks-proceeding … ae2-Paper-round2.pdf

Dataset: mlcommons.org/en/multilingual-spoken-words/

Provided by Harvard John A. Paulson School of Engineering and Applied Sciences