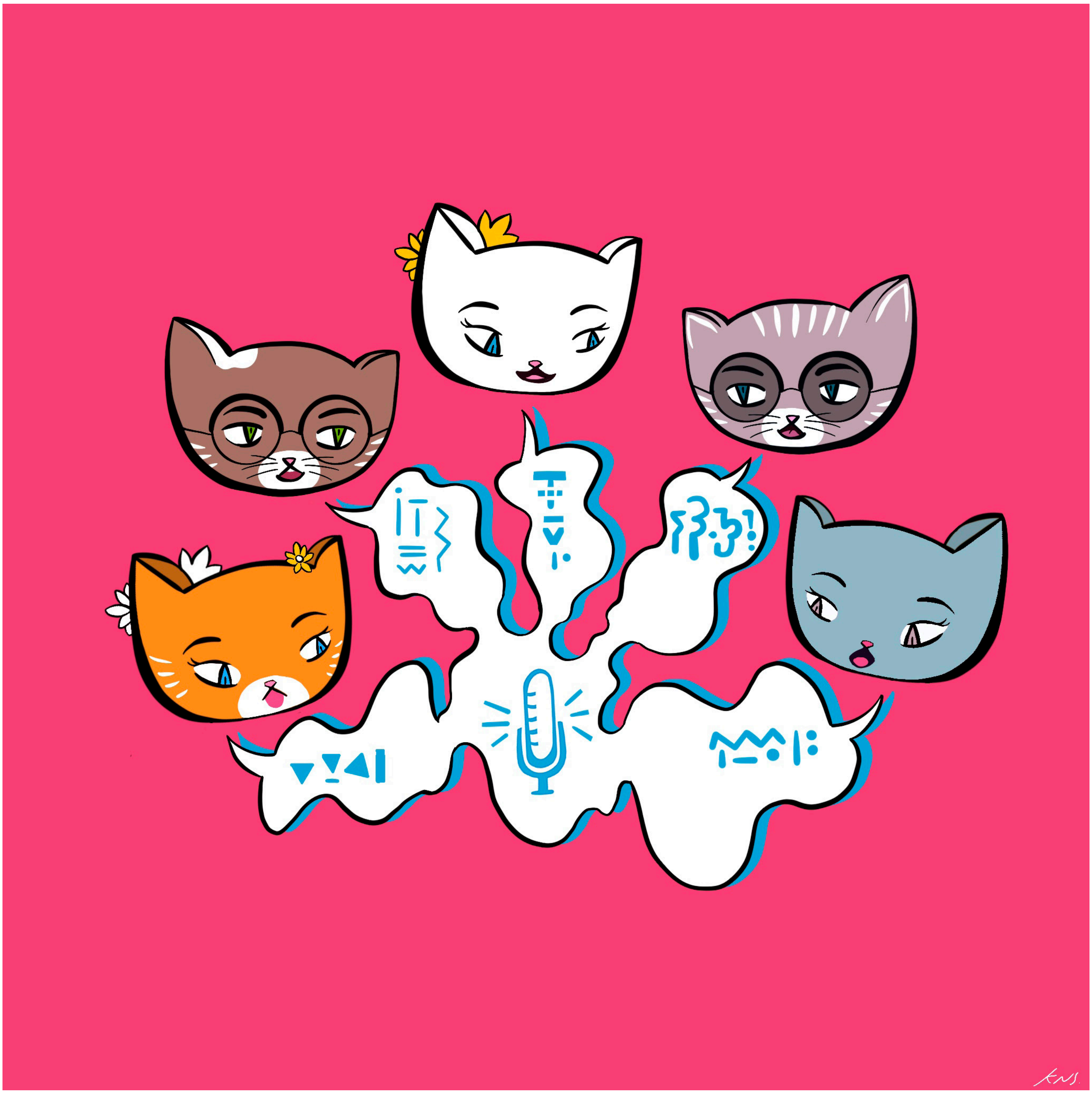

MAKING VOICES HEARD

COMMON VOICE

Case Study

Making Voices Heard
Case Study: Common Voice

Research and Writing **SHWETA MOHANDAS, SAUMYAA NAIDU**
Review and Editing **PUTHIYA PURAYIL SNEHA, TORSHA SARKAR**
Research Inputs **SUMANDRO CHATTAPADHYAY**
Copyediting **THE CLEAN COPY**
Illustration **KRUTHIKA N.S.**
Report Layout and Design **SAUMYAA NAIDU**

CENTRE FOR INTERNET AND SOCIETY
Supported by Mozilla Corporation

Contents

1. About

'... to make voice data freely and publicly available, and make sure the data represents the diversity of real people.'[1]

Common Voice (CV) is an open-source dataset of voice recordings in multiple languages that can be used to train speech-enabled applications. With over 13,905 hours of voice data across 76 different languages as of July 2021,[2] CV strives to create and maintain the largest publicly available voice dataset of its kind. CV believes that the availability of large public voice datasets will help foster innovation and create a healthy market for machine-learning-based speech technologies. In May 2020, CV began data collection for a single-word target segment (the recording of single words in multiple languages) or voice data for single-word sentences (for example yes and no), to be deployed for specific use cases or purposes. The exercise has begun with the digits zero through nine, as well as the words yes, no, hey and Firefox".[3]

2. Methodology and process

CV follows a community-driven model of creating an open-source, multilingual dataset of voice recordings that is openly accessible and usable. At the same time, it has also been working on and navigating various aspects related to privacy of voice data and accessibility for persons with disabilities, which also include complex design challenges and decisions. Some key features of this initiative include:

2.1. Community-driven contribution

"... Providing more and better data to everyone in the world who seeks to build and use voice technology."[4]

Although CV began with creating a voice dataset for English, as most of the team working on it was English-speaking, as of 2021 there are over 76 languages on the platform. CV depends on a community of volunteers and individual users who contribute voice data in order to add new languages to its website and system. One way CV promotes localisation is by localising its website to the languages it wants to add. Before adding a new language, the community has to localise 85% of the website, so that when volunteers from the local language community visit the website, they can easily navigate it, and do not need to rely on English.

[1] "Common Voice by Mozilla," *Common Voice*, accessed January 4, 2022, https://commonvoice.mozilla.org/en/about.

[2] "Common Voice by Mozilla." *Common Voice*, accessed January 4, 2022, https://commonvoice.mozilla.org/en/datasets.

[3] Branson, M., "Help Create Common Voice's First Target Segment," *Discourse*, 12 May 2020, https://discourse.mozilla.org/t/help-create-common-voices-first-target-segment/59587, 3 November 2021.

[4] Roter, G., "Sharing Our Common Voices – Mozilla Releases the Largest to-date Public Domain Transcribed Voice Dataset," *The Mozilla Blog*, 9 February 2021, accessed 3 November 2021, https://blog.mozilla.org/en/mozilla/news/sharing-our-common-voices-mozilla-releases-the-largest-to-date-public-domain-transcribed-voice-dataset/

Then, when the language is active on the site, it is up to the community to submit 5,000 sentences that have been recorded in that language. This indicates two things to CV: a) that there is an active language community that can provide voice recordings, and b) that the barrier to including the language in CV is fairly low.

The recorded material is based on a sentence corpus that CV provides; everybody on the platform is presented with sentences that they can record and submit. These include content such as parliamentary transcripts, Wikipedia articles, and sentences that members of the community have submitted. Two other community members then check to see if the audio matches the sentences. Though this is not a foolproof system, CV reports that it has a rather high accuracy rate. If people record things that are not on the card, they get voted down very quickly. This system of community curation and regulation, therefore, adds a layer of control to the accuracy and quality of content.

"Amazon and Apple, by necessity, choose languages based on what makes sense in the market and makes the most profit."[5]

Key players in the voice-as-product market serve more widely spoken languages, such as English, French, and German, because they have a large user base and hence greater demand. The issue occurs with underrepresented languages, uncommon accents, or the voices of people from underrepresented/marginalised groups – such as those belonging to particular ethnic or gender identities. As a result, large populations remain unrepresented in datasets used to train commercial voice technologies and products. This is the gap that CV is striving to diminish.

CV's data collection differs from that of start-ups and companies like Google and Amazon; here, the sentences are self-recorded by people, and CV does not automatically detect the individual's identity, location, or other data. It does not infer the contributor's demographic based on their browsing data. Community members are also instructed not to identify people who are in the dataset.

2.2. Design process and development

Since it was envisioned as a community-driven experience, the CV team applied experience design practices when conceptualising this database.[6] Like in many design problems, the project began with the identification of a need. This need was for large quantities of publicly available voice data that could be used to train speech-to-text engines. In the design process that followed, the team ideated on creating an open-source voice dataset over the course of several design thinking exercises with Mozilla community members.[7] This resulted in paper prototypes of varying design concepts. CV then gathered in-person feedback on these prototypes to identify which design concepts to proceed on. The initial assumption of the project team was that people would need an ulterior motive to provide

[5] Interview, Common Voice, online, Bangalore, 22 October 2020

[6] Branson, M., "We're Intentionally Designing Open Experiences, Here's Why," *Medium*, 10 September 2018, accessed 3 November 2021, https://medium.com/mozilla-open-innovation/were-intentionally-designing-open-experiences-here-s-why-c6ae9730de54, 3 November 2021.

[7] Branson, "We're Intentionally Designing Open Experiences."

voice data towards this project. However, the team's insight from the research was that most people were open to the idea of voice donation. They also inferred that people wanted to learn more about the need for such voice data collection. Hence, they designed a platform whose prominent feature was collecting voice data.[8]

They developed an interactive model where people could 'teach' a robot to understand human speech by reading sentences to it.[9] This robot has become part of the CV website as a mascot of sorts, even though the interactive teaching model is no longer operational. The alpha version of the CV platform was built "to tell the story of voice data and how it relates to the need for diversity and inclusivity in speech technology".[10] The CV team collected community feedback through tools such as Discourse[11]and Github.[12] They developed further iterations after feedback collection and discourse analysis. The Open Innovation team at Mozilla shared with us that they emphasise prototyping and reiterating. They carried out a user experience (UX) audit of the working prototype and considered community feedback from Github and Discourse. Based on this assessment, they made refinements to the platform.

Following the release of the working version, the CV team conducted another UX audit. They took into account a combination of UX heuristics, competitor evaluation (such as of platforms such as Headspace[13]), and community feedback. They looked at community feedback on Github and Discourse and spoke to the engineers who built CV. Since 2017, the focus has been on improving the platform and primarily enhancing the experience of contributing voice data. Presently, the team is looking at the bigger picture by focusing on fine-tuning the contributors' experience based on the data and research accumulated.

3. Enabling multi-language contributions

Following an iterative design process allowed CV to ask questions, derive insights, and improve its platform. The team observed that the data collected needed to be more diverse in terms of gender, accent, dialect, and language. They held an experience workshop to ideate on how to support multiple languages and enable better-quality voice data contributions.[14] They realised that the platform needed to provide people with a way to contribute in their desired language(s). They also added dedicated language pages and community dashboards. The team also made further enhancements, such as a new profile login experience

[8] Branson, "We're Intentionally Designing Open Experiences."

[9] Branson, "We're Intentionally Designing Open Experiences."

[10] Branson, "We're Intentionally Designing Open Experiences."

[11] "Civilized Discussion," *Discourse*, accessed November 1, 2021, https://www.discourse.org/.

[12] "Where the World Builds Software," *GitHub*, accessed November 1, 2021, https://github.com/, 3 November 2021.

[13] "Meditation and Sleep Made Simple," *Headspace*, (n.d.), accessed 3 November 2021, https://www.headspace.com/.

[14] Branson, M., "Prototyping with Intention – Mozilla Open Innovation," *Medium*, 8 May 2020, accessed 3 November 2021, https://medium.com/mozilla-open-innovation/prototyping-with-intention-33d15fb147c2

and a new contribution experience, to increase the quality and quantity of voice contributions.[15]

Over the course of our interviews, we learned that CV had been designed to be a global project from the beginning. During the initial stages of development, the team ran a design sprint with a paper prototype on the streets of Taipei. It soon became clear that the platform could not be limited to English. They collected feedback from people who did not speak English as a first language, but wanted to contribute to the platform. It was evident from the feedback that CV did not need to design for specific languages, but for people to opt-in and contribute in a language of their choice. The CV interface is basic, but it features a simple mechanism to choose and add a language. Through this research, the team also discovered that there is an audience for language preservation, who wanted to add languages to CV. The team is currently looking at evolving CV for not just major languages but also for lesser-known or less visible languages.

4. Accessibility and access

The team analysed the CV website on Lighthouse,[16] an open-source, automated tool that audits web pages for performance, accessibility, and search engine optimisation (SEO). Their Lighthouse score indicated that they did not perform well in the area of colour contrast. Subsequently, they are working on ensuring that the website matches all accessibility standards. The CV team emphasised on the importance of having a high quality and accessible dataset. The files for English voice data are heavy and difficult to download, so they are working towards improving access. They are also working on creating a web app version of the website for use on devices with limited bandwidth so that contributors are able to utilise it online and offline.

5. Privacy and data collection

“We don’t believe in taking information that we have not specifically been given regardless of what products are available to us.” [17]

With a large number of people providing voice data, there is a need to protect privacy, especially as voices and accents are easily identifiable. As they understand the vulnerability of voice data, the CV team works closely with their trust and legal team to ensure the privacy of their contributors. They also work closely with the technical, legal, and privacy teams to ensure that the websites – and any new additions – comply with their privacy policies. Mozilla also has a data steward programme, which is run by a group of experts in the organisation who have volunteered to be consultants on data collection and best practices in data management and protection. The CV platform itself operates on two primary principles. The first is de-identification to the highest degree possible.

[15] Branson, “Prototyping with Intention.”

[16] “Lighthouse | Tools for Web Developers,” *Google Developers,* 2020, accessed 3 November 2021, https://developers.google.com/web/tools/lighthouse

[17] Interview, Common Voice, online, Bangalore, 25 March 2020

This requires that for any language being recorded, there should be recordings by at least five people so that it becomes harder to identify them. CV also tries to remove identifiers such as sex and age in smaller datasets. The second principle is based on consent – CV does not associate voice with any client-facing data except when they consent to it. The dashboard helps contributors control who can see their profile; they can hide their visibility to others on CV. The team has created the website to be as malleable as possible when it comes to contributors' interactions with it. Contributors do not necessarily need to have a profile to contribute voice data. CV's terms and conditions agreement states that they collect data for research and that they collect personal (voice) information only when people contribute their voices.

6. Design decisions

Currently, the CV website is not voice-activated but based on 'classic' touch-/point-and-click interactions. The CV team feels that it is important to enable some sort of voice detection in the website, as this will allow for the recordings to be more succinct and accurate. The team has also been thinking about the future of the website: what would it look like when people want to donate their own voices on CV? How can CV use the data they collect to tune voice recognition on the platform itself? If they enable this, they would have to rethink the entire user experience, including navigation, actions, and initiators for contributions.

The overall objective of CV's interface design is to simplify the process by which people can contribute. It is meant to have an intuitive design. However, the experience of contributing may not be the same for everyone, and so this objective is difficult to achieve. The team observed that soon there will be a homogenisation of voice interfaces, as has been the case with websites. They note that this is already underway with wake words and voice assistants. An important question to ask here is if CV data and open-source data can make this homogenisation look different. Can they allow people to tinker and play outside of bigger entities and challenge the idea of what voice interfaces should look like?

The design team notes that it is tough to design for responsiveness. Their challenge has been to fit large quantities of information into a small device/screen, and this is exacerbated by the localisation of CV in various languages. It is difficult to design an interface where one cannot control the way the text appears across browsers. When they cannot read a language, it is difficult to troubleshoot. While this is an ongoing challenge, it is a good problem to have, as it shows that CV is growing. They affirm that taking on community feedback is the most critical and rewarding aspect of this work.

7. Challenges

A key challenge in making CV easier for contributors and the community to access is the need for internet connectivity. In addition to this, material for recording comes from sources such as parliamentary transcripts and Wikipedia, which might not reflect the actual reading and speaking styles that people use in their day-to-day lives. As these sources use more formal writing styles, the training model is also skewed towards a formal mechanism as opposed to the casual way people converse in real life. At times, women and others from underrepresented communities find it less than welcoming to engage with projects in the open-source community – including that of CV – because it mostly consists of men. This means that the dataset comprises mostly male voices, and members from diverse gender identities and communities are not adequately represented in the datasets.

8. Future of Common Voice

“We are only seeing an increased interest in Common Voice.”[18]

CV saw a 20% growth in recorded hours during October–December 2020. Additionally, there has been a significant increase in the interest in CV, both from industries as well as communities. In recent years there has been an increase in community-driven contributions, especially from people involved in language preservation and civic duty systems. These individual and community-based initiatives help add more languages into the CV system, which might not have been possible with a centralised system. More recently, CV received two investments worth $1.5 million from Nvidia and $3.4 million from other investors to continue their work with native African languages.

Disclaimer: This is an independent case study conducted as a part of the Making Voices Heard Project, supported by the Mozilla Corporation. The researchers have not received any external remuneration as a part of this case study, and claim no conflict of interest.

[18] Interview, Common Voice, online, Bangalore, 22 October 2020.