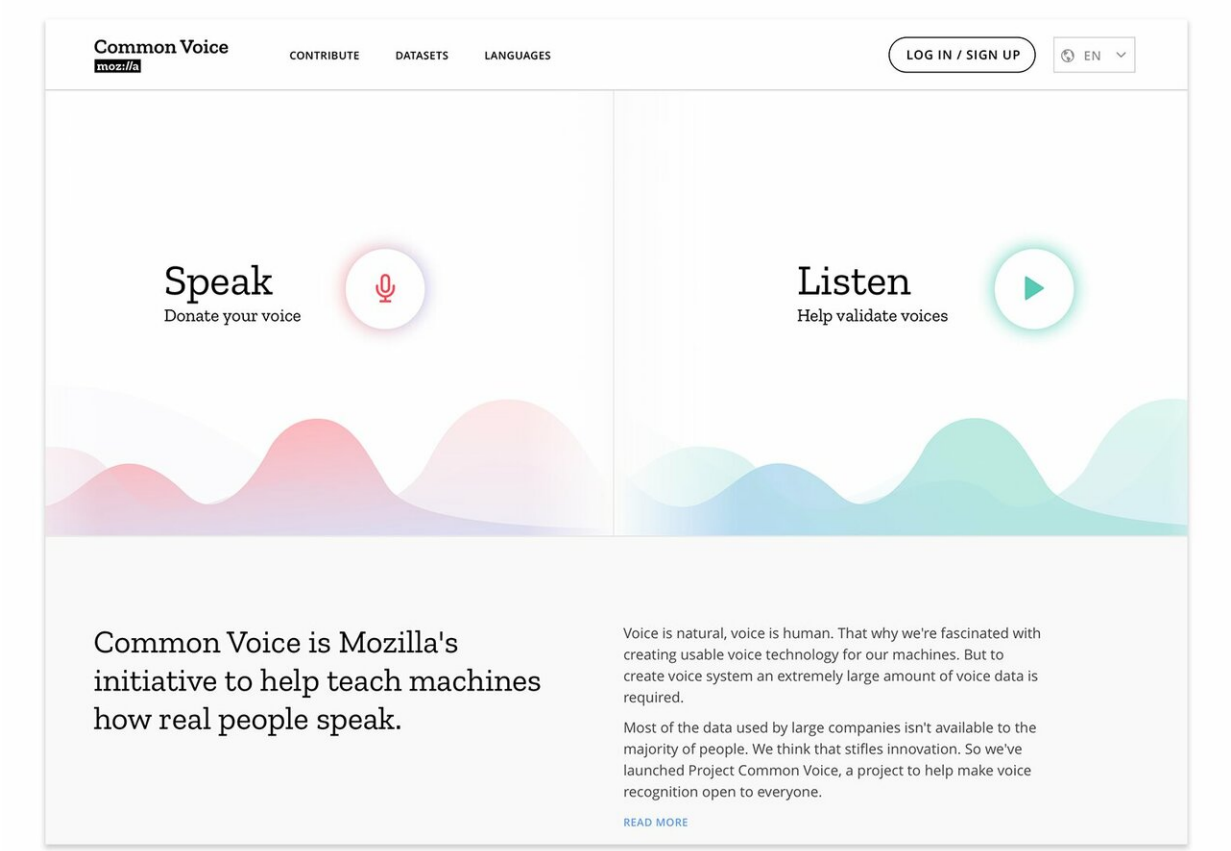
Mozilla initiative helps voice technology players via multi-language dataset

March 1 2019, by Nancy Cohen



This may sound like a mouthful but it really means much. Mozilla is talking about the "largest to-date public domain transcribed voice

dataset." Translation: Over 14,000 people. In 18 languages. Of almost 1,400 hours (1,368 to be exact) of recorded voice. Welcome to an initiative dubbed Common Voice.

This is what the Mozilla announcement said, in the form of a blog on Thursday from George Roter.

"Today, we're excited to share our first multi-language dataset with 18 languages represented, including English, French, German and Mandarin Chinese (Traditional), but also for example Welsh and Kabyle. Altogether, the new dataset includes approximately 1,400 hours of voice clips from more than 42,000 people."

Contributors to the project have professional specialties that range from doctoral candidates in speech recognition to machine learning scientists to a professor of computational linguistics. As such, the effort represents a global community of voice contributors along with what Mozilla credited as "passionate volunteers."

The purpose of Common Voice is to help teach machines how real people speak. In brief, it has evolved into a massive collection of voice clips in dozens of languages. What's next: The full dataset will be available for download on the Common Voice site.

It looks as if the Mozilla team's contributors also worked out the inevitable pain points. The blog mentioned those points. "People who contribute not only see progress per language in recording and validation, but also have improved prompts that vary from clip to clip; new functionality to review, re-record, and skip clips as an integrated part of the experience; the ability to move quickly between speak and listen; as well as a function to opt-out of speaking for a session."

Sounds like fun or an academic sandbox but actually there are more solid

aspirations among those who have contributed to building its corpus.

In 2019, Mariella Moon in *Engadget* has noticed the range of languages now included Dutch, Hakha-Chin, Esperanto, Farsi, Basque, Spanish, French, German, Mandarin Chinese (Traditional), Welsh and Kabyle.

TechRadar's Olivia Tambini, said, "By providing a huge library of human voices in a range of languages for free, Mozilla could be opening the doors for companies that don't have the resources of Apple, Amazon, and Google, to develop their own voice assistants."

Another benefit involves Mozilla itself. Mariella Moon in *Engadget* said, "The organization itself plans to use the clips it collects to improve its Speech-to-Text, Text-to-Speech and DeepSpeech engines."

Roter said, plain and simple, "Our goal is to both release voice-enabled products ourselves, while also supporting researchers and smaller players ."

Note the bragging rights belong to it being the largest, not the only, dataset of its kind. Mozilla wanted site visitors to know that it was the largest, not the only, and also said that in time site visitors can "Look to this page as a reference hub for other open source voice datasets."

If you visit the Common Voice site you get the message about their keen ambition. "We're building," said Mozilla. And what are they building? An "open source, multi-language dataset of voices that anyone can use to train speech-enabled applications."

Contributors can opt-in to provide metadata like their age, sex, and accent. Voice clips in turn are tagged with information useful in training speech engines.

More information: blog.mozilla.org/blog/2019/02/ … ribed-voice-dataset/

voice.mozilla.org/en/datasets