# Cultivating Spoken Language Technologies for Unwritten Languages

Thomas Reitmaier
Swansea University
Swansea, UK
thomas.reitmaier@swansea.ac.uk

Dani Kalarikalayil Raju
Studio Hasi
Mumbai, India
daniel@studiohasi.com

Ondřej Klejch
University of Edinburgh
Edinburgh, UK
o.klejch@ed.ac.uk

Electra Wallington
University of Edinburgh
Edinburgh, UK
electra.wallington@ed.ac.uk

Nina Markl
University of Essex
Colchester, UK
nina.markl@essex.ac.uk

Jennifer Pearson
Swansea University
Swansea, UK
j.pearson@swansea.ac.uk

Matt Jones
Swansea University
Swansea, UK
matt.jones@swansea.ac.uk

Peter Bell
University of Edinburgh
Edinburgh, UK
peter.bell@ed.ac.uk

Simon Robinson
Swansea University
Swansea, UK
s.n.w.robinson@swansea.ac.uk

## ABSTRACT

We report on community-centered, collaborative research that weaves together HCI, natural language processing, linguistic, and design insights to develop spoken language technologies for unwritten languages. Across three visits to a Banjara farming community in India, we use participatory, technical, and creative methods to engage community members, collect spoken language photo annotations, and develop an information retrieval (IR) system. Drawing on orality theory, we interrogate assumptions and biases of current speech interfaces and create a simple application that leverages our IR system to match fluidly spoken queries with recorded annotations and surface corresponding photos. In-situ evaluations show how our novel approach returns reliable results and inspired the co-creation of media retrieval use-cases that are more appropriate in oral contexts. The very low (< 4h) spoken data requirements makes our approach adaptable to other contexts where languages are unwritten or have no digital language resources available.

## CCS CONCEPTS

• **Computing methodologies** → **Speech recognition**; • **Human-centered computing** → *Participatory design*; *Field studies*; *Interaction techniques.*

## KEYWORDS

Speech/language, zero-resource information retrieval, co-creation, field study

This work is licensed under a Creative Commons Attribution International 4.0 License.

*CHI '24, May 11–16, 2024, Honolulu, HI, USA*
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642026

**ACM Reference Format:**
Thomas Reitmaier, Dani Kalarikalayil Raju, Ondřej Klejch, Electra Wallington, Nina Markl, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2024. Cultivating Spoken Language Technologies for Unwritten Languages. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3613904.3642026

## 1 INTRODUCTION

In this paper we report and reflect on three phases of a project to cultivate speech and language technologies in collaboration with a traditional farming community of Banjaras in Western India, who speak Gormati, a language without a native script.

Linguistic research involving minoritised language communities is often targeted at documenting or preserving a language in the face of worrying statistics that as many as 40% of the 7,000+ languages spoken today are endangered[1] and likely to become extinct by 2050. While such efforts to document and preserve are laudable, they miss out on lines of research that reinvigorate and carry forward a minoritised language through digital media [74] and the possibilities now afforded by Artificial Intelligence (AI) in general and speech and language technologies in particular.

In this contribution we weave together ethnographic, creative, and technical methods in partnership with a Banjara community. Inspired by the oral culture and agrarian lifestyle of the community, we showcase how we cultivated an AI model from seed—that is, without drawing on any existing digital language resources in the target language—to drive a simple, mobile information retrieval interface to surface co-produced media related to their farming practices in response to spoken-language queries.

The development methodology at the heart of this contribution initially requires very little data, and can be iteratively improved, allowing for tighter feedback loops between community data contributions and system improvements. This approach therefore resonates with participatory and action research methodologies that emphasise partnership and reciprocity. In India alone there are 424

[1] https://www.ethnologue.com/insights/how-many-languages/

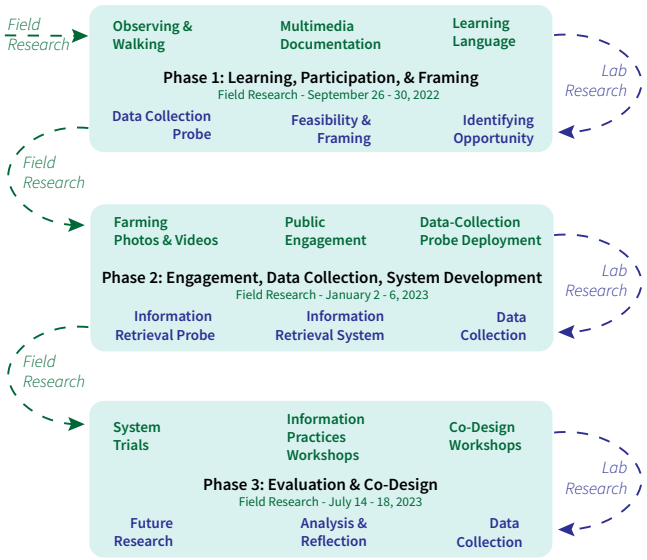


**Figure 1: Schematic outline and timeline of community engagement, design, development, and evaluation activities.**

living indigenous languages of which 304 have no digital language support available and therefore no access to, nor pathways to the development of, spoken language technologies [22]. Our data collection approach and the IR interface we designed and tested in-situ demonstrate the feasibility of our novel approach to speech and language technology development for communities living in oral cultures and speaking languages that are not (commonly) written and have few digital language resources available.

Figure 1 shows a schematic outline of our research and how it is split across three phases, each of which is further subdivided between field and lab research and development. This multi-phased approach afforded us opportunities to build trust and identify, develop, and leverage community assets, which together highlight the value of incremental research approaches [85]. While each phase lists particular methods and probes developed and utilised, the schematic suggests a linear ordering to these that distorts the 'messy' realities in which methods were adapted, assembled, and performed in response to the particular context at hand (see [37]). This also resonates with ICT for Development (ICT4D) research involving people living in oral cultures that emphasises modified, flexible, and opportunistic methods in order to deliver valuable and actionable results [28].

## 2 BACKGROUND

To contextualise our research contributions, we outline related speech-driven systems and efforts to develop digital language resources. We also engage with orality theory to show how the epistemology of writing affects (speech) user interface design. Finally we consider ethnographic and linguistic accounts of Banjara culture and language as well community-centred research on digital repositories and information retrieval that is tailored to the needs and functions of oral, rural, or indigenous communities.

### 2.1 Orality & Written Representations

Ong's seminal work on *Orality and Literacy* [48] interrogates the technology of writing and unpacks the ways in which taken for granted aspects of experience are deeply affected by writing and how these do not generalise to primarily oral—rather than literate—contexts and cultures. The subtleties of Ong's theory are not without critique, especially because of the way it frames signed languages and primarily oral cultures as dependent on or inferior to spoken languages and literate cultures, respectively [8, 15, 26]. However, ICT4D researchers have engaged with the broader arguments of Ong's work to surface a range of pertinent and practical design implications: for instance, by drawing attention to the ways in which oral thought relies on repetition; how information is structured through additive narratives; and, how abstract categories and complex information hierarchies should consequently be avoided when designing user interfaces in such contexts [67]. Goody's argument that written language is not merely speech transcription but a mode of thought reproduction [27] is pertinent to our research too. Consider this paper as an example, which not only contains tabular structures (for instance, the author table at the beginning of this paper, which lists names above affiliations and contact details) and hierarchical ordering (sections, subsections, lists, etc.), but also includes subordinate clauses (such as this one) that exemplify this mode of thinking. According to Goody, knowledge and representation are two sides of the same epistemological coin. This has far-reaching implications, as what we might call the epistemology of the written word also influence the metaphors that structure our experience and thinking (see [35]). Consider how Bidwell juxtaposes the literal and figurative translation of *"are we walking together?"* with *"are we on the same page?"* in reporting her insights of deep design research with rural, oral communities in the Eastern Cape of South Africa [4]. These consequences are not limited to conversation and writing. User interfaces extend and re-produce written thought too [82], just as databases—dominant cultural forms [40] of literate societies across the 'hyperdeveloped' world [72]—are driven by relational, tabular, and hierarchical data structures [20].

### 2.2 Speech Interfaces

Against this backdrop, Speech interfaces in particular have been identified as promising technologies to broaden digital participation of illiterate and semi-literate populations across the Global South [83], especially for those speakers of languages that, like Gormati, do not have a native script. Here Vashistha and Raza [77] give a useful overview of almost 15 years of research, innovation, and impact of Interactive Voice Response (IVR) and Interactive Voice Forum (IVF) applications that—through the widespread adoption of mobile phones—*"have found applications in diverse domains and have profoundly impacted underserved communities in low-resource environments"* [77, p. 570]. Both IVR and IVF systems are accessed by calling a (typically subsidised) phone number, and then navigating a menu of options using speech commands or keypad numbers. An agricultural extension IVR system, for instance, might allow callers to listen to announcements by pressing 1 or saying 'announcements'. Rather than purely disseminating information, IVF applications provide callers with the ability to record messages,

and in the case of Patel et al.'s Avaaj Otalo [50], allow them to post questions and listen to the questions and responses of other farmers and agricultural extension workers. This social element of IVF platforms has driven the success of related systems such as CGNet Swara (for citizen journalism) [45], Polly (for entertainment and jobs) [58], and Sangeet Swara (for social media) [75].

IVR/IVF research has investigated the efficacy of and caller preference for key press vs speech command input modalities (e.g., [50]), but generally do not engage with the nuances of Ong's orality theory and recognise the ways in which IVR/IVF systems impose a hierarchical ordering system. Vashistha and Raza mention that for caller-generated content it is *"difficult to automate categorisation"* [77, p. 596] and how such categorisation is necessary for users and IVF operators/moderators to find and access content.

IVR systems in particular [57], as well as smart-speaker installations [60], have been leveraged to develop digital language resources that many minoritised languages lack. Involving and empowering minoritised language speakers in the crowdsourcing of transcriptions has been a further area of HCI research and user interface innovations [60, 76, 78, 79]. However, these systems have to date targeted regional languages with more established writing systems.

## 2.3 Community & Farming Practices

Our approach of documenting and engaging community members around their farming practices resonates with the Digital Green initiative that leverages peer learning and participatory video to record small-scale farmers across India on effective practices such as vermicomposting or fertiliser application [25]. However, content that is shared through the organisation's agricultural extension platform, while popular and impactful, is stored on a database and accessed through a webpage with a hierarchical navigation system (e.g., by language; by category; by sub category; etc.) that favours larger, regional languages that are more commonly written. Particularly when working in indigenous, marginalised and minoritised contexts, Science and technology studies scholars Verran and Christie alert us to the misplaced dichotomy between traditional (i.e., oral) and modern (i.e., literate) cultures:

> *Traditional cultures are contemporary forms of life just as modern cultures are. They are rich in modes of innovation [...] We can understand traditional cultures as involving nonmodern forms of identity. They have ontologies that make modern assumptions about knowledge and knowing look strange.* [80, p. 73]

They key challenge, then, is to devise appropriate and flexible ways of arranging, storing, and finding digital content that are usable for those working within nonmodern cultures and where this *"becomes a site, a time and place where young and old, with their varying competencies, work together [...] in ways that can empower and educate the young while recognizing older people as knowledge authorities"* [80, p. 74].

## 2.4 Community Repositories & Information Retrieval

HCI researchers have partnered with remote, marginalised, or indigenous communities across the globe to design digital technologies to *"enact culture in the digital age"* [74, p.16] to address specific problems communities face [5] while cultivating sensibilities to cultural and addressing sticky representational issues [44, 70]. Designing with an Aboriginal community, Soro et al. developed a community-notice board with support for both oral and written storytelling, bi-lingual content and different representations of time [70]. Working with Sámi people of the circumpolar north, Moradi et al. designed a web-based digital archive of cultural heritage materials, where a tension emerged around border(less) maps [44]. And in South Africa Bidwell et al. designed two iterations of a community audio repository, to create and share recordings with access to shared tablets, to address *"the difficulty local Xhosa people have in communicating between villages"* [5, p.227]. While the interface to record audio remained unchanged across the iterations, community members found it difficult to find specific recordings, so the revised interface allowed users to tag a recording with photos and record short, annotating abstracts about the recording. The interface to find a recording displays photo tags alongside the recording and autoplays annotations as the user scrolls through their list of recordings. Difficulty in finding voice recordings, in the form of WhatsApp voice messages, was experienced by both Xhosa participants in South Africa and Marathi participants in Maharashtra, India, in Reitmaier et al.'s study; supporting textual search of voice messages was identified as key opportunity for Automatic Speech Recognition (ASR) systems [59].

## 2.5 Ethics, Consent, & Compensation

Those working in the fields of language documentation [7, 21] and ICT4D [16] have highlighted the pertinence of research ethics. The studies presented in this paper were approved by an institutional review board at Swansea University. We also follow best practices within ICT4D research that emphasise long-term engagement and reciprocity [16], and working with local researchers and organisations [16]; and, the practices of linguistic research [7, 21] into unwritten languages, which proposes to establish informed consent orally [21], as well as to place linguists in control of operationalising storage and access to collected data [7]. In following Brereton et al.'s advice and configuring our research approach for reciprocity and engagement, we supported community members when they visited us in the city or when they asked for support or information not directly related to our research [9].

## 2.6 Ethnographic & Linguistic Resources

We also consulted ethnographic accounts of Banjara culture and language elsewhere in India [11, 46] and report on these in the next section. We identified linguistic resources[2] such as word lists [43] and multi-lingual dictionaries [33]. However, we found that community practices, particularly surrounding writing and transliterating, diverged from ethnographic descriptions and so we did not draw on written resources in developing our system. This also means

[2] E.g., http://www.language-archives.org/language/lmn

that our development approach is more likely to be applicable to other contexts where language is spoken, not written, and where knowledge practices are not- or less-influenced by writing systems.

# 3 PHASE 1: LEARNING, PARTICIPATING, & FRAMING

The first phase of this project was not directed towards a particular technological purpose. Rather, it was setup as an opportunity to observe, participate, and learn from an agrarian Banjara community in Jalgoan District, Maharashtra State, India. We report on these activities extensively here to introduce what we learned about the community, its everyday practices, and how we drew on these early experiences to situate and frame the subsequent design methods and approaches of later phases, including identifying spoken language technologies as a nascent design opportunity. These activities and experiences also provoked interdisciplinary discussions across HCI, design, linguistics, and NLP through which we decided to focus on use-cases and data collection around more specific topics (e.g., farming) as opposed to open-ended, unconstrained speech.

## 3.1 Objectives & Methods

The objectives of this phase of research were to lay the groundwork for community partnership, to learn from the community, and to inform and frame subsequent methods and approaches. Previous research involving oral cultures and communities has relied extensively on ethnographic methods for these purposes [6]. The ethnographic methods we utilised are inspired by Lee and Ingold's observation that walking, especially when done alongside others, is a powerful, but often underappreciated form of anthropological engagement: that places are made and best understood through the journeys that people make within and between them; that walking attunes us to multi-sensory and embodied experiences; that walking is fundamental to social life; and that walking together is a particularly sociable type of movement that affords opportunities for shared understanding [38].

We complemented this style of "fieldwork on foot" [4, 38] with audio-visual media (photos, videos, and voice recordings) recorded along the way. Here Pink suggests that mobilising audio-visual media in this manner in general, as well as involving local people in the co-production of videos in particular, are effective methods for uncovering and simultaneously documenting insights [52].

We are also mindful of Brereton et al.'s critique that obtaining the privileged position of ethnographer and observer is difficult, particularly in remote or Indigenous settings and for projects seeking to drive (digital) innovation [9]. In moving "beyond ethnography", processes of engagement, (mutual) learning, and reciprocity should be primary considerations as these underpin valid and sustainable research partnerships [9]. It is in this spirit, rather for the purposes of ethnographic analysis, that we utilise our methods.

## 3.2 Community Background

*3.2.1 Settlement & Infrastructure.* We (one of the authors) were hosted[3] by a family in the community for five days. We slept on the rooftop of a two-room pucca[4] house, owned by one of four brothers. The houses of the brothers (and their families) are clustered around a shared courtyard. All but two of the houses in the cluster are single-roomed, tin-roofed, and constructed of wattle-and-daub. The community has a mains electricity supply, although intermittent power cuts are common. Clean water is only available for 30 minutes every few days, leading to a well-rehearsed choreography of filling every available container. There are no sanitation facilities. On our drive into the community we could see a new telephone mast carrying 4G radio units and antennas, boasting actual 4G speeds (~50Mbit/s) that surpass those of many urban areas by a factor of ten. We observed only a small number of (younger) people with smartphones[5] – most elders either do not own a phone or share a featurephone. The cost of mobile data in India is amongst the lowest globally (~$0.10 per GB)[6]. The smartphone usage we increasingly observed in the community largely revolved around popular culture on YouTube and YouTube shorts.

Walking through the community we learned that the sum of all the surrounding courtyards and hamlets constitute the bounds of the *thanda*—or Banjara settlement of about 6000 people—where many members are related to other members of the thanda [11, p. 45]. Agricultural plots adjoining each hamlet within the thanda are similarly owned by close kin and are generally inherited patrilineally. The thanda is adjacent to, but also distinct from, a Marathi[7] village, a typical settlement pattern. According to our hosts, Banjaras mostly keep to themselves, although there is some trading between the thanda and the village, and Banjara children attend Marathi-language school.

*3.2.2 History & Language.* Our hosts were able to retrace the history of their thanda back to four generations ago. Historically Banjaras led a nomadic life, but were forced to settle by colonial British rule. With the passage of the Criminal Tribes Act of 1871 Banjaras were branded as criminal as their *"nomadic ways of life [...] was regarded as suspicious and [...] difficult to be controlled"* [11, p. 8]. After independence Banjaras were declared a 'Denotified Tribe' and are currently classified as a 'Vimukta Jati and Nomadic Tribe' to recongise their historical marginalisation and make them eligible for special considerations in the state of Maharashtra.

There are 30 million Banjaras in India, but due to their nomadic history and contemporary settlements scattered throughout the country, Banjara culture resists neat classification. Depending on the region they settled in, they are known by at least 26 names (e.g., Banjara, Banjari, Vanajara, Lamban, Lambadi, etc.). Their language is similarly polyonymous, including (but not limited to) Gormati, Gor, Banjari, Lamni, Lambadi, etc. For consistency we use the term "Banjara" to refer to Banjara people/community and "Gormati" as the language spoke by Banjara community members. This follows the conventions of the particular community we partnered with, but also note that we can only speak for the conventions and practices of that particular community and place.

Gormati belongs to the Indo-Aryan language family, but it has many dialects that can vary even from thanda to thanda within the

[3] In each research phase, we compensated community members for their hospitality and time in the form of gifts (e.g. torches, crockery, pressure cooker, ceiling fans, etc.), a more traditional and culturally-appropriate mechanism of exchange (see [42]).

[4] Made of durable materials, unlike the less-permanent buildings described later.

[5] We observed ever greater numbers of phones during each subsequent visit.

[6] E.g., https://www.jio.com/selfcare/plans/mobility/prepaid-plans-list/

[7] The main ethnoloinguistic group of the State of Maharashtra.

same region. While kith and kin speak Gormati to each other within and across thandas, outside of their community Banjara people speak the local language of their region—in this case Marathi—and usually also Hindi [11, p. 53]. Gormati, however, does not have an indigenous script [11, p. 57], although it can be written (transliterated using the closest corresponding letters) through Telugu, Kannada and Devanagari script, depending on the state in which the thanda is located [10, p. 41].

Our engagements within the community nuance such general findings. While we found evidence that transliteration of Gormati using Devanagari script is possible, only a minority of people are able to do this, and it is not a common practice. Furthermore, there is a linguistic difference between younger and older generations. Younger generations are generally fluent in and can mostly read and write Hindi and Marathi: regional and national languages, respectively. Older generations often cannot read or write these languages and may not be fluent in them either.

Therefore, a major barrier to digital participation is text-input (see [18]). However those who are not fluent or literate in Hindi make do and seek assistance from younger family members (e.g., to contact a community member or obtain information) [61].

## 3.3 Community & Farming Practices

Every morning and at different times throughout the day, we went for walks to the surrounding fields to observe, learn, and participate. By walking with our hosts into the fields, we also followed in the metaphorical footsteps of Gupta, who has been walking with rural communities in India to uncover and share grassroots innovations on topics that range from farming, sustainable conservation, animal husbandry to cooking and recipes [29]. Along the way we stopped to greet and chat with other community members. We asked what they were doing and observed them carry out their work in the field: ploughing, weeding, watering, planting. We captured glimpses of such encounters through photos, and videoed community members demonstrating certain aspects of their work to serve as aide-mémoire and later as a means of communicating these experiences with the wider research team, distilled and presented through slide decks. These activities were not directed towards a particular purpose, but were invaluable later to situate design activities. For instance, we developed an intuition of when people were communicating freely and when the language barrier was creating confusion. We mostly chatted in Hindi, but when we sensed confusion, our hosts would translate. Farming is as intrinsic to Banjara culture as the Gormati language, and our hosts were eager teachers of both. They said that if we stayed a few more weeks we would be able to speak Gormati, as we already were picking up certain greetings, phrases, and names of crops.

Both men and women, old and young, work in the fields. Cotton is the primary cash crop; millet, corn, lentils, chillies, and onions are grown for subsistence. Branches cut from trees planted around field boundaries are harvested as tree-hay and fed to cattle and goats along with stalks from corn or millet. Oxen are used to pull carts and for ploughing. Cow's milk is usually sold, but goat milk is served with chai. We mostly ate chapatis made of millet with lentils, prepared by women over wood fires.

When it got too hot (~33 °C), we returned to the courtyard and rested under a neem tree. The weather, cycles of day and night, as well as the needs of animals and crops—rather than the clock—created the rhythm of quotidian activities. Both in the fields and in the courtyard, Banjara women often chant poems and sing songs. And as more community members learned that we were interested in Banjara culture and the Gormati language, they would approach us to capture video of them singing a song. When it got dark, and during periods of down time, we composed some of our own songs inspired by the sonority of the voices we heard throughout the day.

## 3.4 Findings & Implications

Corroborating our direct experiences with ethnographic [11, 46] and linguistic [10] accounts of Banjara culture and language elsewhere in India, we were struck by how these frequently mention transliteration through regional scripts (e.g., using the widespread Devanagari script in Maharashtra). This contrasts with a key finding of our engagements with the community: that Gormati was spoken, rather than transliterated. In fact, the linguistic landscape of the community contained very little writing.[8]

Through discussions with the wider research team across Design, HCI, Linguistics, and NLP disciplines we decided not to pursue lines of inquiry that utilised or surfaced transcriptions or transliterations, mostly because of limited transliteration practices in the community, but also because we wanted our approach to be adaptable to other contexts. Respecting oral practices, rather than imposing tranliteration or writing systems, was a key implication of this phase of research.

These discussions were scaffolded around slide deck presentations containing images and videos we recorded and co-produced while in the community. These slide decks showed the thandas we visited, farming practices we observed, and expressed and communicated glimpses of what we learned about everyday life and the use of Gormati in the community (see [52]).

Our discussions also focused on what Harper [30] refers to as a *"marriage of purpose"* between users and machines that is sensitive to community needs and context, but is also anchored in an understanding of how technology works and what it might realistically deliver – an equally important consideration given the opportunities and hype that currently surround AI [23].

Given current barriers to digital participation especially for older generations and the increasing adoption of smart phones by younger generations (after the installation of a 4G telephone mast), the key finding of our ethnographic engagements throughout phase 1 was that speech and language technologies could have a role to play in the digital expression and sharing of Banjara insight and culture. Through our engagements we also established trusting relationships with community members. Not only had community members proven to be willing (and patient) Gormati teachers, but upon our departure they also expressed a wish for us to return again and establish a nascent partnership. From the technology side, we had to be realistic about what we could deliver, but also needed data to develop Gormati language technologies. Through discussions across the research team, we agreed on a series of **high-level guiding principles**:

[8] Seed and pesticide packaging being notable exceptions.

To alleviate some of the technical complexities of the project and to reduce the amount of required data, given also that there are so few digital language resources in Gormati, we decided to **focus data collection to a couple of specific domains or topics** (e.g., farming). Rather than collecting open-ended unconstrained speech, that one might use when chatting with a friend, we posited that focusing on specific domains would also involve a smaller subset of potential words (~100), and that **gathering 30 hours of speech data** within these domain constraints, would provide enough word repetition to develop a basic language model. To implement these guiding principles and to tighten feedback loops, we also decided that **members of the broader research team should participate in future community visits**.

# 4 PHASE 2: ENGAGEMENT, DATA COLLECTION, SYSTEM DEVELOPMENT

In phase 2 we returned to the community to deepen our partnership and to focus our engagements around more topic-specific use-cases of speech and language technologies. We also experimented with different data-collection methods while in-situ, and used the insights we gained to train community members on how they can continue to contribute data after we left the community. It is this data that we then utilised to train a novel spoken language information retrieval system for Gormati.

## 4.1 Objectives & Methods

The objectives of this phase of research were to engage community members to contribute spoken-language data that match and drive-uses cases for spoken language technologies that support everyday practices. As nearly three months had passed since our phase 1 visit and since another member of the research team was visiting the community for the first time, we began phase 1 by leaning on more ethnographic and audio-visual documentation methods of phase 1 (see Section 3). We did this to tune our senses and sensibilities from urban and research lab environments to those found in rural, oral, and agrarian communities.

Gradually we transitioned our methods to focus also on spoken-language data collection. We took photos and videos (with consent and permission) of farmers conducting the activities they were doing as we passed by and opportunistically recorded videos of people demonstrating other techniques and practices. For instance, upon hearing bees buzzing in a hedge we took a video of the person accompanying us on harvesting honey. To trial different approaches to collecting speech data, we also asked people to narrate (in Gormati) as we were filming.

A central tenet of our design research is to engage and reciprocate, rather than to solely document and collect (see [9]). That is, we wanted to also teach community members how spoken language technologies are able to 'learn' from repeated exposure to particular utterances and how, with time, such a system could identify and match similar phrases. We did this through workshops through which we also recorded further audio data.

## 4.2 Engagement

Returning to the community for five days, as a team of two researchers (one Hindi speaking; one non-Hindi speaking), we settled into the familiar rhythm of accompanying our hosts into the fields and engaging with people as we went along. We again slept on the rooftop of the main house in the courtyard and benefited from being immersed in context and surrounded by community members. We consulted and clarified with our hosts whenever questions emerged. During downtime we wrote up notes, which we also shared with the wider research team, and continually discussed, refined, and reflected on our plans and methods.

**Figure 2: Engaging community members on how speech recognition systems are trained using the metaphor of a learning child.**

We also ran workshops to engage community members on how speech and language technologies are developed or 'trained' and to experiment with different data-collection methods.

Here, we found that the metaphor of how young children pick-up phrases through repeated exposure useful, and would draw on this metaphor to explain how spoken language technologies can make mistakes that can seem childish. We also set up a voicemail box that community members could call and contribute recordings of the different things they did to care for their plots, animals, plants, and equipment. We thought initially that it could serve as a spoken diary that fulfils the domain-constrained vocabulary requirement, but that it could later be queried by community members, for instance, if they needed to remember when something happened. However, when we presented this idea to workshop participants they did not express interest in keeping, and being able to query, such a diary and, to our initial surprise, reported no difficulty in remembering things. However, on post-hoc reflection our surprise likely says more about how we equated memory with written or calendar records, which again shows the deep level at which writing restructures consciousness (see [48, p. 95]).

In preparing for the workshops we worked with one of our hosts to create a consent recording in Gormati that explained that we would be using the data to create an interactive Gormati-language speech-driven application for them. We also created slide decks of the photos we took of different crops and tools, but discussed how we wanted to avoid common approaches to voice user interface development that seek to detect the presence of a predefined keyword (e.g., 'cotton') or keyphrase (e.g., 'watering cotton') [e.g., 56, 66]. We were wary of reifying photographs into objects and

keywords or phrases. On this topic specifically, Ong reminds us that *"an entirely oral language which has a term for speech in general [...] may have no ready term for a 'word' as an isolated item, a 'bit' of speech"* [48, p. 60]. During the workshops, we instead asked community members to play an adapted form of a Wittgensteinian language game [84] and narrate the doings associated with the things (see [32]) pictured.

So we recorded participants narrating those 'doings' – the steps involved in growing the crop or operating the tool pictured in each slide. Here we found a generational divide, whereby older men and women would feel confident in narrating at length, but younger generations were far briefer. Later on we used these narrated slide decks to showcase how an information retrieval system might work. This was done, for instance, by asking participants to say something similar to what they just recorded for the photo of a ox-drawn wagon and then accessing the slide containing that photo through a keyboard shortcut and playing back their recorded narrative. It was difficult to create 'clean' recording environment as the courtyard represents a nexus of both (noisy) activities and (inter)family relations. People had to come and go, so we worked with people when they had time and showed interest, but also did not keep them longer if they had other things to attend to. Participants in workshops found it difficult to imagine use-cases or domains other than farming. They did however mention finding songs and accessing religious ceremonies, which typically also included singing.

## 4.3 Data Collection

We discussed and reflected on our in-situ activities with the wider team and decided to steer away from sung content because this is an application area that would likely be too complex for the capabilities of a system in such a resource-constrained context. Given that community members had observed us recording community narrations in the fields and workshop participants had practised creating photo narrations, we decided to utilise mobile digital storytelling software[9] as a data-collection probe, replicating the process we had started on our laptops. We trained four younger community members on how to use the data-collection probe (see Fig. 3), following the process we trialled and refined during earlier workshops. We created a slideshow template on each phone, which contained 30 photos of crops, animals, and equipment.

We encouraged younger generations, who were more adept at operating their Android smartphones, to help older people to record narrations, a practice which ICT4D researchers refer to as 'intermediation' [61] and which Verran and Christie identify as a site of inter-generational collaboration [80]. To make the process easier for data-contributors, we also explained that it is helpful to record similar content for the same photos; again using the metaphor of a child learning words and phrases through repeated exposure. We loaned a phone to a young lady, to ensure that female voices are represented, as these are often missing from IVR datasets [77]. The three young men used their own Android devices. We paid data-contributors and erred on the side of generosity to ensure that the amount was appropriate and commensurate, and also covered airtime expenses: 4000 rupee (~$50) per contributor.

[9] https://play.google.com/store/apps/details?id=ac.robinson.mediaphone

**Figure 3: Training community members to use the data-collection probe.**

We showed data-contributors how to export narratives from the digital storytelling software and how to share these with the research team. Making this step explicit made sure that we did not accidentally collect data that was not intended for us (e.g., if they were using the software for other purposes). We shared the consent recording from the earlier workshop with the data-contributors and explained how they needed to obtain consent if they recorded someone we had not already obtained consent from earlier.

We stayed in contact with the data-contributors and host family after leaving the community and collated data in batches. While the digital storytelling app exports made it easier to link recorded audio to specific photos across multiple phones and data-contributors, the order of photos in our presentation template meant that data-contributors were creating more narrations for photos that appeared towards the start of the slide deck and fewer for photos that appeared towards the end. In total we collected 3h43m of spoken-language annotations of 30 photos (see Fig. 4).

## 4.4 Findings & Implications

Through our community engagements and workshops, community members were starting to understand how speech technologies learn to pick up and match Gormati words and phrases through repeated exposure to them in the form of community contributed recordings. However, talking to data-contributors after we left the community, we got the sense that the task of annotating photos with recordings, while clearly specified, was still somewhat abstract. That it was unclear how the recordings they were contributing were related to actual Gormati speech technologies and how such technologies might actually work in context. The difficulty in sustaining their engagement and contributions further evidences this, which lead us to substantially scale back our data collection ambitions from 30h to 3h43m.

We posited that an interactive demonstrator system would help to motivate and inspire community members to take on the labour

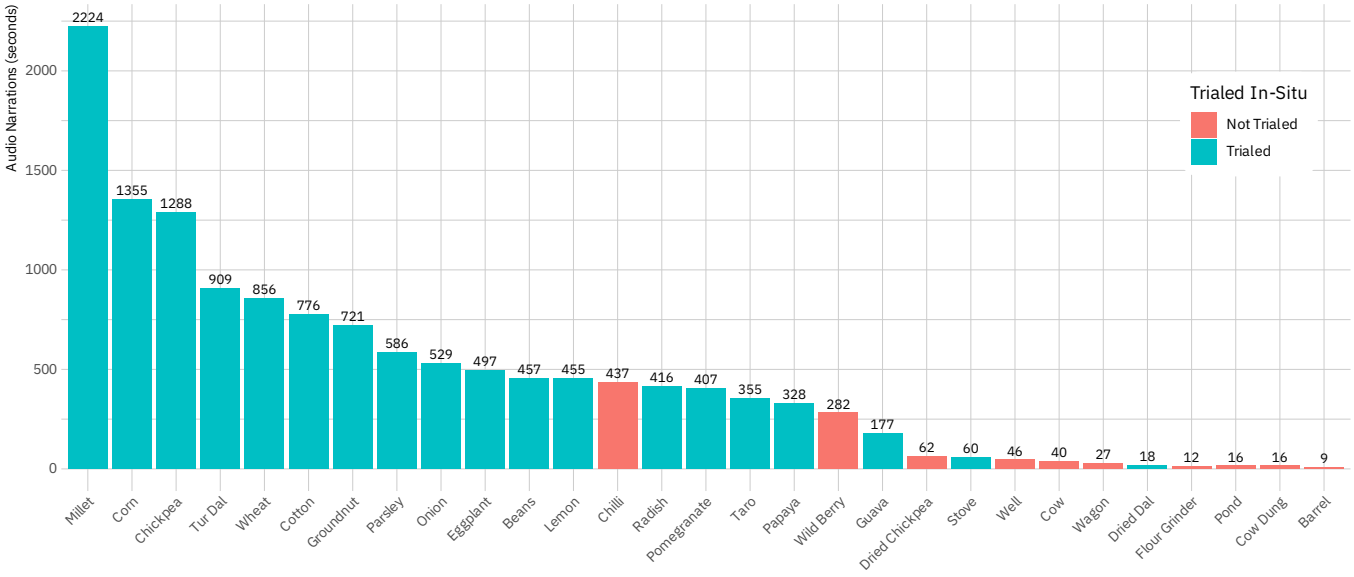


**Figure 4: Distribution of 3h43m of spoken language annotations across 30 photos.**

of contributing data if they could see how these contributions translated into a working system.

## 4.5 Information Retrieval System Development

Meanwhile, in the lab, we (NLP researchers & linguists) reviewed technical literature about spoken content retrieval to find a solution that would allow us to respond to this challenge and build an interactive demonstrator system using only 3h43m of training data.

The most straightforward method for spoken content retrieval is *keyword search*. In this method an automatic speech recognition (ASR) system, trained on manually transcribed data, is used to generate several alternative transcripts of each recording. These transcriptions of all recordings are then used to build a search index, which can subsequently be used to search for a given keyword or keyphrase [2]. The small amount of training data poses a central challenge to building a high-quality ASR system. Applying a multilingual phone recogniser trained on data from well-resourced languages [24, 39], can remedy this problem.

We decided to split our information retrieval system into two components: a phone recogniser and a ranker[10]. We used a *multilingual phone recogniser* trained on transcribed speech data from well-resourced languages to transcribe queries and captions into phone sequences. We then trained the *ranker* to predict how these phone sequences correspond to each photo. Through this decoupled architecture, we were able to rapidly prototype our system and bootstrap the voice user interface with almost zero hours of spoken content in the target language. Furthermore, it allowed us to quickly and iteratively update the system, in anticipation of community members contributing more data during phase 3 of the work.

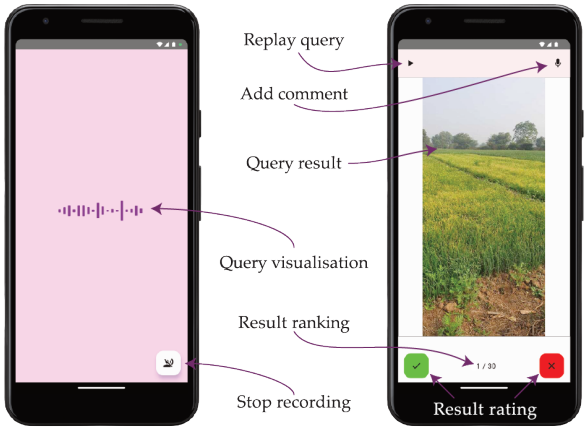


**Figure 5: Media retrieval probe screens for querying (left) and viewing/rating query results (right).**

## 4.6 Probe Development

When we were satisfied with the performance of the IR system in lab conditions, we deployed it as an API so that it could be used and evaluated in the Banjara community. For this purpose, we also created a probe to interact with the IR system through the API, implementing this as an Android app. The interface (see Fig. 5) was kept deliberately simple: after speaking a query the user is shown the top-ranking photo result as generated by our ranker. The user can then rate the result using either the green checkmark or red X-mark buttons, after which the next photo in the ranked list is displayed. After rating all results the user can create a new query.

[10]For technical details on training the multilingual phone recogniser and the ranker see Appendix A and [34].

# 5 PHASE 3: EVALUATION & CO-DESIGN

We returned to the community as a team of four researchers (one Hindi speaking; three non-Hindi speaking) for five days. This time we decided to stay in a nearby town (~30 minute drive), so as not to impose on our hosts, because there were more of us and the monsoon season made it impossible to sleep outdoors.

## 5.1 Objectives & Methods

The objectives of this phase of research were to evaluate the prototype system we developed, which we did quantitatively to avoid participant response bias caused by social and demographic factors [17]. We also wanted to leverage the prototype to engage community members on future use-cases of Gormati speech technologies through a series of design workshops and data collection exercises.

A key challenge here is that the methods that underpin more mainstream forms of user centred or co-design fall short when working with oral communities [28], and in those that are (often also) digitally marginalised [41]. Traditionally, sketches and paper prototypes are used as shared artifacts that facilitate communication between designers and users. However many, especially older, community members cannot read or write and even those that are textually literate tend to treat writing as a more formal, special activity (e.g., letter writing): the opposite of the informal style of writing through which designers involve users in sketching or paper prototyping (see also [6]). Furthermore, digitally marginalised users often do not have a strong sense of digital technology and the types of things computers, or in our case speech and language technologies, can do well. They also have little experience of how digital technology is malleable and can be (re)programmed to look or function differently [41]. Technology probes—simple prototypes, left strategically incomplete and flexible—offer a way of engaging user groups on new (unaccustomed) technological possibilities [31]. Such probes have been successfully used in other, oral contexts as a central component of design workshops to co-design digital storytelling software [6].

Adapting to these constraints and building on successful uses of technology probes in oral contexts, we also leveraged our prototype as a 'dialogical probe' in the design workshops to facilitate a future-oriented design dialogue:

> *As such, the concrete prototype works as [...] a dialogical probe, that supports increasing cross-cultural understanding through dialogue [...] but it will only work because there are people who accompany it and engage in dialogue around it* [69, p. 115].

## 5.2 Evaluation Trial

Before arriving in the community, we ran an evaluation trial with a community member who had recently migrated to the nearest major city (Mumbai). We brought a phone with the probe app installed as well as printouts of the 30 photos listed in Fig. 4. We asked the participant, who was previously a data-contributor, to complete two tasks:

*5.2.1 Task 1: Querying Individual Photos.* For this task, we kept the full deck of photos hidden from the participant. We disclosed the photos one at a time, and after showing a photo, asked them to use the app to record a query that they would expect to bring up the photo.

*5.2.2 Task 2: Querying the Collection.* For the second task, we spread out all of the photos on the floor and pointed to an individual photo. Similar to the first task, the participant was asked to record a query on the app that should bring up that photo as a result. The aim here was to support the query task with contextual knowledge about the full corpus of images.

*5.2.3 Results & Reflections.* Between the two tasks, we found (via the participant's feedback) that knowledge of the corpus did not make a difference to how the participant recorded queries. We also found that photos with only minimal audio annotations were not reliably being returned as results. While the participant represented a best case scenario—being digitally savvy and part of the training dataset—we learned that we would need to focus more on communicating the capabilities of the system before recording user queries, even if this meant giving community members an overview of the corpus of photos. We were, however, cautious about affecting how community members articulated their queries.

Trialling the app ourselves during its development, we used Gormati keywords and keyphrases we picked up from the community, such as 'kapashi' (cotton) and 'bajri' (millet), to test if its bi-directional streaming of audio queries and photo results was working. So, we planned to avoid demonstrating the app ourselves in the community and would instead encourage participants to speak naturally. To accommodate longer queries, we configured the app to only cut off a query after 10 seconds had elapsed. We also adapted the prototype with functions to replay a query and to record audio comments while viewing photo results to allow us to contextualise interactions – for instance to mark queries we might make to test whether the system was working or to indicate why a query failed for external reasons (e.g., loud noises or concurrent speech).

## 5.3 Community Evaluation

On the second day of our visit we recruited eight community members (4M, 4F; aged 20–50), who had not been part of data-contribution during phase 2, to experiment with and evaluate the system. We decided to conduct evaluation sessions inside one of the homes, back-to-back, and all on the same day to minimise community members consulting with one another about the task and how they created queries. We went through ethics and consent with participants as well as introducing the system, what it does, and how it operates. We kept the corpus of photos out of view from participants and then showed an individual photo from the corpus, asking participants, as in our earlier evaluation trial, to say a query that would bring up that photo in the app. We then rated the returned photo results using the rating buttons on the photo results screen (see Fig. 5). After completing the rating step, we showed participants the next photo and asked them to record a query for it.

In some cases we had to retry a photo query: because participants were still thinking about what to say after we had already started recording; because someone had entered the room and was talking

while recording; because the goat tied to the front of the house was bleating to be fed; or, because host family members brought in chai. We noted these interruptions by recording comments on the results screen.

In articulating her alternative account of the relations between plans and situated action in the context of scientific research, Suchman found that the experimenters' expertise lay not in strict adherence to plans and protocols but in being able to continually adapt by drawing on plans as a resource for action [73, p. 185]. During the evaluation sessions, we too found it necessary to adapt. Drawing on the trust and intuitions we formed over the course of three visits, we could sense that early participants were getting tired during evaluation sessions, especially since we had to ask them to repeat queries to compensate for interruptions by people or animals. So, we decided to accommodate participants by excluding photos with less than two minutes of audio moving forward, and hence limiting our evaluation to 19 photos. However, in the moment we mixed up two of the photos and excluded the photos of 'Chillies' (437s) and 'Wild Berries' (282s) by accident and included the photos of 'Stove' (60s) and 'Dried Dal' (18s) instead. The green columns of Fig. 3 show the 19 photos that formed part of the evaluation dataset, while the red columns indicate the long tail of photos we decided not to trial because they had so little data associated with them. A further accommodation we made was to show the photos on a laptop, because older participants had difficulty seeing the printed photos inside the dimly lit homes. Changing this on the fly, we could no-longer rely on simply shuffling the photos in the deck to randomise the order, nor rely on removing cards from the deck to keep track of which ones we had shown to participants. While most photos were shown to between five and eight participants, three photos ('Dried Dal', 'Corn', & 'Stove') were only shown to one or two participants. Despite occasional interruptions and unanticipated accommodations, however, overall we settled into a steady rhythm of querying and rating the returned photo results.

## 5.4 Results

After returning from the community, we iteratively cleaned the data generated during the system evaluations by first removing those queries which contained an audio comment to mark them as excluded (e.g., for testing, needing to be retried, etc.) or was inaudible because the microphone was blocked. Next we marked the remaining 99 queries which contained audible speech for inclusion. Figure 6 shows the distribution of results of the 99 queries that participant-evaluators created on the app in response to being shown one of 19 different photos. The dotted blue line indicates our target of returning the corresponding photo as a top-5 result. Across the 19 photos there were 73 queries that returned the corresponding photo on the app as a top-5 result. The remaining 26 queries did not meet our target threshold.

Subsequently, we recruited a community member to assist with translating a random sample of 40 queries into Hindi using voice recordings. We further transcribed and translated the Hindi audio into English.

Consider our worst results for 'Pomegranate' ($19^{th}$) and 'Onion' ($18^{th}$). In the 'Onion' example, the participant formulated a query surrounding the replanting of field, as the photo showed a freshly cultivated field that had recently been planted out to onions (see Fig. 5). It is likely that the IR model picked up the words surrounding fields and planting, which would overlap with the annotations of the millet photo, which was incorrectly returned as the top result for that particular query. The 'Pomegranate' example also featured a descriptive query about how the fruit on the bush of the photo looked ready for harvest and selling at market. However, the photo of 'Wheat' was returned as the top result. Although these queries produced outlying results, other queries for pomegranate and onions produced top-5 results on five and three occasions, respectively. This style of fluid and descriptive querying is furthermore representative of the larger query dataset, which did not contain keyword or keyphrase queries (e.g., for 'the pomegranate photo').

## 5.5 Community-Based Co-Creation

We leveraged the interactive demonstrations that our probe afforded to engage and involve eleven community members (7M, 4F; 18–68) in the co-creation of media retrieval use-cases that are more appropriate in oral contexts. These occurred during two workshops across two days, one focused on current information seeking practices and the other on uncovering use-cases that support and extend these practices with more useful content than the photos of the current probe. Before, between, and after workshops we experimented with different content generation approaches as well as refining the ways of collecting audio annotation data to drive the IR system with the same participants.

*5.5.1 Evolving the IR System.* After the evaluation sessions, we asked two younger community members, who had been data-contributors in phase 2, to take three additional photos. They sent us photos of a goat, sorghum, and (a different type of) corn. On two phones we created three slideshows, one for each photo, on the digital storytelling data-collection app (see Fig. 3) used during phase 2. Between the two participants, we asked them to create and send us ten spoken annotations for each photo. We used these photos and annotations to retrain the ranker model to showcase how the current IR system can evolve, and we utilised the evolved system during workshops.

*5.5.2 Current Information Practices & Languages.* We split this workshop across two groups to fit more comfortably inside the house: the first was with four younger participants aged 18–22 and the second was with seven older aged participants (aged 30–68). Younger and older generations had different responsibilities in the fields and in the homes, and so were available to participate at different times. The generational divide across groups also expressed itself in terms of smart-/feature-phone usage and non-usage as well as their fluency and literacy in Hindi and Marathi. We asked polylingual participants to translate for one participant in the second group who did not speak Hindi. We structured what ended up being lively discussions around five scenarios/topics, designed to cover a broad range of everyday experiences. To arrive at these, we utilised interpretative research strategies [68] and drew on our observations and lived experiences of previous research phases – documented through field notes and research diaries – to come up with 18 potential scenarios/topics, captured these on post-it notes,

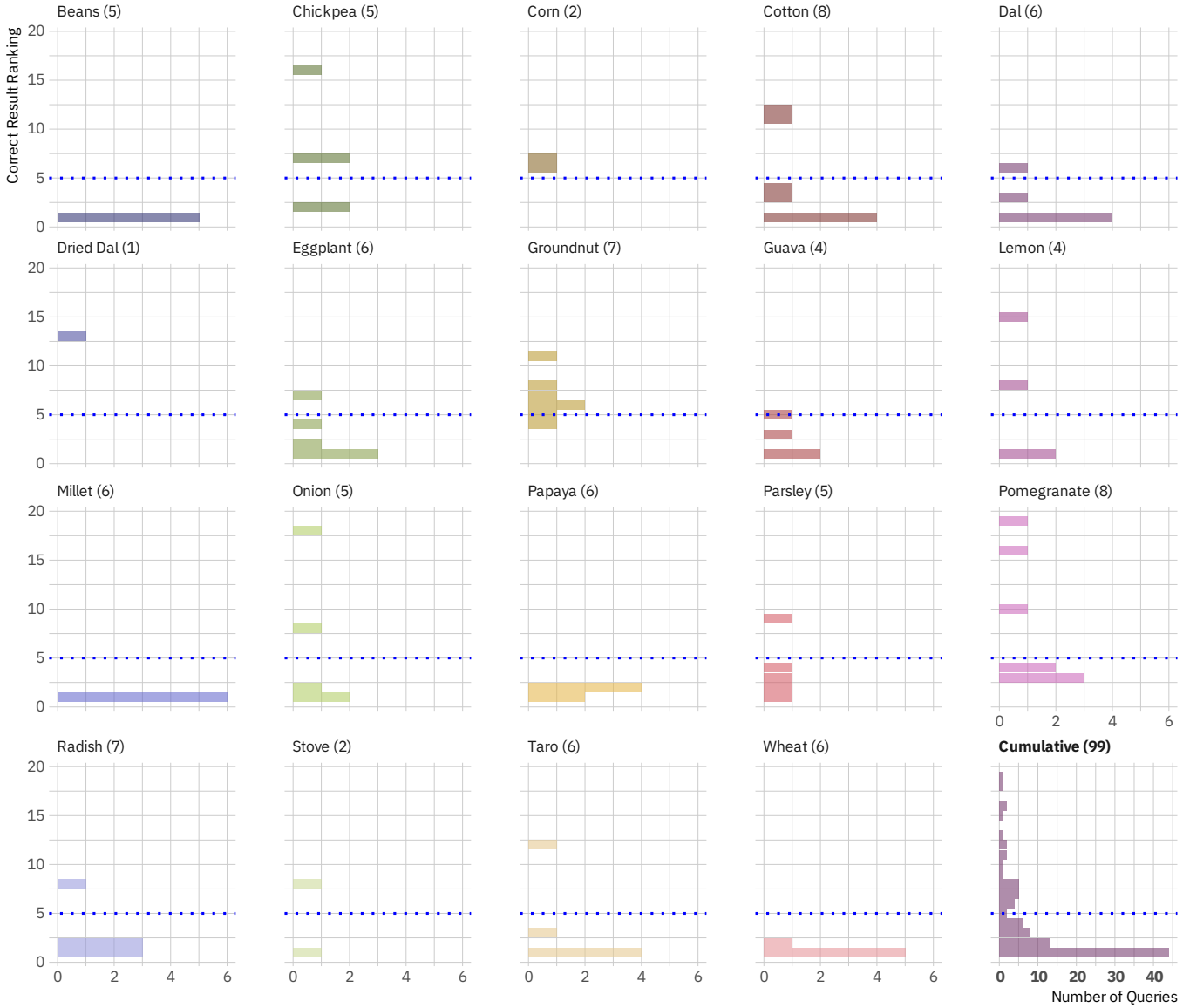


**Figure 6: Distribution of result rankings for 99 queries across 19 photos with top-5 results shown below the dotted blue line. For example the bar plot on the bottom left is captioned 'Radish (7)' to indicate that the photo was of a radish plant and was shown to seven participants who each formulated their own query for that photo. This particular plot shows that across the seven queries the photo of the radish plant was returned as the 8$^{th}$ ranked result once, the 2$^{nd}$ ranked result three times and as the top result a further three times. The 2$^{nd}$ and 1$^{st}$ ranked results are shown below the dotted blue line to indicate that they met our top-5 result target. Finally, a cumulative distribution combining results from all 19 photos is shown in the bottom right corner and highlighted in bold (though note that the x-axis is scaled 0–40 instead of 0–6 here).**

and following discussion between research team members distilled these down to five.

In the *selling cotton* scenario we asked participants to walk us through their line of reasoning for deciding when to sell their cotton harvest. In phase 2 we had observed that one family stored many bags of cotton in the back room of their house, hoping for higher prices later in the year. Some participants were not involved in this process and deferred to and trusted other family members in their decisions. Those involved in the process said that the internet was not a helpful resource: internet prices were characterised as 'fake' – more than what is actually offered from buyers in the area. Instead they call buyers and middlemen in the area, if possible pooling together, so the buyer will come collect the cotton harvest from many families in a single vehicle. But this arrangement often falls through on either side, and then the buyer will not come to collect the crop. Other families drop the cotton off themselves at depot in

the nearby town to command a higher price, but need to cover the cost of transport.

On the topic of *seeking health information*, community members mentioned that they visit physicians in a nearby town and generally do not consult online information. They also reported on their experiences of Covid-19 and how healthcare workers came to administer vaccinations. They received a digital copy of an English vaccination certificate and those with smart-phones would facilitate receiving certificates for those without access.

On the topic of seeking help and information to manage *crop diseases and pests* participants consulted agricultural shops in the nearby town and preferred to go in-person rather than call. Sometimes they will show a photo they have taken of the issue, and usually the people working in the shop will make appropriate recommendations.

Across these three scenarios and especially when information was sought outside the Banjara thanda, participants reported using Marathi, which led us to the topic of *language preferences and perceptions*. Participants preferred, and indeed cared deeply about, their language, but were also pragmatic about speaking different languages, and saw it as necessary to live in a society with lots of different people. However, within the community they always speak Gormati with each other. Elders mentioned that Gormati is a strong and stable language, but also acknowledged that their children learn more and more Marathi in order to speak with outsiders. In their view, they should stick to Gormati. Younger participants prefer Ahirani and Marathi songs, and claimed that songs in those languages are more melodic than Gormati. However, older participants did not share this view.

This led us to the topic of *multimedia*. Here older participants generally relied on those with smartphones to facilitate access. For instance, this was achieved by asking children to play a song from YouTube. Young people demonstrated how they use voice recognition, keyword search, and code-switching on YouTube to search for: *"Gor Banjara Song"*[11]. They explained that you need to use the English alphabet to find content on the internet, and also adjusted their querying style, from fluid Gormati queries that we observed while evaluating the IR system, to using keywords. Top results[12] are of high production value, with well-designed title cards that help identify and differentiate songs.

Participants mentioned that they would like to see more Gormati videos on topics in the following areas—farming, recipes, songs, comedy, and religion—and to record and share their own songs as well as videos with recipes or showing effective farming practices. Younger participants already create video content, but often delete it from their phones to conserve space and choose not to upload it as it does not match the production value of the videos they like to look at online.

*5.5.3 Community-Generated Media Content.* Between the workshops we experimented with generating the type of media content participants mentioned earlier: filming community-members making chapati, cooking lentils, weeding, and ploughing. Participants in the videos narrated what they were doing and, at our encouragement, repeated their demonstrations and narrations. For instance, when demonstrating how to make chapati, the person in the video made multiple chapatis and demonstrated and narrated each step multiple times: dosing and shaping the dough, cooking and flipping the chapati on the stove, and manipulating the cooked chapati to make it more pliable. We also encouraged participants—mothers and older farmers as well as their younger adult children—to make their own videos on project phones. We also found that women in particular wanted to record and share songs. However, unlike phase 2 where we precluded sung content, because it would be technically too difficult for our recogniser to cope with, we encouraged these and later explored ways for community members to contribute spoken, rather than sung, annotation data for the IR system.

*5.5.4 Design workshop.* We met with the same participants the next day to think about use-cases for spoken language technologies, as embodied and exemplified by the current probe, that support and extend their current practices. This time we had arranged to meet with the older group first, so that we could feed back their insights to the more digitally savvy, younger group.

With the *first group*, we started by showing some of the videos they recorded earlier and discussed that they would be of interest both within and outside of the community. We asked them to imagine how they would find those videos if they were living in a different Banjara community hundreds of kilometres away. They mentioned they would ask their children to help and that songs would be of particular interest to them. In the media gallery of one of the two phones we had lent, we tried to locate one of the songs that participants had recorded, but initially could not locate it across similar-looking thumbnails. We also checked on the second phone, until we finally found the video after a more systematic check on the first device. We used this difficulty as an opportunity [28] to show our IVR probe again, demonstrate how it can make it easier to find content from spoken descriptions, and showcase how it had been extended with new photos since the previous day. We also explained that the IVR probe can be changed in future to include videos and song content, but that it can only understand spoken Gormati. We then asked participants to imagine that many songs were on the IVR probe and tell us how they would find a particular song. After some discussion in the group they said that they could either explain the song in words or say the first line of its lyrics.

With the *second group* we began by discussing how the elder group had shown an interest in making videos—of their farming, their cooking, and their songs—and considering whether this would be of any value to them. The group said that if they knew the people in the videos they would look at them, and suggested they might laugh initially, but if the content is useful they could see others looking at them. We asked about another nearby community creating such videos and to imagine what these videos would be of. They expressed interest in seeing how different communities create fertiliser from cow manure, or stock ponds with fish. They also mentioned how their fathers are very skilled at particular aspects of farming, such as inter-cropping and keeping an ox-drawn plough straight; videos of these could be shared within the community and with other communities. They mentioned that videos could also be shared via WhatsApp, which led us to enquire how the messaging app is used in the community. Within their group participants tended to use WhatsApp to forward images and videos and to

[11] Gor is another way of referring to Banjara culture/people.
[12] https://www.youtube.com/results?search_query=Gor+Banjara+Song

send very short messages—e.g., 'hi', 'what's up', etc.—transliterated using an English keyboard. This was the only evidence we saw of transliteration practices.

*5.5.5 Revised Data Collection Methodology.* On our last day in the community we worked with data-collectors from phase 2 as well as community members who had shown interest in creating video content. We loaned phones to an older farmer, and to two sisters-in-law who wanted to perform and share their songs. Other people either had their own phones, or could borrow one from a family member or use one of the phones we had lent. The community-generated video content from earlier in our visits already contained some spoken narrations, but not enough for the ranker model of our IR system, and in the case of sung content, would need to rely entirely on spoken annotation. Trialling this with participants, we learned that young people appreciated being able to listen to the original narrations of the content videos, as they found it harder to record annotations if they lacked the knowledge and/or confidence to describe what was being demonstrated in the video. They found it easier to start by 'repeating' what was already said, but also found ways of integrating their own knowledge and experience of the topic once they started speaking. These recordings were therefore *not* verbatim repetitions used by systems such as 'ReSpeak' [78] to develop transcriptions for written languages. However for our use-case, annotations featuring repetitions with variations are a more useful training resource for the IR system ranker than verbatim ones.

We asked phase 2 data-collectors about their experiences using the digital storytelling software, and they mentioned that it was frustrating to only be able to export the entire digital story slideshow, even if they only wanted to share one new audio annotation. We also wanted to explore a different data collection methodology, given that phase 2 audio annotations were unevenly distributed across photos (see Fig. 4). We suggested that they could also try using WhatsApp voice messaging for this purpose, and set up a group between devices to demonstrate this. We shared a farming video to the group, and participants found it easier to respond to that video with a voice message containing their spoken annotation. This refined method also leverages participants' familiarity with the platform (see [36]). Following this, we settled upon and further demonstrated and agreed on the following (ongoing) data collection process:

- A video (e.g., on farming, cooking, songs, etc.) is shared to the WhatsApp group;
- Participants record and send audio annotations for that video to the same WhatsApp group;
- All audio received is assumed to be related to that video;
- After enough (10–20) annotations have been received, a new video is shared, and the process repeats; and
- Researchers would be included in the WhatsApp group to collect video and audio annotation data and to encourage use.

Finally, we established formal participant consent to being included in the WhatsApp group, to participate, for us to use the video content for a community repository and the spoken annotations to improve the IR system. To date, community members have contributed ten further videos with 48 minutes of audio annotations.

# 6 DISCUSSION & FUTURE RESEARCH

While the installation of new 4G phone masts in the community at the start of our research show how digital divides surrounding internet access are increasingly being addressed, written epistemological assumptions still pervade UI paradigms (e.g., information hierarchies) [82] and are also prevalent in the dominant keyword query paradigm (see [47]) of speech-driven user interfaces and IVR systems. In short, there are still barriers to digital participation that, as our research demonstrates, could be alleviated through speech and language technologies. Here, our research clearly shows how participants in our study—and presumably this extends to similar oral communities—used fluid and descriptive queries rather than keywords and keyphrases. Supporting this distinct interaction style – especially in oral contexts and for unwritten languages – is crucial to unlocking laudable efforts to create content for and with minority language communities, such as the Spoken Web [1] or Digital Green [25].

Current digital inequalities are furthermore splitting open new divides between Global North and South in terms of access to AI technologies. These technologies are often trained on datasets generated by digital platforms [14], which are inaccessible to many minoritised communities in the Global South [81]. In the case of speech and language technologies, this means that the language communities who would stand to benefit the most from this interaction paradigm are simply being cut out of the conversation. Responsible and human-centred [12] forms of AI innovation, as our research shows, have a tremendous role to play in closing this gap and is a critical area for HCI research to contribute [30].

The farming practices we observed while in-situ demonstrate creativity, innovation, and resilience in the face of a changing planet characterised by less predictable and more extreme weather patterns. Not only are these practices never recorded in any datasets generated, for instance, by discussions on online platforms, but we also miss out on engaging with both traditional and contemporary forms of knowledge and practice [80] in the design process of AI. More research, collaboration, engagement, and partnerships are required to bridge these gaps and to ensure more equal representation so that the tremendous opportunities of AI benefit and address the needs of diverse communities across the world, and not just those in the Global North.

Our contribution also speaks to, and is shaped by, NLP research. We have outlined the pipeline and decoupled architecture we used to develop the IR system (see Appendix A) so that more technical researchers might reproduce and build on our results. Technical advances within NLP research, particularly to support so-called 'low-resource' or 'zero-resource' languages, are often organised and structured through competitions associated with major conferences (e.g., [19]) using existing datasets that are far removed from everyday experience and therefore unlikely to benefit those language communities directly. Here our research contributes an adaptable blueprint for NLP researchers to engage with communities from 'day zero'. This blueprint is paired with a development method that supports, and is supported by, these engagements to build interactive systems from scratch – without any existing digital language resources in the target language. The evaluation results of our IR system show that we met our top-5 target metric 74% of the time

and demonstrates the feasibility of this approach. Critically this iterative and incremental engagement and development approach, not only facilitates collaboration across HCI and NLP, but also affords tighter feedback loops for communities that build momentum, motivate and engage data contributions, to ultimately co-create more meaningful systems that are matched with appropriate data.

A key finding of our research is just how important demonstrator systems are to translate the abstract and somewhat ineffable concept of 'speech and language technologies' into something practical and concrete, especially in oral context and in communities with less technological familiarity. In our design workshops, the IR system also functioned as an engagement probe to demonstrate the interactive capabilities of spoken language technologies, how these can be iteratively improved, and critically engage community members to uncover use-cases to suit their community and practices.

We are currently building a communal tablet-based system to store and access videos, that is driven by the IR system. In order to include a video (e.g., of a song, showing a farming technique, etc.) on the tablet, it needs to be supported by 10 audio annotations explaining the content of the video (e.g., the lyrics or meaning of the song or explanation of the technique shown). These ten recordings are a compromise between what community members can deliver before the task becomes too tedious and the needs of the ranker system to operate effectively. The supporting audio annotations are then used to train the ranker, so the videos can then be accessed by community members through spoken queries on the tablet. The IR system combined with a community-operated tablet then function as a novel form of storing and accessing information: where videos are stored and accessed through oral descriptions, rather than tags, categories, hierarchies, or meta-data typically used in database systems (see [20]).

## 7 CONCLUSION

We began our research by accompanying our hosts in an agrarian, Banjara community in Jalgaon District, Maharashtra, India for a walk [38] into the fields. Along the way we learned about Banjara culture, farming, and cooking practices, and picked-up some Gormati phrases too. We also challenged community members who engaged with us to learn about spoken language technologies, to contribute data, and to experiment and feed back on systems, so that together we can cultivate speech and language technology from seed to support their language and oral practices.

Orality theory [48, 67] afforded us a critical lens that brings into focus the written assumptions, epistemology, and representational practices [27] of user interfaces and content repositories more generally and speech interfaces, such as IVR [77], specifically. Developing oral alternatives to these is a substantial technical undertaking and contribution of our work, especially for unwritten languages without digital language resources. Here we had to balance a sensitivity to community context and unfamiliar oral practices while also being anchored in a firm understanding of how spoken language technology works, what it might realistically deliver, and the data that is required for its development [30].

Time spent in the community was essential to mediate between these demands, to experiment with different approaches, to adapt and act opportunistically [28], and to deal with the vicissitudes [71] of such 'data work' [62] in general. These vicissitudes required us to scale back our ambitions as we only had access to very limited amounts of data (< 4h). We therefore tailored the information retrieval system to utilise a multilingual phone recogniser and a ranker that can be trained independently. This decoupled architecture supports the development of interactive information retrieval systems from scratch that can be seeded with as little data as is available. As more media content and annotation data becomes available only the ranker needs to be retrained. Compared to the multilingual phone recogniser, the ranker can be retrained quickly and without much computational resource, supporting both iterative and more sustainable practices.

A trouble that we identify with NLP research is that it falls short on engagement methods especially when developing for minoritised language communities where the very concept of spoken language technologies is abstract and ineffable. Taken together our research contributions create an adaptable blueprint for engaging with communities from 'day zero', paired with a development method that supports, and is supported by, these engagements to build interactive systems from scratch – without any existing digital language resources in the target language.

## ACKNOWLEDGMENTS

We thank community members, study participants, and data contributors for their contribution to this work. All source code data for the systems reported in this paper is available at https://github.com/unmute-tech/. This work was supported by Engineering and Physical Sciences Research Council grant EP/T024976/1.

## REFERENCES

[1] Sheetal K. Agarwal, Anupam Jain, Arun Kumar, Amit A. Nanavati, and Nitendra Rajput. 2010. The Spoken Web: A Web for the Underprivileged. *ACM SIGWEB Newsletter* 2010, Summer (June 2010), 1:1–1:9. https://doi.org/10.1145/1796390.1796391

[2] Cyril Allauzen, Mehryar Mohri, and Murat Saraclar. 2004. General Indexation of Weighted Automata - Application to Spoken Utterance Retrieval. In *Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004*. Association for Computational Linguistics, Boston, Massachusetts, USA, 33–40. https://aclanthology.org/W04-2907

[3] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*. 2278–2282. https://doi.org/10.21437/Interspeech.2022-143

[4] Nicola J. Bidwell. 2012. Walking Together to Design. *interactions* 19, 6 (Nov. 2012), 68–71. https://doi.org/10.1145/2377783.2377797

[5] Nicola J. Bidwell, Thomas Reitmaier, and Kululwa Jampo. 2014. Orality, Gender and Social Audio in Rural Africa. In *Proceedings of the 11th International Conference on the Design of Cooperative Systems*, Chiara Rossitto, Luigina Ciolfi, David Martin, and Bernard Conein (Eds.). Springer, 225–241. https://doi.org/10.1007/978-3-319-06498-7_14

[6] Nicola J. Bidwell, Thomas Reitmaier, Gary Marsden, and Susan Hansen. 2010. Designing with Mobile Digital Storytelling in Rural Africa. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 1593–1602. https://doi.org/10.1145/1753326.1753564

[7] Steven Bird. 2018. Designing Mobile Applications for Endangered Languages. In *The Oxford Handbook of Endangered Languages*, Kenneth L. Rehg and Lyle Campbell (Eds.). Oxford University Press, 0. https://doi.org/10.1093/oxfordhb/9780190610029.013.40

[8] Jan Branson and Don Miller. 1998. Nationalism and the Linguistic Rights of Deaf Communities: Linguistic Imperialism and the Recognition and Development of Sign Languages. *Journal of Sociolinguistics* 2, 1 (feb 1998), 3–34. https://doi.org/10.1111/1467-9481.00028

[9] Margot Brereton, Paul Roe, Ronald Schroeter, and Anita Lee Hong. 2014. Beyond Ethnography: Engagement and Reciprocity as Foundations for Design Research

out Here. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1183–1186. https://doi.org/10.1145/2556288.2557374

[10] Sunder Bukya. 2021. Suffix Negatives of Banjara Language. *Indian Journal of Applied Research* 11, 1 (2021). https://doi.org/10.36106/ijar

[11] J. J. Roy Burman. 2010. *Ethnography of a Denotified Tribe: The Laman Banjara*. Mittal Publications.

[12] Tara Capel and Margot Brereton. 2023. What Is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–23. https://doi.org/10.1145/3544548.3580959

[13] Corinna Cortes and Vladimir Vapnik. 1995. Support vector machine. *Machine learning* 20, 3 (1995), 273–297.

[14] Kate Crawford. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, New Haven.

[15] T. A. Czubek. 2006. Blue Listerine, Parochialism, and ASL Literacy. *Journal of Deaf Studies and Deaf Education* 11, 3 (mar 2006), 373–381. https://doi.org/10.1093/deafed/enj033

[16] Andy Dearden. 2013. See No Evil? Ethics in an Interventionist ICTD. *Information Technologies & International Development* 9, 2 (2013), pp. 1–17.

[17] Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. 2012. "Yours Is Better!": Participant Response Bias in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1321–1330. https://doi.org/10.1145/2207676.2208589

[18] Devanuj and Anirudha Joshi. 2013. Technology Adoption by 'Emergent' Users: The User-usage Model. In *Proceedings of the 11th Asia Pacific Conference on Computer Human Interaction (APCHI '13)*. ACM, New York, NY, USA, 28–38. https://doi.org/10.1145/2525194.2525209

[19] Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, Ashish Mittal, Prasanta Kumar Ghosh, Preethi Jyothi, Kalika Bali, Vivek Seshadri, Sunayana Sitaram, Samarth Bharadwaj, Jai Nanavati, Raoul Nanavati, Karthik Sankaranarayanan, Tejaswi Seeram, and Basil Abraham. 2021. Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages. In *Interspeech 2021*. 2446–2450. https://doi.org/10.21437/Interspeech.2021-1339 arXiv:2104.00235 [cs, eess]

[20] Paul Dourish. 2014. NoSQL: The Shifting Materialities of Database Technology. *Computational Culture* 4 (2014).

[21] Arienne M. Dwyer. 2008. Ethics and Practicalities of Cooperative Fieldwork and Analysis. In *Essentials of Language Documentation*, Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel (Eds.). De Gruyter Mouton, 31–66. https://doi.org/10.1515/9783110197730.31

[22] D. Eberhard, G.F. Simons, and C.D. Fennig. 2023. *Ethnologue: Languages of the World* (26th ed.). Sil International, Dallas, Texas.

[23] M. C. Elish and danah boyd. 2017. Situating Methods in the Magic of Big Data and AI. *Communication Monographs* 85, 1 (2017), 57–80. https://doi.org/10.1080/03637751.2017.1375130

[24] Eric Le Ferrand, Steven Bird, and Laurent Besacier. 2021. Phone Based Keyword Spotting for Transcribing Very Low Resource Languages. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*. Australasian Language Technology Association, Online, 79–86. https://aclanthology.org/2021.alta-1.8

[25] Rikin Gandhi, Rajesh Veeraraghavan, Kentaro Toyama, and Vanaja Ramprasad. 2009. Digital Green: Participatory Video and Mediated Instruction for Agricultural Extension. *Information Technologies & International Development* 5, 1 (April 2009), pp. 1–15.

[26] James Paul Gee. 1986. Review of Orality and Literacy: From The Savage Mind to Ways with Words. *TESOL Quarterly* 20, 4 (1986), 719–746. http://www.jstor.org/stable/3586522

[27] Jack Goody. 1977. *The Domestication of the Savage Mind*. Cambridge Univ. Press, Cambridge.

[28] Trina Gorman, Emma Rose, Judith Yaaqoubi, Andrew Bayor, and Beth Kolko. 2011. Adapting Usability Testing for Oral, Rural Users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 1437–1440. https://doi.org/10.1145/1978942.1979153

[29] Anil K. Gupta. 2016. *Grassroots Innovation: Minds On The Margin Are Not Marginal Minds*. Random House India.

[30] Richard H. R. Harper. 2019. The Role of HCI in the Age of AI. *International Journal of Human–Computer Interaction* 35, 15 (Sept. 2019), 1331–1344. https://doi.org/10.1080/10447318.2019.1631527

[31] Hilary Hutchinson, Heiko Hansen, Nicolas Roussel, Björn Eiderbäck, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, and Helen Evans. 2003. Technology Probes: Inspiring Design for and with Families. In *Proceedings of the Conference on Human Factors in Computing Systems - CHI '03*. ACM Press, Ft. Lauderdale, Florida, USA, 17. https://doi.org/10.1145/642611.642616

[32] Tim Ingold. 2010. The Textility of Making. *Cambridge Journal of Economics* 34, 1 (Jan. 2010), 91–102. https://doi.org/10.1093/cje/bep042

[33] Sri. R V Karnan. 2016. Banjara Dictionary. https://play.google.com/store/apps/details?id=org.rvm.ban.mle.dict.

[34] Ondrej Klejch, Electra Wallington, and Peter Bell. 2022. Deciphering Speech: a Zero-Resource Approach to Cross-Lingual Transfer in ASR. In *Proc. Interspeech 2022*. 2288–2292. https://doi.org/10.21437/Interspeech.2022-10170

[35] George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

[36] Daniel Lambton-Howard, Robert Anderson, Kyle Montague, Andrew Garbett, Shaun Hazeldine, Carlos Alvarez, John A. Sweeney, Patrick Olivier, Ahmed Kharrufa, and Tom Nappey. 2019. WhatFutures: Designing Large-Scale Engagements on WhatsApp. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14.

[37] John Law. 2004. *After Method: Mess in Social Science Research*. Routledge, Oxon, UK.

[38] Jo Lee and Tim Ingold. 2006. Fieldwork on Foot: Perceiving, Routing, Socializing. In *Locating the Field: Space, Place and Context in Anthropology*, Simon Coleman and Peter Collins (Eds.). Routledge.

[39] Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, and Metze Florian. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8249–8253.

[40] Lev Manovich. 2001. *The Language of New Media*. MIT Press, Cambridge, MA.

[41] Gary Marsden, Andrew Maunder, and Munier Parker. 2008. People Are People, but Technology Is Not Technology. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 366, 1881 (Oct. 2008), 3795–3804. https://doi.org/10.1098/rsta.2008.0119

[42] Marcel Mauss. 2000. *The Gift: The Form and Reason for Exchange in Archaic Societies*. W.W. Norton, New York, NY.

[43] Eleanor J. McAlpine. 2000. India and Nepal Word Lists. https://www.sil.org/resources/archives/81509.

[44] Fatemeh Moradi, Linnea Öhlund, Hanna Nordin, and Mikael Wiberg. 2020. Designing a Digital Archive for Indigenous People: Understanding the Double Sensitivity of Design. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society (NordiCHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3419249.3420174

[45] Preeti Mudliar, Jonathan Donner, and William Thies. 2012. Emergent Practices Around CGNet Swara, Voice Forum for Citizen Journalism in Rural India. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development (ICTD '12)*. ACM, New York, NY, USA, 159–168. https://doi.org/10.1145/2160673.2160695

[46] D. B. Naik. 2011. *The Art And Literature Of Banjara Lambanis: Their Art and Literature* (1st edition ed.). Abhinav Publications.

[47] Douglas W. Oard. 2012. Query by Babbling: A Research Agenda. In *Proceedings of the First Workshop on Information and Knowledge Management for Developing Region (IKM4DR '12)*. Association for Computing Machinery, New York, NY, USA, 17–22. https://doi.org/10.1145/2389776.2389781

[48] Walter J. Ong. 2012. *Orality and Literacy: The Technologizing of the Word* (30th anniversary ed.; 3rd ed ed.). Routledge, London ; New York.

[49] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

[50] Neil Patel, Deepti Chittamuru, Anupam Jain, Paresh Dave, and Tapan S. Parikh. 2010. Avaaj Otalo: A Field Study of an Interactive Voice Forum for Small Farmers in Rural India. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 733–742. https://doi.org/10.1145/1753326.1753434

[51] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[52] Sarah Pink. 2008. Mobilising Visual Ethnography: Making Routes, Making Place and Making Images. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 9, 3 (2008).

[53] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur. 2018. Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. In *Proc. Interspeech 2018*. 3743–3747. https://doi.org/10.21437/Interspeech.2018-1417

[54] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

[55] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. In *Proc. Interspeech 2016*. 2751–2755. https://doi.org/10.21437/Interspeech.2016-595

[56] Fang Qiao, Jahanzeb Sherwani, and Roni Rosenfeld. 2010. Small-Vocabulary Speech Recognition for Resource-Scarce Languages. In *Proceedings of the First ACM Symposium on Computing for Development (ACM DEV '10)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/1926180.1926184

[57] Agha Ali Raza, Awais Athar, Shan Randhawa, Zain Tariq, Muhammad Bilal Saleem, Haris Bin Zia, Umar Saif, and Roni Rosenfeld. 2018. Rapid Collection of Spontaneous Speech Corpora Using Telephonic Community Forums. In *Interspeech 2018*. ISCA, 1021–1025. https://doi.org/10.21437/Interspeech.2018-1139

[58] Agha Ali Raza, Mansoor Pervaiz, Christina Milo, Samia Razaq, Guy Alster, Jahanzeb Sherwani, Umar Saif, and Roni Rosenfeld. 2012. Viral Entertainment as a Vehicle for Disseminating Speech-Based Services to Low-Literate Users. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development - ICTD '12*. ACM Press, Atlanta, Georgia, 350. https://doi.org/10.1145/2160673.2160715

[59] Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2022. Opportunities and Challenges of Automatic Speech Recognition Systems for Low-Resource Language Speakers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3491102.3517639

[60] Thomas Reitmaier, Electra Wallington, Ondřej Klejch, Nina Markl, Lea-Marie Lam-Yee-Mui, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2023. Situating Automatic Speech Recognition Development within Communities of Under-heard Language Speakers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3491102.3517639

[61] Nithya Sambasivan, Ed Cutrell, Kentaro Toyama, and Bonnie Nardi. 2010. Intermediated Technology Use in Developing Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 2583–2592. https://doi.org/10.1145/1753326.1753718

[62] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445518

[63] Bernhard Scholkopf, Kah-Kay Sung, Christopher JC Burges, Federico Girosi, Partha Niyogi, Tomaso Poggio, and Vladimir Vapnik. 1997. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE transactions on Signal Processing* 45, 11 (1997), 2758–2765.

[64] Tanja Schultz, Ngoc Thang Vu, and Tim Schlippe. 2013. Globalphone: A multilingual text & speech database in 20 languages. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 8126–8130. https://doi.org/10.1109/ICASSP.2013.6639248

[65] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Vol. 39. Cambridge University Press Cambridge.

[66] J. Sherwani, N. Ali, S. Mirza, A. Fatma, Y. Memon, M. Karim, R. Tongia, and R. Rosenfeld. 2007. HealthLine: Speech-based Access to Health Information by Low-Literate Users. In *2007 International Conference on Information and Communication Technologies and Development*. 1–9. https://doi.org/10.1109/ICTD.2007.4937399

[67] Jahanzeb Sherwani, Nosheen Ali, Carolyn Penstein Rosé, and Roni Rosenfeld. 2009. Orality-Grounded HCID: Understanding the Oral User. *Information Technologies & International Development* 5, 4 (Dec. 2009), pp. 37–50.

[68] Robert Soden, Austin Toombs, and Michaelanne Thomas. 2024. Evaluating Interpretive Research in HCI. *Interactions* 31, 1 (Jan. 2024), 38–42. https://doi.org/10.1145/3633200

[69] Alessandro Soro, Margot Brereton, Jennyfer Lawrence Taylor, Anita Lee Hong, and Paul Roe. 2016. Cross-Cultural Dialogical Probes. In *Proceedings of the First African Conference on Human Computer Interaction (AfriCHI'16)*. Association for Computing Machinery, New York, NY, USA, 114–125. https://doi.org/10.1145/2998581.2998591

[70] Alessandro Soro, Anita Lee Hong, Grace Shaw, Paul Roe, and Margot Brereton. 2015. A Noticeboard in "Both Worlds" Unsurprising Interfaces Supporting Easy Bi-Cultural Content Publication. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. Association for Computing Machinery, New York, NY, USA, 2181–2186. https://doi.org/10.1145/2702613.2732713

[71] Susan Leigh Star. 2010. This Is Not a Boundary Object: Reflections on the Origin of a Concept. *Science, Technology & Human Values* 35, 5 (Sept. 2010), 601–617. https://doi.org/10.1177/0162243910377624

[72] Lucy Suchman. 2002. Practice-Based Design of Information Systems: Notes from the Hyperdeveloped World. *The Information Society* 18, 2 (2002), 139–144. https://doi.org/10.1080/01972240290075066

[73] Lucy Suchman. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions* (2 ed.). Cambridge University Press, Cambirdge, UK.

[74] Jennyfer Lawrence Taylor, Alessandro Soro, Paul Roe, Anita Lee Hong, and Margot Brereton. 2018. From Preserving to Performing Culture in the Digital Era. In *Digitisation of Culture: Namibian and International Perspectives*, Dharm Singh Jat, Jürgen Sieck, Hippolyte N'Sung-Nza Muyingi, Heike Winschiers-Theophilus, Anicia Peters, and Shawulu Nggada (Eds.). Springer, Singapore, 7–28. https://doi.org/10.1007/978-981-10-7697-8_2

[75] Aditya Vashistha, Edward Cutrell, Gaetano Borriello, and William Thies. 2015. Sangeet Swara: A Community-Moderated Voice Forum in Rural India. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 417–426. https://doi.org/10.1145/2702123.2702191

[76] Aditya Vashistha, Abhinav Garg, and Richard Anderson. 2019. ReCall: Crowdsourcing on Basic Phones to Financially Sustain Voice Forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13.

[77] Aditya Vashistha and Agha Ali Raza. 2023. Voice Interfaces for Underserved Communities. In *Introduction to Development Engineering: A Framework with Applications from the Field*, Temina Madon, Ashok J. Gadgil, Richard Anderson, Lorenzo Casaburi, Kenneth Lee, and Arman Rezaee (Eds.). Springer International Publishing, Cham, 589–611. https://doi.org/10.1007/978-3-030-86065-3_22

[78] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2017. Respeak: A Voice-based, Crowd-powered Speech Transcription System. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1855–1866.

[79] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2018. BSpeak: An Accessible Voice-based Crowdsourcing Marketplace for Low-Income Blind People. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13.

[80] Helen Verran and Michael Christie. 2014. Postcolonial Databasing? Subverting Old Appropriations, Developing New Associations. In *Subversion, Conversion, Development*, James Leach and Lee Wilson (Eds.). The MIT Press, Cambridge, Massachusetts. https://doi.org/10.7551/mitpress/9727.003.0005

[81] Marion Walton. 2014. Pavement Internet: Mobile Media Economies and Ecologies for Young People in South Africa. In *The Routledge Companion to Mobile Media*, G. Goggin and Larissa Hjorth (Eds.). Routledge, London, UK.

[82] Marion Walton, Vera Vukovic', and Gary Marsden. 2002. 'Visual Literacy' as Challenge to the Internationalisation of Interfaces: A Study of South African Student Web Users. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems - CHI '02*. ACM Press, Minneapolis, Minnesota, USA, 530. https://doi.org/10.1145/506443.506465

[83] Frederick Weber, Kalika Bali, Roni Rosenfeld, and Kentaro Toyama. 2008. Unexplored Directions in Spoken Language Technology for Development. In *2008 IEEE Spoken Language Technology Workshop*. 1–4. https://doi.org/10.1109/SLT.2008.4777825

[84] Ludwig Wittgenstein. 2009. *Philosophical Investigations* (rev. 4th ed.). Wiley-Blackwell, Malden, MA.

[85] Marisol Wong-Villacres, Aakash Gautam, Deborah Tatar, and Betsy DiSalvo. 2021. Reflections on Assets-Based Design: A Journey Towards A Collective of Assets-Based Thinkers. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 401:1–401:32. https://doi.org/10.1145/3479545

# A TRAINING MULTILINGUAL PHONE RECOGNISER AND RANKER

We trained the multilingual phone recogniser as described in [34]. In particular, we used 20 hours of English data from LibriSpeech [49] and 20 hours of German, French, Spanish, Polish and Russian from GlobalPhone [64]. We trained a small time-delayed neural network [53] acoustic model with Kaldi [54]. We used lattice-free maximum mutual information objective function [55], mapped phones to X-SAMPA, and used the mapped phone sequences as training targets. To improve the robustness of the acoustic model in cross-lingual phone recognition, we used features extracted with the pretrained self-supervised model XLS-R [3] instead of the traditional MFCC features. In particular we used the 300M parameter version of XLS-R and we used representations from the 18th layer as we found that this layer contained the most usable information for cross-lingual phone recognition. During decoding we used a

bi-gram phone language model trained on the multilingual phone recogniser's phonetic transcripts training dataset.

We implemented the ranker as a Support Vector Machine (SVM) [13] with the radial basis function kernel [63] using the Scikit-Learn toolkit [51]. Our SVM model used uni-grams, bi-gram and tri-gram phone sequences with term-frequency inverse-document-frequency weights [65] as features to predict corresponding photos. We used the default SVM parameters and selected the n-gram range using cross-validation.