Common Voice for Pakistan: Developing an Open Speech Corpus for Low-Resource Pakistani Languages

Meesum Alam Francis M. Tyers
Indiana University Bloomington, USA
meelam@iu.edu, ftyers@iu.edu

Abstract

Pakistan is home to more than 70 languages, of which around 30 are endangered. Most Pakistani languages remain absent from modern speech and text technologies, with resources concentrated on Urdu and a few major languages. Through Mozilla's Open Multilingual Speech Fund, this paper documents a one-year project to develop an open, community-driven speech corpus for 39 indigenous languages of Pakistan. The dataset includes locally authored texts, everyday language, poetry, and folk songs, creating a culturally grounded resource. The project not only supports automatic speech recognition, but also promotes linguistic preservation and digital inclusion.

Keywords: Common Voice, Pakistan, Endangered Languages, Speech Corpus, Low-resource ASR

1. Introduction

Pakistan is a linguistically diverse country, with roughly sixty-eight local, five regional, and one national language recorded by Ethnologue (Nesbeitt, 1999), along with additional undocumented varieties and dialects across diverse regions. More than 25 languages in Pakistan are endangered due to migration, urbanization, and other factors, including the dominance of Urdu and English in education and no policy for indigenous languages (Rahman, 2006). These low resources languages also lack institutional support needed, and there exists no corpus or dataset representing the communities in the digital age. Such corpora are important for survival and revitalization.

Open source corpus projects such as Mozilla's *Common Voice*, provide a inclusive route for low-resource languages to enter the digital ecosystem. This allows native speakers to contribute under open licenses, democratizes dataset creation and fosters inclusion of low resource languages in speech technology research. This project extends the model to Pakistan's multilingual context and adds on 39 underrepresented languages to the digital world.

The following sections explain the related work on the development of datasets for low-resourced languages, insight into the corpus creation for this project, challenges faced during the one year span of the project, some immediate impacts and conclusion.

2. Mozilla Common Voice

Developing speech technology for the world's diverse languages is important and requires large, open, and high-quality corpora. Yet most languages still lack the necessary resources for automatic speech recognition (ASR), speech-to-text, or translation. The Mozilla *Common Voice* (CV) project, launched in 2017, addresses this global gap through community-driven data collection process. Speakers record short sentences and validate others' recordings, making it one of the world's largest open multilingual collection of speech corpora with more than 200 languages and tens of thousands of contributors (Ardila et al., 2020b).

Common Voice was designed to democratize access to speech data through open access, participation, and transparent licensing. This was a response to proprietary datasets, limited to major languages, CV distributes data under Creative Commons (CC0 or CC BY-NC-SA) licenses, ensuring reuse for research while maintaining privacy and ethical standards. On this platform, community members record sentences, and samples require at least two positive votes before inclusion in the validated set, which is divided into training, development, and test splits (Ardila et al., 2020b). This decentralized crowdsourcing model enables equal participation across high- and low-resource languages of the world.

Many regional initiatives show the transformative potential of the Common Voice approach. The *Catalan Common Voice* effort, within the AINA Project (2022–2024), collected over 2,000 validated hours through institutional support, gender and accent diversity (Guasch et al., 2024). In *Latvia*, the *BalsuTalka.lv* campaign expanded the Latvian corpus from 18 to 270 hours in one year through strong community engagement, university collaboration, and diaspora participation, they also incorporated the endangered Latgalian variant (Bicevskis et al., 2024). Similarly, the **Kuvost** project built a Kurdish–English speech translation corpus of over 1,000 hours from CV data, combining volunteer transcription, expert review, and multilingual adaptation (Ali et al., 2025). These projects highlight the scalability and cultural adaptability of community

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3355–3359
11–16 May 2026.

led corpus creation.

Inspired by above mentioned projects, we approached to **Common Voice Pakistan** with the aim to extend this model to 39 local languages, many of which are endangered or digitally excluded. With Urdu and English dominating formal domains, languages such as Saraiki, Balochi, Khowar, Kalasha, Hindko, and Torwali remain underrepresented. Adopting the CV framework offers an ethical and sustainable way to collect inclusive speech data. The approach prioritizes: (1) *data diversity* by capturing regional phonology and accent variation; (2) *community engagement* through mobilizing universities, cultural organizations, and activists through localized campaigns; and (3) *open access and sustainability* with the leveraging CC0 licensing to ensure long term reuse of the data.

By following these globally tested strategies, the Pakistani Common Voice project helps transform marginalized languages of the country, their inclusion into digital spheres, and community owned speech technologies for all.

3. Corpus Creation

In the recent update of the Common Voice project, there have been added two ways to collect speech data for a particular language; scripted speech where a contributor records already uploaded sentences in the database and spontaneous speech where a contributor sees a prompt and records the speech, later transcribes, edits and validates it (Mozilla Common Voice Project, 2025). We followed the first approach for the 38 languages and only used both for Ushojo language because the language did not have much textual resources.

The corpus was collected collaboratively with native speakers, authors, and community organizations across the languages. Instead of a top down design, we collected materials from the local authors of each language community. Since the project includes languages with number of speakers between 500 to a million, it was preferable to take an adaptable approach to collect textual corpus for each language. In the first phase, the writers, publishers, field researchers and community activists were approached to collect the already established textual corpus of the language. This approach worked really well for medium resourced languages and we were able to collect novels, poetry, short stories and weekly magazine to be added in CV dataset. After the collection of these texts, the data was preprocessed; converted pdf to txt, we used a python script to shorten sentences into a length of 15 words that is acceptable by CV, and reviewed by the community members to be uploaded to CV. For the languages with no or minimal textual work, we asked to the community members to write sentences around their daily life, anecdotes and stories about their culture and communities. These sentences created by language consultants were uploaded to the CV platform for recording purposes. We ensured that the language consultants are using the scripts recommended and supported by the communities, for example, Latin script was used for Wakhi and Brushaski as the community and younger generation understand it better as compared to the Perso-Arabic script. We used keyboards developed by the communities or by the local organization such as Forum for Language Initiative and Institute of Applied Linguistics and with community collaboration.

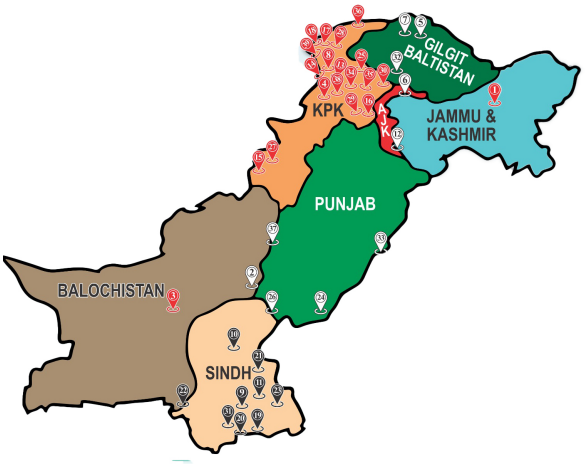

Figure 1: Map of Pakistan showing 39 language locations collected under the Common Voice Pakistan project.

In the spontaneous speech; the prompts were translated into the local language and uploaded to CV database for the recording. The language consultants recorded those prompts, transcribed, edited, and later two other language consultants reviewed these transcriptions for the validation purposes.

Throughout this corpus creation process, we ensured that language consultants and volunteers reviewed every text to ensure grammaticality, naturalness, and the inclusion of as many semantic domains as possible. This method produced sentences that sound authentic to community life rather than machine generated examples. The resulting corpus blends spoken and literary registers, creating a balanced linguistic sample that genuinely represents communities.

List of Languages (ISO 639-3 codes):
1. Balti (*bft*), 2. Eastern Balochi (*bgp*), 3. Brahui (*brh*), 4. Kateviri (*bsh*), 5. Brushaski (*bsk*), 6. Bateri (*btv*), 7. Dawoodi/Domaaki (*dmk*), 8. Dameli (*dml*), 9. Gurgula (*ggg*), 10. Goaria (*gig*), 11. Kachi Koli (*gjk*), 12. Gojri (*gju*), 13. Gawri (*gwc*), 14. Gawar-Bati (*gwt*), 15. Hazargi

(*haz*), 16. Northern Hindko (*hno*), 17. Khowar (*khw*), 18. Kalasha (*kls*), 19. Koli Parkari (*kvx*), 20. Koli Wadiyari (*kxp*), 21. Loarki (*lrk*), 22. Lassi (*lss*), 23. Dhatki (*mki*), 24. Marwari (*mve*), 25. Indus Kohistani (*mvy*), 26. Oadki (*odk*), 27. Ormuri (*oru*), 28. Palula (*phl*), 29. Pahari (*phr*), 30. Kohistani Shina (*plk*), 31. Sindhi Bhil (*sbn*), 32. Shina (*scl*), 33. Sansi (*ssi*), 34. Torwali (*trw*), 35. Ushojo (*ush*), 36. Wakhi (*wbl*), 37. Khetrani (*xhe*), 38. Kalkoti (*xka*), 39. Yadgha (*ydg*).

4. Dataset Overview

The Common Voice Pakistan dataset represents first of it's type as being the most comprehensive open speech resource ever compiled for the country. The comprises of the languages from the south of the country to the very north region including the Kashmir as shown in Figure 1. The dataset covers 39 languages spanning multiple linguistic families, including Indo-Aryan, Iranian, Dardic, Turkic, and isolate. In total, the dataset contains approximately 530 recorded hours and 493 validated hours of speech, drawn from over 139,000 sentences contributed by 1,058 speakers, the detailed stats for each language can be seen in Table 1. The average clip length ranges from three to five seconds. Each language dataset includes gender, age, and metadata to support balanced acoustic modeling and socio-phonetic analysis. The metadata includes language name, family, number of speakers, script, domains of the text, contributors information and sample text. All recordings and metadata are released under the Creative Commons CC0 license, ensuring access and reuse for research and technology development of these languages.

The Common Voice Pakistan data produced through this project is publicly accessible through the Mozilla Data Collective at https://datacollective.mozillafoundation.org/.

5. Challenges and Discussion

This one year project produced over 530 hours of recorded speech and nearly 500 validated hours, 139,000 sentences and representing more than a thousand unique speakers nationwide. Languages with larger speaker populations such as Balti, Khowar, and Hindko showed higher participation, while critically endangered languages like Dawoodi/Domaaki and Ormuri benefited from targeted recruitment to ensure inclusion. In the targeted recruitment, we arranged short sessions with the community members and awarded them with small gifts upon completion of 100 sentences recorded.The project gained print and social media support and soon spread across the platforms such as community gathering and local universities. In addition to contributing the dataset, the project spread an awareness about the indigenous communities in the country. It also fostered local technical literacy and encouraged continued volunteer engagement beyond the funded phase.

Nevertheless, the whole process of developing the Common Voice Pakistan dataset presented a range of linguistic, technical, and organizational challenges. Some languages lacked standardized orthographies or established writing traditions, the Khetrani community used Saraiki writing system, Brushu Marka an academy of Brushaski language used Latin script for Brushaski, Wakhi also used Latin script. We supported community decision on the use of script and encourage the standard and practicing ones.

The other concern was achieving demographic balance across gender, age, and dialect groups, particularly in remote regions with limited connectivity or recording infrastructure. The open and crowd sourced nature of Common Voice also meant that recordings were captured on heterogeneous devices, resulting in variable audio quality and background noise that required continuous post processing and human validation. Coordination across 75 language consultants and volunteer teams demanded persistent communication and careful data management.

Despite these constraints, the community driven approach proved highly effective, enabling sustainable data creation and strengthening communities ownership. The experience supports the idea of how an open infrastructure, participatory design, and community trust can jointly advance linguistic preservation and equal access to data.

6. Conclusion and Future Work

The Common Voice Pakistan initiative demonstrates how open, community-centered collaboration can create meaningful digital resources for underrepresented languages. Through the engagement of native speakers, writers, and local organizations, the project built a foundation for inclusive technologies that reflect Pakistan's linguistic and cultural diversity. The resulting data covering 39 languages and over 500 hours of validated speech shows that data creation is possible even under limited infrastructure, provided that communities are given ownership and support. Beyond its technical outcomes, this project spread awareness of these smaller communities within the country and beyond, people started taking interest in these languages, we shared the list of language consultants around the country for research, and above all we learned that the ethical imperative of linguistic equity in artificial intelligence ensuring that smaller languages can participate in the global digital ecosystem.

We are continuously contributing to the common

Table 1: The table presents the recorded and validated hours, language vitality, sentences, and speakers engaged in the project.

ISO	Language	Vitality (UNESCO)	Rec. (h)	Val. (h)	Sentences	Speakers
bgp	Eastern Balochi	Vulnerable	14	13	6,997	25
hno	Northern Hindko	Vulnerable	11	11	2,349	36
xhe	Khetrani	Sev. Endangered	11	11	5,040	11
trw	Torwali	Def. Endangered	19	19	7,770	27
bsk	Brushaski	Vulnerable	12	11	2,447	27
bft	Balti	Vulnerable	19	18	7,968	156
kls	Kalasha	Sev. Endangered	11	11	3,912	23
khw	Khowar	Vulnerable	21	18	7,046	49
dml	Dameli	Def. Endangered	11	11	5,670	5
oru	Ormuri	Crit. Endangered	18	17	7,355	12
phl	Palula	Def. Endangered	30	22	4,745	20
scl	Shina	Vulnerable	11	11	3,300	39
dmk	Dawoodi / Domaaki	Crit. Endangered	11	11	4,139	10
ush	Ushojo	Sev. Endangered	6.6	6.6	1,161	32
lss	Lassi	Vulnerable	11	10	2,029	15
plk	Kohistani Shina	Def. Endangered	17	13	4,657	10
wbl	Wakhi	Vulnerable	16	13	5,493	13
gwt	Gawar-Bati	Def. Endangered	13	13	3,719	5
xka	Kalkoti	Def. Endangered	11	11	1,980	9
bsh	Kateviri (Bashgali)	Sev. Endangered	11	11	2,646	14
gju	Gojri	Vulnerable	11	11	3,852	6
mvy	Indus Kohistani	Def. Endangered	26	23	6,634	56
gwc	Gawri (Kalam Kohistani)	Def. Endangered	16	12	5,574	22
mki	Dhatki	Vulnerable	11	11	2,055	12
gig	Goaria	Sev. Endangered	11	11	2,005	20
ggg	Gurgula	Sev. Endangered	13	13	2,005	24
gjk	Kachi Koli	Vulnerable	12	11	2,004	22
mve	Marwari	Vulnerable	11	11	2,003	20
odk	Oadki	Sev. Endangered	12	12	2,047	20
kvx	Koli Parkari	Vulnerable	12	12	2,025	22
kxp	Koli Wadiyari	Vulnerable	12	12	2,079	22
phr	Pahari	Vulnerable	15	15	2,077	63
ydg	Yadgha	Sev. Endangered	12	11	1,882	15
sbn	Sindhi Bhil	Sev. Endangered	11	11	2,001	21
lrk	Loarki	Sev. Endangered	12	12	2,006	20
btv	Bateri	Def. Endangered	11	11	1,053	16
ssi	Sansi	Sev. Endangered	11	11	2,007	21
brh	Brahui	Vulnerable	11	11	3,095	18
haz	Hazargi	Vulnerable	11	11	1,361	7
Total (39)	–	–	**532.6**	**493.6**	**139,000+**	**1,058**

voice project by adding more hours and validating the recorded ones, we plan to focus on expanding coverage to additional languages and improving data balance across gender and dialects. We are also working on developing baseline ASR using open-source frameworks. The Common Voice Pakistan aims not only to advance multilingual speech technology but also to preserve the voices and oral traditions that define the country's cultural identity.

7. Acknowledgements

We thank the language communities across Pakistan, especially the speakers, language consultants, and volunteers who contributed to the creation and validation of this multilingual speech corpus. We also acknowledge the Mozilla Foundation and the Common Voice initiative for providing the open infrastructure that enabled this project, as well

as the researchers, collaborators, and community partners who supported data collection and corpus development. This work is dedicated to the indigenous language communities of Pakistan, whose voices and cultural heritage deserve representation in the future of language technology.

8. Bibliographical References

Voxforge: Free speech recognition (linux, windows and mac). https://www.voxforge.org/. Accessed 2025-09-18.

2024. Unesco atlas of the world's languages in danger (interactive platform). https://en.unesco.org/silkroad/languages-and-endangered-languages-along-silk-roads. Accessed 2025-09-18.

Hemn Ali, Peshraw Ismael, Mohammed Attia, Loïc Barrault, Amira Eshky, Hend Al-Khalifa, and Benoît Sagot. 2025. Kuvost: A kurdish-english speech translation corpus based on common voice. In *Proceedings of the 22nd International Workshop on Spoken Language Translation (IWSLT 2025)*, pages 91–102, Bangkok, Thailand. Association for Computational Linguistics.

Anonymous et al. 2024. Wer we stand: Benchmarking urdu asr models. *arXiv preprint arXiv:2409.11252*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020a. Common voice: A massively-multilingual speech corpus. In *Proceedings of LREC 2020*, pages 4218–4222.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Reuben Henretty, Mayra Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020b. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 4218–4222, Marseille, France. European Language Resources Association (ELRA).

Rosana Ardila, Megan Branson, Kelly Davis, et al. 2019. Common voice: A massively-multilingual speech corpus.

Kaspars Bicevskis, Lauma Pretkalnina, Inguna Skadina, Ingus Grinbergs, and Arturs Znotins. 2024. Balsutalka.lv: Mobilizing the latvian community for large-scale speech data collection. In *Proceedings of the 14th International Conference on Language Resources and Evaluation (LREC 2024)*, pages 1620–1627, Torino, Italy. European Language Resources Association (ELRA).

Alexis Conneau, Min Ma, Simran Khanuja, et al. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. Pakistan: Languages, literacy, maps, endangered languages. Ethnologue: Languages of the World. Accessed 2025-09-18.

Mark J. F. Gales, Kate M. Knill, Anton Ragni, and Shakti P. Rath. 2014. Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *SLTU 2014*, pages 16–23.

Miquel Guasch, Salvador Climent, Marta R. Costa-jussà, Miquel Borràs, Jordi Codina, and Jordi Hernàndez. 2024. Catalan voices for everyone: Building a large-scale, inclusive speech corpus through the aina project. In *Proceedings of the 14th International Conference on Language Resources and Evaluation (LREC 2024)*, pages 1652–1660, Torino, Italy. European Language Resources Association (ELRA).

A. Hannun, C. Case, J. Casper, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. In *arXiv preprint arXiv:1412.5567*.

Christopher Moseley, editor. 2010. *Atlas of the World's Languages in Danger*. UNESCO Publishing. See also the interactive Atlas website.

Mozilla Common Voice Project. 2025. Spontaneous speech — prompts. https://commonvoice.mozilla.org/spontaneous-speech/beta/prompts. Accessed: 2025-10-24.

Sarah L Nesbeitt. 1999. Ethnologue: Languages of the world.

Vineel Pratap, Andros Tjandra, Bowen Shi, et al. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

Tariq Rahman. 2006. Language policy, multilingualism and language vitality in pakistan. *Lesser-known languages of South Asia: Status and policies, case studies and applications of information technology*, pages 73–106.

George Roter. 2019. Sharing our common voices: Mozilla releases the largest public-domain transcribed voice dataset. Mozilla Blog.