Social Computing for Social Good in Low-Resource Environments

Aditya Vashistha

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Richard J. Anderson, Chair

James A Fogarty

Kurtis L. Heimerl

Program Authorized to Offer Degree:

The Paul G. Allen School of Computer Science and Engineering

University of Washington

Abstract

Social Computing for Social Good in Low-Resource Environments

Aditya Vashistha

Chair of the Supervisory Committee:

Professor Richard J. Anderson

The Paul G. Allen School of Computer Science and Engineering

Mainstream social computing technologies—like social media platforms, online discussion forums, or crowdsourcing marketplaces—have transformed how people participate in the information ecology and digital economy. They empower mostly urban, affluent, and literate people, and improve their reach to information and instrumental needs. However, these technologies currently exclude billions of people worldwide who are too poor to afford Internet-enabled devices, too remote to access the Internet, or too low-literate to navigate the mostly text-driven Internet. To enable these communities to access and report information, global development researchers and practitioners have used Interactive Voice Response (IVR) technology to create voice-based social computing services (or *voice forums*). These services let users access, report, and share information via ordinary phone calls. However, challenges in managing local language audio content, high cost of voice calls, and technical difficulties in setup makes these services difficult to scale despite their demonstrated impact.

This thesis will present three systems that I built to address these scalability, sustainability, and

replicability concerns. Sangeet Swara is a social media voice forum that uses community moderation by its low-income, low-literate users to manage and moderate audio content recorded in local languages. Respeak is a voice-based crowdsourcing marketplace that enables voice forum users to complete speech transcription tasks vocally to subsidize their cost of voice calls. IVR Junction is free and open source toolkit that enables global development organizations to easily build, set up, and maintain voice forums. Together, these systems fulfill my vision of building scalable, sustainable, and replicable voice-based social computing services that enable people without literacy, smartphones, or the Internet to participate in informative dialogues at both community and global scales.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express deep gratitude and admiration for my advisor, Richard Anderson. Thank you Richard for your advice, support, encouragement, and friendship. I could never finish this journey without you. Thank you for sharing your profound wisdom, encouraging me to tackle big problems, and giving me infinite resources and freedom. I am deeply inspired by your dedication to students and drive to positively impact the world. Thank you for being always available—weekdays and weekends alike—24x7 in all these years. You are my biggest champion! I hope to make you proud, always!

I am forever grateful to Gaetano Borriello for advising me during the first two years of my PhD. Gaetano, thank you for investing your time, energy, and resources in me. I miss you dearly and wish I could spend more time with you! Your life (and untimely death) taught me important life lessons. I hope to take your legacy forward, and will feel accomplished if I can be half of what you were: an endearing human, prolific researcher, and groundbreaking innovator.

I would never come to grad school without the support of two exceptionally genius Microsoft researchers. Thank you Bill Thies and Ed Cutrell for inspiring me and motivating me to pursue a Ph.D. Bill, I am forever indebted to you for your advice and mentorship. Your selflessness, ingenuity, and wisdom are awe-inspiring. Ed, thank you for motivating me to *"get the license"* to travel any path of my choice. Thank you for patiently listening to my ideas, motivating me to work hard, and lending a helping hand when I needed it the most.

I am very fortunate to receive encouragement and guidance from several UW faculty who went

above and beyond to support my research. Thank you James Fogarty, Katharina Reinecke, Kurtis Heimerl, Richard Ladner, and Julie Kientz for shaping my thesis and helping me find an exciting job. Also, thank you Jon Froehlich, Franzi Roesner, and Arvind Krishnamurthy for feedback on my thesis and job talks. I am also thankful to Elise DeGoede and Julie Svendsen for troubleshooting every hurdle I faced in grad school.

My friends and colleagues at UW made this journey very memorable! Thank you ICTD students and alumni for your collaborations, insights, and friendship. Thank you for all the fun and laughter during cricket matches, hikes, lunches and dinners, happy hours, bowling, and countless other hangouts. Thank you Waylon Brunette, Sam Castle, Rohit Chaudhri, Nicola Dell, Abhinav Garg, Philip Garrison, Mayank Goel, Lilian de Greef, Samia Ibtasam, Esther Jang, Matt Johnson, Naveena Karusala, Neha Kumar, Shrirang Mare, Jared Moore, Trevor Perrier, Fahad Pervaiz, Pooja Sethi, Sam Sudar, Galen Weld, and Matt Ziegler. I would not have survived and thrived without you. I am also thankful to Ravi Karkar, Shrainik Jain, and Maaz Ahmad for energizing me with conversations over countless cups of tea and coffee.

I could never accomplish anything without the support of my family. Thank you for your unconditional love, unwavering support, and immeasurable sacrifices. I dedicate this thesis to my parents, Asha Vashistha and Rajendra Kumar Vashistha, who always encouraged me to follow my dreams and fly high! Thank you for inspiring me with your hardwork, dedication, and selflessness. I am very fortunate to have the love and support of my adorable sisters, Rashi Natha and Astha Vashistha. You are my role models and pillars of strength. Thank you for lovingly taking care of my family duties when I was absent. Also, gratitude to my in-laws and brother-in-law, Satya Prakash Natha, for being a constant source of motivation. I am blessed to receive the love of the children in my family. Unnati, Aadhya, and Aavya, you bring me immense joy, happiness, and peace. My life has no meaning without you. Last but not least, my amazing wife, Rashmi, supported me tirelessly through thick and thin with great love and care. Thank you for celebrating my successes and giving me strength

in struggles. You inspire me to be the best version of myself in all roles of life. This dissertation is as much the fruit of your labor as it is mine.

My dissertation work was supported by a range of organizations, including industry research labs, governmental agencies, social enterprises, funding agencies, and grassroots organizations. Thank you Microsoft Research, Facebook, USAID, Humanity United, Access Now, PATH, Enable India, CGNet Swara, Nehru Yuwa Sangathan Tisi, Grameen Vikas Sanasthaan, and Rajasthan Netraheen Kalyan Sangh for investing in my research.

DEDICATION

To my family for their unconditional love, unwavering support, and immeasurable sacrifices.

Chapter 1

INTRODUCTION

Social computing technologies—like social media platforms, discussion forums, and crowdsourcing portals—have transformed how people communicate with each other. In addition to improving access to information, news, and entertainment, they have impacted governance [58, 107], politics [80, 93, 126, 170], civil society movements [151, 156], crisis response [125, 155], marketplaces [62, 84, 135], and healthcare [54, 78, 127], among other parts of our lives. Although concerns regarding data misuse, privacy breaches, and their overuse have grown recently [116, 144], these technologies are continuing to soar, mostly among literate, urban, and connected communities, all across the world. However, despite their promises (and pitfalls), these technologies are currently excluding billions of people worldwide who are too remote to access the Internet, too low-literate to navigate the mostly text-driven Internet, or too poor to afford Internet-enabled devices.

For example, only 45% people in developing countries and 20% in least developed countries used the Internet in 2018 compared to over 80% in developed countries [26]. Just between India and Pakistan, there are nearly 1.1 billion people offline. Although 70% of their populations have access to mobile phones, most people still use basic or feature phones, making it difficult to extend existing social computing technologies on these devices running custom operating systems. Even when people can afford smartphones and the Internet, low literacy skills prevent 26% of adults in India and 42% of adults in Pakistan from using text-based interfaces. Most South Asian languages and dialects are still unsupported by the advances in natural language processing ruling out the use of voice interfaces like Siri and Alexa. These connectivity, literacy, and socioeconomic barriers result in "utility gaps" [74], limiting mobile phone use to making and receiving voice calls.

Recognizing these structural limitations, HCI researchers have used Interactive Voice Response (IVR) technology to create *voice-based* social computing services (or *voice forums*) that let users call a toll-free phone number to record voice messages in their local language and listen to messages recorded by others. For example, Avaaj Otalo [131] lets Indian farmers share agricultural information with each other and ask questions to experts. The users call a toll-free number, and press 1 on the phone keypad to ask a question, press 2 to listen to announcements, and press 3 to listen to radio programs. Voice forums like Avaaj Otalo overcome connectivity barriers by using ordinary phone calls, literacy barriers by using local language speaking and listening skills, and socioeconomic barriers by using toll-free (1-800) lines. Because of their accessible and usable design, these services have found applications in diverse domains—such as health information systems [91, 173], civic engagement services [33, 119], and rural information portals [52, 140, 141, 169]—and have profoundly impacted marginalized communities in low-resource environments. However, the following three limitations significantly impede their potential to scale.

1. **Content moderation**: Users of these services record audio content in local languages with no speech corpus and recognition models, making it extremely difficult to moderate, search, and index these services.

2. **Financial sustainability**: To be accessible to low-income users, these services use expensive toll-free lines. The resultant cost of voice calls poses a huge burden to sustainability, putting these services at risk of being shut down.

3. **Replicability**: These services are technically challenging to build and maintain for global development organizations, thereby making it hard to replicate them in new contexts. Moreover, they are completely disconnected from mainstream social computing platforms, impairing information exchange between local and global communities.

My PhD thesis explored solutions to these three bottlenecks to enable people without literacy, smartphones, or the Internet to participate in informative dialogues at both community and global scales.

The remainder of this chapter summarizes my work to address these bottlenecks and describes the high-level contributions made in this dissertation.

1.1 Using Community Moderation to Scale Voice Forums

Most voice forums rely on manual moderation by a dedicated team of moderators to identify poor-quality content and categorize audio posts. Aside from challenges in training moderators to understand community expectations, the cost of manual moderation gets prohibitively expensive as the content on these services starts to scale.

To overcome the content moderation challenge, I designed, built, and deployed Sangeet Swara—the first community-moderated, voice-based social media service that lets its users record, listen to, vote on, and share songs, poems, and other cultural content [161]. As users listen to messages, Sangeet Swara requests them to annotate the quality and category of the content by pressing phone keys (for example, press 1 to upvote or 2 to downvote the message). It then uses collaborative filtering techniques to rank, order, and categorize audio messages based on users' votes. I designed new community moderation algorithms for ranking and filtering audio content because of differences in the properties of voice and text (e.g., audio content is more difficult to skim than text), and in the features of IVR-based and text-based interfaces (e.g., the former tracks content users skip more accurately than the latter).

In an eight-month deployment, Sangeet Swara received broad and impassioned usage by marginalized people in rural and peri-urban India: it received 53,000 phone calls from 13,000 callers who submitted 6,000 voice messages in 11 languages as well as 150,000 votes. Sangeet Swara also found unexpected uptake among blind people, who were invested in building and maintaining a community and used the service to expand their social network to distant locations [162]. Community moderation by callers, nearly four-fifths of whom were first-time users of a social networking platform, was 98% accurate in content categorization, made meaningful distinctions between high- and low-quality posts, and performed judgments that were in 90% agreement with expert modera-

tors. This research demonstrated that low-income, low-literate users of social media voice forums can moderate and categorize audio content in local languages themselves, thereby addressing the content moderation challenge. Chapter 3 describes the design, implementation, deployments, and evaluation of Sangeet Swara.

1.2 Using Voice-based Speech Transcription to Financially Sustain Voice Forums

While a few voice forums sustain themselves through advertising, external grants, and partnerships with mobile network operators (MNO) or governments, these alternatives are often beyond the reach of bottom-up development-focused voice-based services; advertising requires a massive initial investment to attain a scale that is lucrative for advertisers; external funding opportunities are unpredictable; partnerships with MNOs and governments are seldom possible.

To overcome the financial sustainability challenge, I examined whether low-income users of these services could complete useful work on their mobile phones to offset their participation costs on services like Sangeet Swara. Since literacy and connectivity barriers render mainstream crowdsourcing marketplaces such as Mechanical Turk unfeasible in this context, I designed and built Respeak—the first voice-based crowdsourcing marketplace that pays users to transcribe audio files using their speech [163, 166, 167]. To transcribe an audio file, Respeak sends small audio segments to multiple users and pays them via mobile airtime for each submitted transcript. Instead of typing the transcript on a phone's keyboard with constrained physical space, users re-speak (i.e., repeat) audio content into an off-the-shelf speech recognition engine and submit the speech recognition output as a transcript. Once multiple users submit transcripts for a particular segment, Respeak combines the transcripts using sequence alignment algorithms to reduce random speech recognition errors.

I conducted multiple cognitive experiments, usability studies, and experimental evaluations to evaluate the feasibility, usability, and acceptability of Respeak. For example, I investigated how audio segment length and presentation order affects content retention and cognitive load on Respeak

users, and whether speaking is indeed a more efficient and usable output medium for transcription than typing. I examined accessibility and usability barriers in Respeak and compared them to those in mainstream microtasking platforms. I also examined how different phone types, channel types, and modes to review transcripts affect task accuracy and completion time.

After incorporating insights from these evaluations into Respeak's design, I deployed it to 73 low-income students [166], blind people [167], and rural residents [163] for nearly two months by partnering with Indian Institute of Technology Bombay (IIT Bombay), Enable India, and Nehru Yuwa Sangathan Tisi (NYST), respectively. Collectively, users transcribed 70 hours of audio data by completing 50,000 micro tasks with an average accuracy of 70% and earned ₹31,000 at an hourly rate that exceeds the average hourly wage in India. Respeak then merged transcripts from multiple users to produce transcription with over 90% accuracy at one-fourth of the market rate, generating sufficient profit to subsidize participation costs of other voice-based services. My analysis indicated that one minute of crowd work on Respeak could subsidize eight minutes of airtime on services like Sangeet Swara. This research also demonstrated the feasibility of a crowdsourcing marketplace that is accessible via ordinary phone calls from the most basic phones. Chapter 4 describes the design, implementation, deployments, and evaluation of Respeak.

1.3 Building a Toolkit to Replicate Voice Forums

To enable global development organizations with limited technical resources to replicate voice-based social computing services like Sangeet Swara, I designed and built IVR Junction and released it as free and open source software [168]. Using services deployed on IVR Junction, basic mobile phone users can record and listen to messages, and their voices can be heard by a global community on Facebook or YouTube. IVR Junction's distributed architecture makes it resilient to network blackouts by repressive regimes to curb dissent and cost-effective due to its use of geographically distributed local access points instead of expensive long-distance phone calls to a centralized server.

Many organizations have used IVR Junction to create voice-based social computing services in

South Asia and Africa; these services have received 110,000 phone calls from nearly 25,000 people who spent 6,100 hours to access, report, and share content. For example, the office of the President of Somaliland used IVR Junction to establish a direct communication channel between parliamentarians and indigenous people in a region with fragile political institutions and polarized media [87]. In five months, the users recorded over 4,300 audio messages that were also indexed on the official website of the parliament of Somaliland. The US Agency for Global Media used IVR Junction to provide a three-minute news broadcast and receive feedback during the Mali civil war. Similarly, women's rights activists in India used it as a voice petition platform after a gang rape incident that sparked international outrage. Chapter 5 describes the design, implementation, and deployments of IVR Junction.

1.4 Benefits and Pitfalls of Voice Forums

In addition to designing, building, and deploying voice forums, I also examined how people in low-resource environments use them. In particular, I investigated how different user groups perceived benefits and limitations of these services. For example, the user analysis of Sangeet Swara found surprisingly high adoption from low-income blind people. To examine the reasons for this, I conducted the first analysis of how low-income blind people in India use mainstream social media services and why they gravitate towards social media voice forums like Sangeet Swara [162]. I found that most of this population does not explore mainstream platforms due to: severe financial constraints that impede their access to smartphones and the Internet; difficulties in understanding the language and accent of screen reader software; cross-cultural usability issues; and a lack of training and help. The few blind people who use these platforms struggle with inaccessible features, like the lack of captions on photos, the lack of screen reader commands for Facebook chatting, and the difficulties in searching for friends. In contrast to mainstream platforms, Sangeet Swara was accessible because of its reliance on basic phones, toll-free lines, and a speech interface. Sangeet Swara offered several benefits to them, for example, it helped them gain self-confidence and knowledge, discover other blind people in distant locations, and build social capital by sharing informative content with them.

Voice forums, like any other social platform, come with their own pitfalls. They end up reflecting the existing sociocultural norms and values of the society, including its shortcomings and biases. For example, while Sangeet Swara served as an instrument of inclusion for low-income blind people, it failed to create a welcoming environment for female users; only 7% of Sangeet Swara users were women despite its accessible and usable design. To explore the reasons for this, I conducted an in-depth examination of the use of social media voice forums by women and men [164]. I found that women users faced systemic discrimination and harassment in the form of posts that contained abuses, threats, flirtatious behavior, and blackmailing. Most women were extremely hesitant to object to harassment directed at them, primarily due to deep-rooted patriarchal values that discourage them from arguing and questioning others. Most male users perceived women as objects of desire. They condoned the unruly behavior of other men and disapproved of objectionable messages less strongly than did women. Using a feminist HCI lens [56, 57], I proposed how these services could be re-designed to provide an equitable and inclusive platform to women. This included recording women-friendly IVR prompts, subsidizing women's participation through non-monetary incentives, and creating new filters to mask women's personal identities. Chapter 6 presents the analysis of benefits and limitations of voice-based social computing systems for people in low-resource environments.

Finally, Chapter 7 discusses the relevance of the dissertation findings and presents new challenges as well as big frontiers in building social computing systems for social good in low-resource environments.

1.5 Contributions

Over the course of this thesis, I have **built** scalable, sustainable, and replicable social computing systems for people who face literacy, socioeconomic, and connectivity barriers (e.g., [161, 166, 168]). I have **systematized** how these new users produce, consume, and share content in offline and online social spaces (e.g., [162, 164, 165]). I have **deployed** social computing technologies to achieve *social*

good by improving people's access to information and instrumental needs (e.g., [87, 167]).

In particular, I made the following contributions:

- I built the first community-moderated social media voice forum to connect people in low-resource environments. I demonstrated that voice forum users, most of whom are low-literate people, rural residents, and blind people, can moderate local language audio content in their digital community without any outside support.

- I built the first voice-based crowd-powered speech transcription marketplace for voice forum users who lack literacy skills as well as access to Internet-connected devices. I demonstrated that low-income students, blind people, and rural residents—the predominant users of voice forums—can vocally transcribe audio files with high accuracy. I showed that the profits from crowd work can provide earnings as well as airtime to voice forum users, thereby addressing the financial sustainability challenge.

- I built a free and open source toolkit that makes it easier for global development organizations to build and set up voice forums. I showed how several governmental agencies, social enterprises, and grassroots entities use the toolkit to deploy voice forums in low-resource environments.

- I analyzed the benefits and limitations of the voice-based social computing systems that are designed to achieve social good for marginalized people in low-resource environments.

Taken together, these contributions demonstrate the feasibility of building scalable, sustainable, and replicable voice-based social computing systems that address instrumental and information needs of people in low-resource environments. In addition to solving technical challenges to provide people with access to social computing systems, this thesis enrich the understanding of how these new users produce, consume, and curate local content, how they rely on offline and online social networks

to meet their information needs, and what incentives motivate them to share information. These insights will play a pivotal role in informing the design of future social computing systems. Finally, this thesis contributes to a growing discussion about the positive as well as negative impact of social computing on society, and offer suggestions to make these systems more diverse, inclusive, and impactful.

Chapter 2

RELATED WORK

Mobile phones have profoundly impacted how people worldwide interact with each other. They have made significant inroads especially in developing regions that account for nearly 80% of the world's 8 billion mobile phone subscriptions [26]. However, unlike developed regions, most subscribers in developing regions own basic or feature phones instead of smartphones. Most of them use their phones primarily for making and receiving voice calls due to barriers highlighted in Chapter 1. The World Wide Web has enabled rich communities of user-generated content in *resourceful* environments, lending a platform for users to share news, entertainment, and information with each other. But in *low-resource* environments, is it possible for low-income, low-literate people to participate in an informative dialogue in the absence of Internet connectivity?

Global development researchers and practitioners have responded to this challenge by using interactive voice response (IVR) technology to create voice-based social computing services (or voice forums) that lets users call a phone number to access, report, and share information in their local languages. These services have found applications in diverse domains due to their accessible and inclusive design, and have profoundly impacted marginalized communities in low-resource environments. In this chapter, I present related social computing research that influenced this dissertation work. In particular, I discuss how these voice forums evolved in the last two decades, what challenges plagued their growth, and what solutions were used to address their pain points.

2.1 Evolution of Voice Forums

The first wave of voice-based services focused on improving information access for people in low-resource communities. For example, HealthLine enabled low-literate community health workers in Pakistan to retrieve relevant information by speaking out pre-defined commands [149]. While initial efforts like HealthLine allowed users to only consume information, subsequent services took the form of voice forums and enabled marginalized communities to also produce and share information. This included Avaaj Otalo (an agriculture discussion forum) [131], CGNet Swara (a citizen journalism service) [30], Mobile Vaani (a social media service) [33, 118], and IBM's Spoken Web (a user-generated information directory) [52]. These services allowed users to report, access, and share information via ordinary phone calls. For example, CGNet Swara [30] enables rural communities in India to report and listen to locally relevant news, grievances and cultural content. The users call a toll-free phone number, and press 1 on the phone keypad to record a new message in their own language, and press 2 to listen to messages recorded by others. Recorded messages are fact-checked, published on a website and the forum, and viewed by activists, government actors, and the mainstream media. Since its inception, CGNet Swara has received over 600,000 phone calls, 6,500 reports, and resulted in resolution of over 300 grievances, thereby positively impacting lives of low-resource rural residents.

The success of these initial voice forums demonstrated their great potential to enable information access and connectivity among underserved populations in diverse HCI4D contexts. However, the vast majority of these services ran into the hurdles of user training and technology adoption. Nearly a decade back, the biggest roadblocks to designing voice forums were usability, motivation, and spread [98, 106]; target populations faced difficulties in using even the simplest of speech-based telephone interfaces, they did not exhibit interest or trust in using such services, and it was difficult to advertise and spread such services to under-connected people. Researchers tried to overcome these barriers by conducting in-the-lab-training as well as door-to-door field campaigns, but it was quickly realized that these approaches were not scalable.

The second wave of voice forums focused on addressing these concerns. For example, Raza et al. used a ludic design approach to train users, and promote usability and spread. They built Polly, a voice-based entertainment service that lets users make a short audio recording, apply funny voice modifications to it, and share it with their friends via automated voice calls [139]. They deployed Polly to five low-income people in Pakistan in early 2012. Within a year, Polly spread virally to over 165,000 users via 636,000 calls without any outreach efforts. Polly's playful design trained users to navigate IVR interfaces, and also led to its viral adoption. Raza et al. then used Polly to share instrumental information with users to aid their socioeconomic development [141]. In an initial test, 34,000 Polly users listened to 728 job advertisements nearly 386,000 times within a year.

Over the last seven years, Polly has been successfully used in multiple countries to rapidly spread useful information to underserved populations. In 2014, at the peak of the Ebola crisis in West Africa, Polly-Santé (meaning "Polly-Health") was deployed as an emergency disaster-response service in Guinea to spread reliable information about prevention, symptoms, and cure of Ebola [171]. The information originated from the Centers for Disease Control and the service was funded by the U.S. Embassy in Conakry. A key hurdle to information dissemination in the Guinean context is great linguistic diversity and the lack of a widely understood common language. Fortunately, this is not a major impediment for voice forums. Polly-Santé was launched in 11 local languages and reached more than 7,000 local mobile phone users within a few months. In 2014, Polly was also used by `Babajob.com`—an online job portal in India—to advertise a voice directory of available jobs to thousands of low-literate job seekers [138].

Since 2016, Polly has been active in Pakistan as a gateway to maternal health information for underconnected expectant parents. Polly advertises a hotline called Super Abbu (meaning "Super Dad") that allows expectant parents to record health questions that are answered by volunteer doctors. Such private and anonymous access to trained gynecologists allows parents to ask questions around pregnancy and childbirth that are often considered sensitive and even taboo topics in the local context. The service specifically targets fathers to promote paternal participation and allow them to share their experiences with their peers. In its initial deployment, Super Abbu reached 21,000

users—96% of them were men—in just two months, uncovering a pent-up demand for maternal health information and giving the target population an agency to anonymously access culturally sensitive yet lifesaving reliable information.

In the last decade, many more voice forums have emerged to meet information and instrumental needs of people in diverse HCI4D contexts, including health [50, 68, 91, 95, 173], civic engagement [87], agriculture [143, 150], education [86, 108], employment [169], and social media [76]. Together, these voice forums and others like CGNet Swara [30], Polly [141], and Mobile Vaani [33] have attracted millions of calls and audio recordings.

The growth of voice forums have also motivated HCI4D researchers to investigate several design elements carefully, for example, how input via speech or DTMF keys affect navigation [130], whether adaptive UI is more usable than fixed UI [55], how data collection on a live call compare with recorded prompts [69], and whether people prefer posts from experts or peers [132]. Others have examined the impact of these services on people in low-resource environments (e.g., rural communities [109, 110, 118], people with disabilities [76, 162]). More recently, researchers have used these services as a means to rapidly collect spontaneous speech corpora containing content diversity, speaker diversity, and natural speech elicitation for low-resource languages and accents [137]. However, despite their demonstrated impact on marginalized communities and language research in low-resource environments, voice forums lack the potential to scale, sustain, and replicate because of the following three challenges:

1. **Content moderation:** Since users of these services record audio content in local languages with no speech corpus and recognition models, it is extremely difficult to manage content on these services.

2. **Financial sustainability:** Since these services pay for expensive toll-free lines to be accessible to low-income callers, the resultant cost of voice calls poses a huge burden to financial sustainability.

3. **Replicability:** Since these services are technically challenging to build and maintain for global development organizations, they are very hard to replicate in new contexts. Also, they often operate in silos, impairing information exchange between local and global communities.

In the following sub-sections, I present prior research efforts to address these three bottlenecks to make voice forums more scalable, sustainable, and replicable.

2.2 Managing Content on Voice Forums

Since voice forums generate audio content in low-resource languages and accents unsupported by advancements in natural language processing, it is very difficult to automate categorization and moderation of posts and responses. This makes it very hard for users to browse data and providers to regulate these services. Also, voice forum users listen to audio content in a sequential manner and are unable to skim the audio unlike textual content. In addition, voice forums are often deployed in low-resource environments where most people lack technical know-how of social media, making them particularly vulnerable to disinformation and fake news [115]. For example, recent acts of mob violence in India have been attributed to fake WhatsApp messages [61, 82], which can also be easily recorded and shared on voice forums. These reasons makes content moderation critical to present users with respectful, accurate, and high-quality recordings.

Large voice forums typically employ a dedicated team of moderators, who screen recordings, offer feedback to contributors on how to record good posts, and perform tagging, categorization, and moderation of audio posts. For example, both CGNet Swara [30] and Gram Vaani [33] currently employ 10–15 full-time moderators. Although manual moderation is highly accurate, it becomes difficult to scale as these services grow, due to high cost, delayed response, and challenges in hiring moderators who are familiar with local context.

Several news and social networking sites like Reddit, Slashdot, and Stack Overflow draw on collaborative filtering and community moderation algorithms to manage user-generated content [83,

142, 148, 157]. They use community votes and recency to determine high quality and contextually relevant user-generated content. Since most of the content on these platforms is textual, a number of natural language processing techniques have been employed to categorize, annotate, and moderate content, and even predict emotions. However, no prior work has focused on using community moderation on a voice forum, which is different in several ways from community moderation on a text-based forum. For example, audio content is more difficult to skim than textual content, meaning that users may lose patience in hearing and ranking lower-ranked posts. An IVR system can also track exactly what a user listened to and what content they skipped, which is difficult to do on a webpage. Finally, the limited affordances of an IVR interface and the limited technology skills of voice forum users add more constraints to the design of community moderation algorithm.

Exploring mechanisms to scale community moderation and content curation has received limited attention in the HCI4D community. Most closely related to the dissertation work is a system called Gurgaon Idol, where voice forum users rated audio recordings to influence eventual playback on a community radio station [98]. However, this research focused on the usability of the recording and voting interface, and the training of users, rather than influencing the playback order or enabling users to perform moderation tasks. In this thesis, I design new community moderation algorithms that are well-suited for voice forums and examine whether low-income, low-literate voice forum users can manage user-generated content on their community themselves.

2.3 Financial Sustainability of Voice Forums

There is a growing concern among global development researchers and practitioners about the high operating costs of voice forums [111]. Providers of these services often pay for the cost of acquiring toll-free lines so that low-income people can access these services for free. However, this cost becomes prohibitive as the usage increases, often putting these services at risk of being shut down. For example, Polly was discontinued several times because of the lack of resources to meet growing call volumes [139]. Many voice forums rely on external funding to subsidize the cost of voice calls,

however, the unreliable nature of grants and awards makes this approach unsustainable. For example, the founder of CGNet Swara expressed frustrations on how limited funding to subsidize phone calls may cause them to *"shut down completely"* [13].

A few voice-based services such as Kan Khajura Tesan [34]—an on-demand entertainment service in India from a consumer goods company with USD 5 billion revenue—and Mobile Vaani [33]—a social media voice forum with over 5 million users in central and north-eastern India—have used advertising revenues to subsidize the cost of voice calls. These services advertise products and services that cater to low-income consumers in rural and peri-urban areas (e.g., small sachets of washing powder, toothpaste, soap). Although these services are existential proof of advertising as a viable approach to financially sustain large-scale voice forums, the initial investment required to gain critical mass for advertising is often beyond the reach of bottom-up development-focused voice forums.

Some voice forums such as Ila Dhageyso [87]—a service to connect citizens with government officials in Somaliland—and 3-2-1 service [27]—a phone call-based search engine in Africa—have partnered with government agencies and mobile network operators to subsidize the cost of voice calls. Although such partnerships greatly reduce the burden of voice call costs, building and maintaining such partnerships is seldom possible due to mismatch in goals, expectations, and values.

Recent years have seen advances in the availability of low-cost smartphones and affordable 2G connectivity in developing regions. Smartphones are quickly leapfrogging traditional desktop computers because of their low-cost, portability, and intuitive touchscreen interfaces. Banking on these advances, some researchers and practitioners have created voice-based smartphone applications that mimics IVR application, but uses data channel instead of voice channel to upload and download voice messages. For example, instead of accessing CGNet Swara via an expensive phone call, D'Silva et al. have designed a smartphone application that downloads voice messages over the mobile Internet and plays them locally on the device [75]. In addition to reducing operational cost by 25 times, the smartphone application also enables offline access. Once an audio file has been downloaded by the application, it can be played locally without any connection to the server. In addition

to enabling replay of content in regions with intermittent connectivity, this avoids the costs incurred by repeatedly streaming the same content to the same user (a very common practice today). Saving audio files locally also enables users to propagate them via Bluetooth and SD card sharing, extending their reach to local feature phone users for free. Rural activists have been using the application since September 2014. They provide intermediate access [145] to rural beneficiaries who lack access to smartphones and Internet connectivity. Although this approach is promising, it can address the financial sustainability challenge only when a majority of people in low-resource environments use smartphones and the Internet.

Given these limitations in existing approaches to financially sustain voice forums, there is a need to find alternatives to reduce the burden of phone calls on voice forum providers. This thesis examines: (1) would users be willing to pay for their phone calls to access a voice forum; (2) can profits from paying users be used to subsidize participation of low-income users, and (3) can profits from crowd work by voice forum users be used to subsidize their participation costs?

2.4 Challenges in Replicating Voice Forums

Despite the enthusiasm surrounding voice forums, the unfortunate reality is that it remains quite complex to install and configure them. Many services—such as CGNet Swara [30], Avaaj Otalo [131], and Phone Peti [99]—utilize open-source platforms like Asterisk [1] or FreeSWITCH [6] for the telephony interface, and require hosting a Web server to connect with moderators. Although tractable for technology researchers, using these platforms requires Linux expertise that is usually beyond the reach of many non-profits and non-governmental organizations.

Most voice forums have a centralized architecture that provides a single access point (or calling number) for users. This makes it difficult to scale voice forums and extend them in new geographic locations. For example, if a non-profit organization would like to scale a voice forum operating in region A to another region B, then either users living in B would have to make an expensive

long-distance phone call to the access point in region A or the organization would have to set up a local service in B, thereby disconnecting people in two locations. Also, most voice forums are disconnected from mainstream social computing systems like Facebook and Twitter where people with smartphones and Internet connectivity communicate with each other. As a result, most voice forums operate in Silos, impairing information exchange between different local communities as well as global audience.

While there exist many toolkits for building and replicating voice forums, none of the open-source platforms offer distributed and scalable operations, catering to both local callers and global audience on the Internet. Freedom Fone [5] and Awaaz.De [2] are built on FreeSWITCH, and integrate a voice forum with an Internet site for viewing audio recordings. Similarly, the IBM Spoken Web project proposes a "World Wide Telecom Browser" that acts as a single access point as the user browses content hosted on separate servers [51]. None of these projects support distributed access between synchronized local servers. Also, they do not connect voice forums with mainstream social media portals like Facebook or Twitter.

Recent years have seen tremendous growth in organizations that support software-as-a-service model. These advances have also impacted how voice forums are build and set up. For example, cloud telephony systems—like Twilio [11], Tropo [10], Exotel [4], KooKoo [8], and engageSPARK [3]—encapsulates the implementation details from users, and makes it *very* easy for organizations that lack technical expertise to build and maintain voice forums. Although these systems are robust and usable, they are very expensive to use. Also, they store audio content in their own servers, raising concerns related to security, privacy, and data abuse. Finally, these systems do not synchronize content across distributed call centers, making them less scalable and replicable. This thesis contributes to these ongoing efforts by presenting a toolkit that makes it easy to build distributed and connected voice forums for people in low-resource environments.

Chapter 3

COMMUNITY MODERATION OF VOICE FORUMS

The Internet has transformed the way we conduct our lives and connect with others. Because of its profound impact on society, it is often considered as the new Industrial Revolution [53, 94]. However, in the current form, it excludes billions of people worldwide who are too poor to afford it, too low literate to use it, or too remote to access it. Voice forums have emerged as an inclusive and accessible alternative to the Internet for people living in low-resource environments and as-yet unconnected communities. In recent years, voice forums have been used to address information and instrumental needs of marginalized people, including low-literate [108, 143], rural [50, 68], disabled [76], refugee [158], indigenous [87, 119], and many other communities [91, 95, 173]. Together, these services have received millions of calls and voice messages in local languages.

However, one bottleneck that has prevented voice forums from rivaling the scale of large Internet websites is the process of content curation and moderation. In order to ensure availability of respectful, accurate, and high-quality recordings, large voice forums typically employ a dedicated team of moderators who tag, categorize, and moderate posts. However, manual moderation is difficult to scale for a range of reasons. For example, if these platforms grow by orders of magnitude, it would be very difficult to manage the cost and quality of manual moderation.

In this chapter, we examine if community moderation can be used, instead of moderation by experts, to manage user-generated content on voice forums. To examine its feasibility, acceptability, and usability, we first create a vibrant virtual community that is inclusive of low-income users in rural India. We then examine:

- Do participants value their interactions with the community?
- Can the community moderate itself without outside assistance?
- Can it be financially sustainable?

Drawing on lessons learned from prior voice forums, we design and build Sangeet Swara: a social media voice forum that uses *community moderation* to overcome the limitations of a dedicated moderator team [161]. Sangeet Swara enables people in rural India to share songs, poems, jokes, and other cultural content. In addition, it relies on users to categorize the content they hear on the system and rate its quality. These ratings, in turn, influence the order that recordings are played to other listeners, thereby improving the overall user experience. While community moderation has been successfully used on Internet websites, such as Reddit and StackOverflow, to date it has not been used to influence the playback priority on a voice forum. Extending community moderation to an IVR platform involves several unique challenges, including the limited affordances of the interface and users' limited experience with technology, especially in rural India.

In this chapter, our primary contribution is the design and 11-week deployment of Sangeet Swara, which we evaluate along three dimensions: the engagement of users, the accuracy of community moderation, and financial sustainability. We find that users were highly engaged, with over 25,000 calls and 5,000 recordings from over 1,500 people. The service found unexpected uptake among people with visual impairments, who were especially passionate about building and maintaining the community. We show that community moderation was 98% accurate in categorizing the content and gender of posts; it also made meaningful distinctions between high-quality and low-quality posts, and made judgments that were in 90% agreement with researchers on a sample of recordings. We also conducted an automated phone interview with 204 users, and offer qualitative findings regarding their perceptions of the service as well as the strengths and limitations of community moderation.

As a secondary contribution, we also advance the dialogue surrounding financial sustainability of

voice forums in low-resource environments. Up until now, voice forums have relied on expensive toll-free lines in order to make them accessible to low-income callers. However, as usage scales, toll-free lines become too expensive to sustain [139]. The most direct solution to this problem is for users to pay for their own calls: an experiment that we tried in two different contexts. In Sangeet Swara, we eventually disabled the toll-free access, but found that even the most fervent rural users were unable to bear the cost of phone calls. As a follow-up experiment, we also describe Talent Hunt: an adaptation of Sangeet Swara to a higher-income context where we experimented whether profits from paying users could be used to cross-subsidize participation of low-income users. Although Talent Hunt received considerable use, including about 27,000 phone calls from 12,000 different callers, this usage was driven mainly by a promotional contest and disappeared as soon as prizes to attract participation were awarded.

In the following sections, we describe the design of Sangeet Swara, its deployment in rural India, and its evaluation using the three metrics mentioned above. We then present the design and deployment of Talent Hunt. Finally, we discuss the lessons learned from these deployments and the implications for future voice forums.

3.1 Sangeet Swara Design

Prior research has demonstrated that entertainment content drives technology adoption by low-income people in the developing world [136, 153]. In fact, even voice forums that are intended for other purposes often see many recordings of songs, religious verses, and other performances [87, 120, 131]. Recognizing the appeal of entertainment, we designed Sangeet Swara, a social media voice forum accessible via phone calls, where callers could record songs, poems, jokes, and other cultural content. A key aspect of the system is that it *ranked* the recordings based on feedback from the community. The ranking aimed to order the posts according to what was most likely to be enjoyed and appreciated by listeners. There was a single, global ranking computed across all recordings and all listeners in the system. In addition to the rank order, the system calculated a

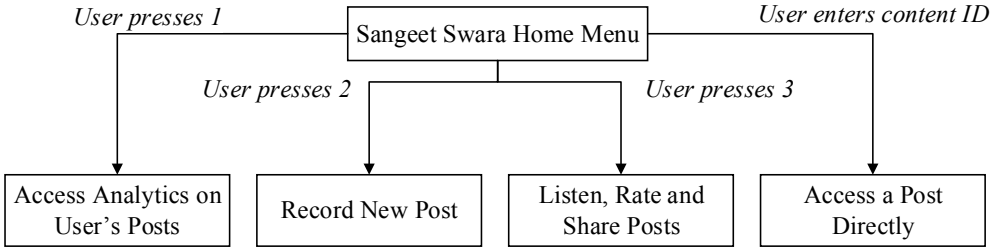

Figure 3.1: High-level call flow of Sangeet Swara.

separate *playback order* that determined which post a listener heard at a given time. The playback order balanced the interests of listeners (who desired to hear high-quality posts) with the interests of content contributors (who desired to have as large of an audience as possible). Both the rank order and playback order were dynamically updated based on listeners' ratings of the content. We give more details on these orderings later.

3.1.1 Call flow

Sangeet Swara relied on key press (DTMF) navigation: users listened to an audio menu and indicated their selection by pressing a digit. Although not as expressive as free-form speech, DTMF interactions have been shown to be robust and also preferred among users in rural India [130].

The high-level call flow is illustrated in Figure 3.1. The first thing that callers heard was an 8-second folk music excerpt, followed by a greeting in a male voice: "Friends, welcome to Sangeet Swara. You can record and listen to songs, poems, and jokes. Please note, it is free to call on this number." Then they were asked to select between the following options:

1. **Check on your posts.** Users who had recorded at least one post could listen to all of their recordings and also learn what rank they had obtained in the system. For users who had not recorded anything, this option was omitted.

2. **Record a post.** Users were encouraged to introduce themselves as part of their recording. We

restricted the length of recordings to 60 seconds (plus a 10-second buffer for the introduction). After recording a post, users received an SMS with a unique five-digit numeric ID for that post. This ID could be shared and used to jump directly to the post (details below).

3. **Listen to other posts and rate them.** When users chose to listen to posts, first we played the top ranked post (introduced as "the best message on the basis of community votes"). Playing the best message first ensured that callers heard at least one high-quality recording per call. It also encouraged friendly competition to be featured in the top spot. After the first post, users listened to other posts in the playback order computed by the system. In advance of playing each post, the system announced its current rank among all posts recorded to date. Since the rank order and playback order were different, users listened to an unpredictable mixture of highly-ranked and low-ranked content.

   After listening to a recording, users were required to give feedback by pressing a key for "like" or "dislike." Users could also interrupt the playback of a post to offer an early judgement, in which case the remainder of that post was skipped. Each user had only one vote to count towards a given post; if they played a post twice, they could change their vote but not increase it. Users could also press a key to receive an SMS that was suitable for sharing with friends. The SMS contained the unique ID of the post and instructions for accessing it on Sangeet Swara.

4. **Jump directly to a post.** Users could directly jump to a post by entering its ID number at the main menu. The first digit of the ID number was always different than the other options at the main menu, enabling users to make the jump immediately without navigating through any other menus.

Our design of menu prompts respected the lessons learned from prior IVR systems in low-income communities [69, 106, 117, 131]. The prompts were recorded in the local language and accent of the target area (North Indian Hindi), with slow and clear diction by the speaker (a male). Prompts

explained each possible action before the corresponding key press; keys had consistent meanings across all menus; multi-digit inputs were avoided as much as possible; and invalid key presses led to explanatory error messages.

We used iterative prototyping to refine the system in advance of deployment. In a formative lab evaluation, 28 callers placed 236 calls over a period of 3 weeks. To understand usability barriers, we performed participant observation and conducted five unstructured interviews. This led to several improvements. For example, before posting a recorded message, the system played it back and asked for confirmation from the user.

Towards the end of our field deployment, we also augmented the call flow with an additional feature. We identified regular users (those who had called at least ten times) and notified them, at the beginning of the call, that they were now a "senior member" of the Sangeet Swara community. Commensurate with this distinction, we asked them to take on a new responsibility, which was to answer one pre-recorded question at the beginning of each phone call. As detailed later, we used these questions both to conduct surveys of the users and to take users' help in categorizing the content recorded by others. The survey questions solicited free-form audio responses, while the categorization questions were multiple choice. Senior members could advance to the main menu of Sangeet Swara only after answering the question posed.

The curation tasks performed by callers on Sangeet Swara are related to crowdsourcing efforts in developing regions. For example, Jana (formerly known as txtEagle [77]), mClerk [88], and MobileWorks [123] enabled users to earn money by completing small tasks on low-cost mobile phones, using either mobile Internet or SMS. In contrast, Sangeet Swara administers tasks via an IVR interface, where callers categorize and moderate audio posts.

3.1.2 Rank Order

The rank order aims to sort posts by increasing order of quality, as determined by users' upvotes and downvotes. There are two criteria that contribute to the rank ordering:

- High scores: a post with a higher ratio of upvotes to downvotes is likely to be of higher quality.
- High confidence: for comparable ratios of upvotes to downvotes, we have more confidence that a post is good if more people have voted on it.

Following Reddit's algorithm for sorting comments [121], we integrate both of these concerns by calculating the lower bound of the Wilson score confidence interval for a Bernoulli parameter:

$$\frac{\hat{u} + z^2_{\alpha/2}/2n - z_{\alpha/2} * \sqrt{[\hat{u}\,(1-\hat{u}) + z^2_{\alpha/2}/4n]/n}}{1 + z^2_{\alpha/2}/n}$$

Here, $\hat{u}$ is the fraction of upvotes, n is the total number of votes, and $z_{\alpha/2}$ is the $(1 - \alpha/2)$ quantile of a standard normal distribution. We used the lower bound of a 95% confidence interval ($\alpha = 0.05$) to compute the rank score. The post with highest score was assigned the top rank.

3.1.3 Playback Order

The playback order refers to the sequence in which a user listens to posts. The playback order needs to balance the following competing criteria:

- Listeners want to hear good content. This prioritizes posts with a large fraction of positive votes.
- Contributors want their posts to receive a fair ranking. This prioritizes posts with a small number of total votes (since more votes lead to a more accurate assessment of quality).

To balance these concerns, we calculate a post's playback priority according to the following formula:

$$\frac{U}{U+D} * (1 - DF)^{U+D}$$

Here, U is the number of upvotes for the post, D is the number of downvotes for the post, and *DF* is a discount factor that serves to balance the concerns of listeners and contributors. We used a discount factor of 0.333 after analyzing a range of values and their impact on example scenarios. We initialized posts with a single upvote to avoid division by zero. The first term of the equation represents the priority of playback according to listeners (criterion 1), while the second term captures the priority for contributors (criterion 2). Larger numbers represent a higher priority.

When a user elects to listen to posts, the system plays the highest priority posts that the user has not yet voted on. If the user has voted on all the posts, then attention is restricted to posts the user has liked in the past, and playback proceeds in rank order instead of playback order.

Our calculation of priority is similar to other rankings that reconcile the competing metrics of quality and recency. In our context, recency corresponds to the total number of votes that a post has received to date, rather than the time elapsed since the recording.

3.2 Sangeet Swara Deployment

Sangeet Swara was deployed for eleven weeks in India. In order to lower the barriers to participation, we launched the service using a toll-free (1-800) number. However, as toll-free lines become expensive at a large scale, we also wanted to explore if users would pay for the phone calls. Thus, we moved the service to a regular number after about seven weeks.

To create awareness about Sangeet Swara in rural and small-town India, we posted a message on CGNet Swara, a voice forum for citizen journalism that has considerable reach in rural areas. The post was accessible to CGNet Swara callers for two days, during which time it was heard by 393 unique callers. Out of those, 73 people placed a call to Sangeet Swara. In order to help Sangeet Swara feel familiar to prior users of CGNet Swara, and to set the standards for the community, we

Deployment Duration	11 weeks
Language of Prompts	Hindi
Calls	25,381
Callers	1,521
Posts	5,376
Contributors	516
Average Call Duration	4.96 mins
Total Plays of Posts	198,898
Upvotes	40,590
Downvotes	99,150
Share Requests	773
Direct Jumps to Post ID	7,871

Table 3.1: Usage statistics for Sangeet Swara.

seeded Sangeet Swara with fifteen songs and poems that appeared previously on CGNet Swara.

Sangeet Swara had significant uptake. As summarized in Table 3.1, the system received over 25,000 phone calls from over 1,500 people. There were about 5,000 posts recorded, about 200,000 playback events, and about 140,000 votes cast. Figure 3.2 depicts the usage over time. As detailed in the next section, usage was highest among low-income people from rural and peri-urban areas of northern and central India. Also, the service saw a high uptake among people with visual impairments.

Unfortunately, usage of the system dropped dramatically when we converted the toll-free lines to regular lines. We revisit the question of financial sustainability later, both in the context of Sangeet Swara as well as Talent Hunt.

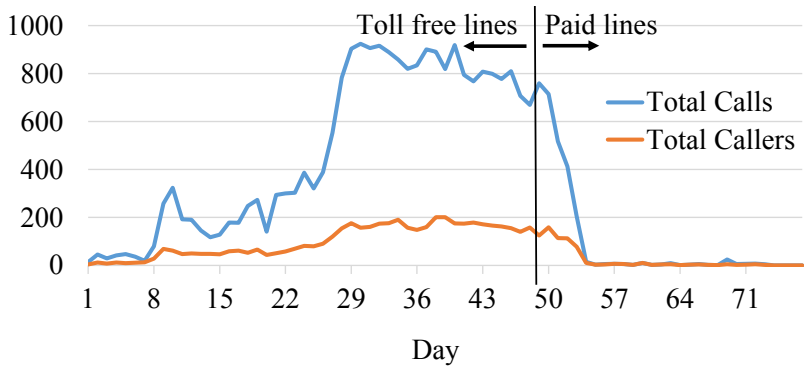

Figure 3.2: Call statistics for Sangeet Swara.

3.3 Evaluating the User Experience

For community moderation to work, users need to value the community and should have the desire to improve it. Thus, we evaluated users' experience of Sangeet Swara, including the worth they attached to the system.

3.3.1 Methods

We used a mixed methods approach spanning qualitative and quantitative analyses. Our primary tool was an automated phone survey, which presented a single pre-recorded question to regular users each time they called. There were nine questions about basic demographic data (age, gender, education, technology exposure, etc.) and six open-ended questions probing the background of the listener (e.g., "tell us about yourself"), their conception of Sangeet Swara (e.g., "how would you describe Sangeet Swara to a friend"), the quality of community moderation, and the strengths and weaknesses of the platform. The survey was live for ten days. All responses were provided in audio format and then translated, transcribed, and analyzed using open coding and axial coding. A total of 204 people (out of 409 regular users) answered one or more of the survey questions, and each question was answered by at least 100 people. On average, free responses were 36 words long.

We performed a content analysis of 100 randomly selected posts. When we learned of the prevalence of visually impaired users, we conducted ten semi-structured telephone interviews to understand their experience in more detail. We used open coding to analyze the audio posts, interviews, and survey responses. We also studied call logs to understand usage patterns.

3.3.2 Results

Sangeet Swara led to high levels of community engagement, with users becoming devoted champions of the system. Similar to other platforms for user-generated content [124], the top 10% of callers placed 70% of total calls. The top 10% of content authors were responsible for 60% of all messages.

User Analysis

Surprisingly to us, Sangeet Swara found broad and impassioned usage by visually impaired users; of those users who told us "something about themselves," 26%[1] voluntarily disclosed that they were visually impaired. The uptake by visually impaired users was organic; although voice is a natural mode of interaction for the visually impaired, we did not anticipate this usage and played no role in promoting it. Our interviews with visually impaired users led to a broader study on their creation, consumption, and sharing of educational content [160].

Our users were predominately young men: 94% were male (average age=25 years) and 6% were female (average age=22 years). The youngest user was an eight year old and the oldest user was a 52 year old man. About half of users were from rural areas or small towns, while others were from larger cities. Users came from a broad range of educational backgrounds: 16% held or were pursuing a master's degree, 40% held or were pursuing a bachelor's degree, 24% were in high school, and 17% were in middle school. Two respondents were in primary school, and one described himself as uneducated. Our users came from a variety of vocations: 54% were students, 17% were teachers, 6%

[1]Because different questions were answered by different numbers of users, we report the percentage of users answering a given question.

were working in private jobs, 5% were unemployed, 4% were musicians, and 4% were farmers. The 42% of users who were employed reported a median annual income of USD 960 with a maximum of USD 7,000.

SMS use was fairly common, with 61% reporting its use. However, most respondents had little experience with the Internet: only 16% had used an email account at least once, and the same fraction had used Facebook. Many users had never heard of Facebook. For example, when we asked them whether they have a Facebook account, three people said: "*We don't have a Facebook account, but we have an account in Bank of India*." On further investigation, we found that they had limited technical expertise, and associated the word "account" with bank accounts instead of accounts on Internet-based services.

Content Analysis

Our open categorization of 100 randomly sampled posts found generic messages (N=36), songs (N=21), poems (N=16), users' introduction about themselves (N=6), songs played from another playback device (N=6), instrumental performances (N=4), jokes (N=3), blank messages (N=3), questions (N=3), and current news (N=2). Eighty-eight messages were recorded by males, two were recorded by females and one message was recorded by a group consisting of both males and females. We did not categorize the gender of blank messages and messages containing playback from other devices. Fifty-six people reported their location while recording the messages. Most of the messages came from the states of Madhya Pradesh (N=25), Rajasthan (N=11), and Uttar Pradesh (N=8). The audio quality was satisfactory for 97 messages.

Songs and poems accounted for about half of the content, and were in a variety of styles. The songs spanned recognizable hits, folk music, and original pieces; solos and duets; a cappella and pieces with instruments. The top 50 posts (analyzed in a later section) are available for listening at `http://soundcloud.com/sangeet-swara/`.

The other half of posts emerged as general social media. Many messages were wishing well to the 'friends' users made on Sangeet Swara. There were greetings, good morning messages, and good night messages for other users, and responses from one user to another. Many messages were about recent topics of national or regional interest. For example, there were five messages about the 2013 North India floods. Eight users gave their phone number and encouraged others to contact them for chatting. Male users often recorded compliments for female contributors, praising their beautiful singing and sometimes requesting their phone number.

While the vast majority of posts on Sangeet Swara were respectful in tone, we found and deleted 22 posts containing abusive language or derogatory comments. As seen in prior forums such as Avaaj Otalo [131], users took an active role in policing the system, e.g., by urging others to record cultural content and to avoid abusive comments. Although we did not evaluate community-based flagging and deletion of unwanted posts, this feature would be important at scale.

Value Offered to Users

Many users attributed great value to their interactions on Sangeet Swara. They recorded strong positive sentiments about the service and shared interesting anecdotes about how Sangeet Swara was impacting their lives. They considered it to be a platform where people show their creativity, voice their opinions, and record interesting content. This sentiment was often strongest among visually impaired users:

> *My mother and father are laborers. You are like my father, my god. I want to thank you again and again, this small kid wants to respect you from the bottom of my heart. I listen to abundant good content on Sangeet Swara. I never got the opportunity to hear such content elsewhere. I am in love with Sangeet Swara since the first day.*
>
> P1 (Male, Student, Visually impaired, Madhya Pradesh)

Visually impaired people used Sangeet Swara to showcase their talent, build social capital, and share information. Some of them considered Sangeet Swara to be a platform *"to learn and understand the principles of life."* A few of them considered it to be a conduit for national peace and integrity, and believed that *"it is proving the mantra of India: Unity in diversity."* One user said:

> *I am blind so I couldn't get educated. I want to thank you from the bottom of my heart because you enabled all blind people to get in touch with each other and show our talent. No matter how much I praise, it won't be enough.*
>
> P2 (Male, Uneducated, Visually impaired, Madhya Pradesh)

Many participants considered Sangeet Swara to be a platform for promoting poor musicians from rural parts of India. They thought of it as a stage for such musicians to showcase and improve their talent, to overcome stage fright, and to step toward India-wide recognition and fame.

> *You can put your hidden talent on the forefront. People don't feel that anyone is listening and thus, they can perform without any hesitation. A performer will feel as if he is alone but a lot of people listen to it later on Sangeet Swara. People with stage fright can present their talent on this wonderful platform. No one can mock you. You will get to meet new people. People in far-off locations will hear you.*
>
> P3 (Male, Teacher, 26 years, Madhya Pradesh)

Many participants appreciated the voice medium, stating that it makes the process of information curation and dissemination much simpler than text-based alternatives. They considered Sangeet Swara to be an inclusive portal for low-literate people, visually impaired people, and tribal people in India.

> *Sangeet Swara is trying to get talent from people in villages and towns. It is a channel for talented people who never got an opportunity to show their talent. Sangeet Swara is trying to get recognition and provide a channel for such people.*
>
> P4 (Male, Musician, 22 years, New Delhi)

Sangeet Swara also helped people build self confidence. Three people reported feeling important when they use Sangeet Swara. Some people felt that Sangeet Swara was also playing a role in improving their grammar, vocabulary and communication skills.

> *I get a lot of knowledge by using Sangeet Swara. Some people record questions, which increases our knowledge. We get to listen to things we have never heard. We learn new vocabulary and sometimes new accents as well. I feel great when people vote for me and give me feedback, be it a good feedback or bad. I consciously think of ways to improve my messages.*
>
> P5 (Male, Student/Farmer, Uttar Pradesh)

3.4 Analysis of Community Moderation

Our goal in this section is to assess the feasibility, acceptability, usability, and efficacy of community moderation in Sangeet Swara. We start with three quantitative approaches: evaluating the accuracy of crowd categorization tasks, comparing the top 50 posts to the bottom 50 posts, and comparing the ranking of posts to an "expert" ranking. We then analyze qualitative feedback from the users.

3.4.1 Categorizing the Posts

Before making a judgement regarding the quality of a post, a basic task one might expect from a moderator is to categorize the content along various dimensions, such as the type of recording,

gender of the contributor, language of the post, etc. Allowing listeners to search or filter content according to these metadata would be an important feature of a scalable voice forum, even though we did not implement such functionality in Sangeet Swara. Especially for most South Asian languages and accents such as Hindi, it is very difficult for current speech recognition technologies to automate or assist with such tagging and categorization tasks.

As described previously, we designated regular users of Sangeet Swara as "senior members" and sometimes asked them to help categorize messages at the beginning of the phone call. Our content analysis found that becoming a "senior member" had a strong positive effect on users. The designation made them feel privileged, honored, and grateful. They felt more accountable for improving content quality, casting votes diligently, and performing tasks:

> *I am now a special person on Sangeet Swara. I have to categorize posts. Please don't use any abusive language on this forum. Please don't say anything wrong because they have made me a senior member and if you do anything wrong then I will tell them.*
>
> (Post on Sangeet Swara)

We asked users to categorize posts along two dimensions: content type and gender. To classify content type, users pressed a key to indicate if the recording was (1) a song, (2) a joke, (3) a poem, or (4) none of the above. To classify gender, users indicated if the speaker was (1) a male voice, (2) a female voice, or (3) they couldn't tell if it was male or female.

Senior members were asked to categorize the top 50 recordings. Whenever senior members called, one of the tasks was randomly presented to them. In total, users completed 3,704 categorization tasks. For each post, we received at least 33 judgements of content type and 40 judgements of gender. The tasks were offered to 291 users, out of which 146 completed the task. The top 20% of workers performed 66% of the tasks.

Task Type	Offered	Done	Response Rate	Accuracy
Content type	1704	1551	91.0%	98%
Gender	2000	1895	94.7%	98%

Table 3.2: Results of categorization tasks done by community.

For each categorization task, we aggregated the responses from the crowd and selected the majority answer as the community response. Before inspecting the community responses, a researcher categorized all the posts by the same criteria. We calculated the crowd's *accuracy* as the fraction of judgements that agreed with the researcher's and the *response rate* as the percentage of users who completed a task when it was offered to them.

Table 3.2 shows the results of categorization tasks done by users. The community showed high accuracy (98% agreement with researcher) on both content and gender classification. For each task type, only one message led to disagreement (a Bollywood hip-hop song for content type, and a muffled voice for gender). The response rate was 95% for gender, and 91% for content type. We speculate that the gender task was easier, leading to more responses.

Many users recorded messages to share their feedback about the tasks. Some users requested more variety of tasks, while others recorded messages critiquing the recordings they categorized. Although the majority of users were excited about helping Sangeet Swara by performing tasks, two people recorded complaints. For example, one of them recorded:

> *I don't want to do any tasks. I just want to listen to the content right away.*
>
> (Post on Sangeet Swara)

	Content Type				Gender			Inaudible		Language		Duration (seconds)		
	Song	Joke	Poem	Misc	Male	Female	Not Sure	Yes	No	Hindi	English	Mean	Median	Mode
Top 50	16	7	23	4	30	20	0	0	50	49	1	48	49	70
Bottom 50	10	0	2	38	46	0	4	4	46	48	2	40	35	70

Table 3.3: Analysis of the top 50 and bottom 50 posts.

3.4.2 Top 50 vs. Bottom 50 Analysis

To examine whether the community applied consistent criteria for desired content, we compared the top 50 messages with the bottom 50 messages (out of a total of about 5,000 messages). If community moderation was successful, the desired content would rise to the top and the poor content would sink to the bottom. We analyzed the posts on several dimensions, including content type, gender, audibility, language, content duration, and the geographic region of caller.

The results of the comparison appear in Table 3.3 (with the exception of geography, which is presented later). We found a significant difference in content type in the top 50 and bottom 50 posts ($\chi^2(3, N = 100) = 53.5, p < 0.0001$). In the top 50 posts, only four posts were in the miscellaneous category as opposed to 38 such posts in the bottom 50. Most of the posts in the bottom 50 were personal messages for another user (N=15), information about other IVR services (N=7), comments on others' posts (N=5), and blank or nonsensical messages (N=4). This demonstrates that the community was successful in promoting songs, poems and jokes to top positions, while pushing messages deviating from the intended usage of Sangeet Swara to bottom positions.

We also found a significant difference in gender between the top 50 and bottom 50 posts ($\chi^2(2, N = 100) = 27.3, p < 0.0001$). Of the top 50 posts, 40% were recorded by females: twenty times the fraction of female recordings in our random sample of content (2%). In contrast, the bottom 50 posts did not contain any recordings by females. This trend corroborates our user and content analyses, in which we found that most users were male and offered special attention, flirting, and adulation to female contributors.

The top 50 and bottom 50 posts did not show significant variations in language, duration, or inaudible posts. However, it is worth noting that the bottom 50 messages contained four inaudible posts while all of the top 50 posts were audible.

We also tabulated the approximate geographical location of callers based on their caller ID[2]. The majority of content authors belonged to similar locations in the top 50 and bottom 50 posts: Rajasthan (N=10, M=14), Madhya Pradesh (N=9, M=14), Uttar Pradesh (N=10, M=7) and Delhi (N=7, M=1).

3.4.3 Community Ranking vs. Researcher Ranking

As an additional validation that community moderation resulted in a meaningful ranking of content, we compared the ranking of messages on Sangeet Swara to a ranking determined by a group of researchers. If these rankings differ, it does not prove that Sangeet Swara rankings are invalid, as the differences could be due to varying tastes of the demographic groups. However, if the rankings agree, it provides additional evidence that the community can perform its own moderation tasks without relying on outside assistance.

In order to compare the judgements of users and researchers, we restricted our attention to songs (the most frequent content type). Restricting attention in this way allowed a more direct comparison of quality, without conflating user preferences for one content type over another. Our experimental design asked researchers to compare a pair of songs, and to see if their preference matched the relative rank of those songs on Sangeet Swara. We prepared 20 pairs of songs from Sangeet Swara. The first ten pairs consisted of one song ranked in the top 20, and one song ranked in the middle 10. The second ten pairs consisted of one song ranked in the top 20, and one song ranked in the bottom 10. We randomized the order of the pairs, and the order of songs within pairs.

For each pair of songs, we asked three researchers (1 male, 2 female, Indian natives, average age = 28 years) to select the one they liked more. Researchers were instructed to focus on the quality

[2]In India, a phone number reveals the geographic region in which a SIM card was purchased, but not the region where it is currently located.

of the singing, doing their best to ignore any variations in language or (if the song is well known) any preference for the original version. Researchers did not know the ranking of songs on Sangeet Swara, and rankings by each other. We used a majority vote to determine the researchers' ranking of a given pair. We compared the researchers' vote with the community ranking to measure agreement between them.

When comparing top-ranked and bottom-ranked posts, 90% of song pairs received the same ranking by researchers and Sangeet Swara users. This amount of agreement is unlikely to happen by chance (a binomial test of 10 trials, each with 50/50 chance of agreement, leading to at least 90% of judgements in either direction, yields $p = 0.02$).

When comparing the top-ranked posts and posts with middle ranking, only 60% of pairs received the same ordering from researchers and Sangeet Swara users. There are several possible interpretations of this result. The top and middle posts were more similar in quality, requiring more subtle distinctions. For example, the agreement among experts was only 75% for this dataset compared to 100% for top-ranked and bottom-ranked posts. As song preferences are highly variable, we may have obtained a higher match if we had used a larger group of researchers. It is also possible that distinctions between these songs were more sensitive to the listener's background or demographic, which differed between our researchers and Sangeet Swara users. While we cannot rule out the possibility that Sangeet Swara users were less careful or less capable to compare the songs, this assertion is not supported by our other observations (such as the high accuracy on categorization tasks).

3.4.4 Qualitative Views of Community Moderation

To understand users' feelings about community moderation, we included a question on this topic in our automated telephone survey. The question asked, "*When you listen to posts, Sangeet Swara tells you the rank of the post. Do you feel that good posts are ranked higher on Sangeet Swara and bad posts are ranked lower?*" We received 126 responses, which were transcribed, translated, and analyzed in two different ways.

The first analysis was a coarse-grained sentiment analysis. The largest category of responses (36%) were neutral or difficult to classify. However, 35% of respondents generally agreed that good content was ranked higher and bad content was ranked lower. A slightly lesser fraction (29%) were not satisfied with the quality of community moderation.

To understand users' views with more nuance, we analyzed the transcripts using open coding and axial coding, arriving at several themes. We found that many people understood that their votes decided the rank and also influenced the playback order. These users emphasized the need to vote honestly:

> *Some messages are really good and their rank is also good. However, around 10% messages aren't good and yet they have a good rank. It is not the fault of the system. The voters should understand which message should be taken to a high level and which to a low level. I would like to tell all listeners that they should listen to messages carefully and then vote honestly. Each vote is precious.*
>
> P6 (Male, Student, 19 years, Uttar Pradesh)

Many people agreed that the quality of community moderation is good and the rank of good quality content is generally higher than the rank of bad quality content. For example:

> *The good songs are higher ranked and the bad songs are lower ranked. I am happy that you decided to rank the posts by our votes.*
>
> P7 (Male, Student, 15 years, Jharkhand)

We found some people who believed that Sangeet Swara administrators decided the rankings. They did not understand how their votes influenced the rank and playback order:

> *Whatever rank the system chooses is right. The good messages have higher rank and the bad messages have lower rank. I trust you that you will never favor anyone. You will categorize the messages properly, give good rank to good messages and will depress bad messages.*
>
> P8 (Male, Telephone operator, Visually impaired, 42 years, Madhya Pradesh)

Some people didn't agree with the ranks assigned to posts, and were unhappy with the quality of community moderation. A few people put the blame on Sangeet Swara administrators for inappropriately assigning the rank:

> *I think you don't listen to the messages. Some messages are very good but have low rank and some messages are useless but they have good rank. Either you are confused or there is some fault somewhere in your system.*
>
> P9 (Male, Student, 19 years, Uttar Pradesh)

Others felt that careless voting by users is responsible for poor quality moderation:

> *Not all the messages appear to have the right rank. I think the reason behind that is voting by the community. I think at many places people do not vote responsibly. They just want to go ahead in the playback list and they don't care whether they are pressing 1 or 2.*
>
> P10 (Male, Government employee, Uttar Pradesh)

Eight respondents demonstrated a lack of understanding between "rank" and "playback order." For example:

> *The rank which is told at Sangeet Swara, I don't understand it. Sometimes the rank is 2511, and the next message is ranked 3303 and then the next to it is 1127. Sometimes it increases and decreases and I don't understand it.*
>
> P11 (Male, Teacher, Jharkhand)

Although the distinction between rank and playback order is necessary to ensure fair playback and voting policies, there could be better and simpler ways to communicate the rankings to users. For example, instead of reading the numeric rank, a prompt could say "this is a new post, and we really need your opinion", "this post is an old favorite", "this post is liked by some people, but more input is needed", and so on.

To summarize, although community moderation demonstrated effectiveness from a quantitative standpoint, our qualitative analysis reveals that there is room to improve on how the system is understood and appreciated by users.

3.5 Evaluation of Financial Sustainability

To create a voice forum that can scale and sustain without outside assistance, moderating the content is only part of the equation. The other challenge is financial sustainability. In particular, there needs to be a way to support the cost of phone calls as the usage of the system grows.

Given how deeply many users seemed to value Sangeet Swara, we thought that a subset of users may be willing to pay for their own phone calls, thereby sustaining the system without external funding. Thus, after spending about USD 3,000 on seven weeks of toll-free support, we planned a switch to a regular line (which costed users the same as a normal phone call – 1 to 2 cents per minute, based on their mobile plan). This required users to call a different phone number, which we announced as part of the welcome message for the five days preceding the change. We are confident that users understood the change in number, because the forum was immediately inundated with emotional

requests to continue the toll-free service. For example:

> *I am very sad. Please, don't change the number. I fold my hands and request. Please consider my request. Not only me, everyone wants it to be toll-free. If it is a paid number then people won't be able to use it. Please don't reject our plea. Sorry sir. I fold my hands and pray, please don't change the number. Please cancel the announcement.*
>
> P12 (Male, Student, Visually impaired, 19 years, Uttrakhand)

Unfortunately, this student's prediction was correct. As illustrated in Figure 3.2, the usage steeply declined without the toll-free lines. Within four weeks, it died out completely. We will say more about this result as part of our closing discussion.

3.6 Follow-up Experiment: Talent Hunt

Although users of Sangeet Swara derived significant value and meaning from the system, they were unable to pay for their own phone calls, which limited the scale we could achieve. As a follow-up experiment, we wanted to see if similar value could be delivered to a slightly higher-income group that might be able to afford to make phone calls without toll-free lines. If successful, such an experiment could grow into a very large ecosystem, giving more opportunities for monetization and cross-subsidization of low-income users.

To explore this idea, we adapted Sangeet Swara to deploy Talent Hunt: a voice forum of songs, poetry, jokes, and other cultural content targeting college students in urban India. The infrastructure and call flow for Talent Hunt was the same as Sangeet Swara, though prompts were recorded in English (arguably the most common language for college students in India) instead of Hindi. The biggest difference was in the incentives offered to users of the system. Instead of using toll-free lines, we promoted participation by awarding a smartphone (Nokia Lumia 710) to the authors of top-ranked posts. We made one award per week for the first six weeks. After six weeks, we still announced one

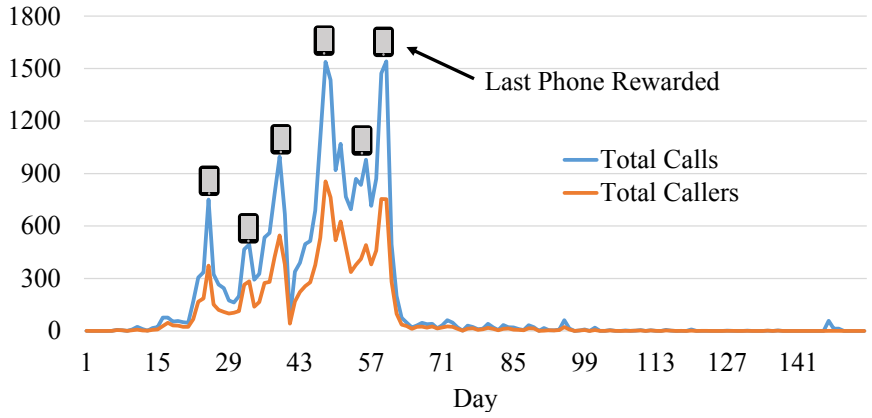

Figure 3.3: Call statistics for Talent Hunt.

winner per week (and featured winners in the main menu), but did not award any phones. Our hope was that the material prizes would be sufficient to seed interest in the forum; future participation would be sustained by social recognition and attachment to the community. We promoted Talent Hunt using posters (in college campuses), email, social media, and targeted outreach by student volunteers.

Figure 3.3 shows the usage of Talent Hunt over time and Table 3.4 compares the usage of Talent Hunt with Sangeet Swara. Although Talent Hunt received calls from 11,751 people (7 times more than Sangeet Swara), unfortunately this usage was driven entirely by the material awards. As soon as the last phone was awarded, participation dropped to zero. We ended up spending more on the phones (USD 1,517) than the users collectively spent on the airtime (USD 1,305[3]).

The award-based incentive structure also had deleterious effects on the quality of the voice forum. Despite the large number of callers, only 368 posts were recorded, which is 15 times less than on Sangeet Swara. Rather than building a supportive community of participation and sharing, users of Talent Hunt were often calling only to vote for a friend. Among ten of the top-rated posts, 99.5% of votes obtained were direct votes, in which the voter jumped directly to the post by entering its ID

[3] Assuming a call rate of USD 0.02/minute.

	Sangeet Swara	Talent Hunt
Deployment Duration	11 weeks	22 weeks
Language of Prompts	Hindi	English
Calls	25,381	27,514
Callers	1,521	11,751
Posts	5,376	368
Contributors	516	304
Average Call Duration	4.96 mins	2.37 mins
Total Plays of Posts	198,898	42,383
Upvotes	40,590	10,832
Downvotes	99,150	3,375
Share Requests	773	251
Direct Jumps to Post ID	7,871	11,868

Table 3.4: Usage statistics for Sangeet Swara and Talent Hunt.

from the main menu: evidence that most callers were trying to support someone they knew in real life.

We also conducted semi-structured qualitative interviews with ten users who were among the top vote recipients. These users indicated that they strategically mobilized large groups of people to vote on their posts. Some of them made announcements in classrooms, on social media sites, and even at a wedding in order to gather more votes.

The quality and quantity of user engagement was also much lower than Sangeet Swara. Out of 11,751 Talent Hunt users, 78% users called at most twice. By the standards established in Sangeet Swara, no member of Talent Hunt would have been designated a senior member. As most of the users were college students in tier 1 and tier 2 cities, they also had access to social media platforms such as Facebook, WhatsApp, Twitter, etc. As a result, Talent Hunt offered less value to them.

To summarize, our experience with Talent Hunt showed that offering financial awards for top-ranked recordings brought several hazards to a community-moderated voice forum. Patterns of participation and voting became grossly distorted by users who sought to help their friends. Moreover, the incentive failed to seed long-term participation. Although users were able to pay for their own calls during the contest period, the usage expired as soon as the contest was over.

3.7 Discussion and Conclusion

This chapter describes lessons learned from Sangeet Swara, a community-moderated voice forum in rural India. The system evoked a passionate response from users, particularly those with visual impairments, who discovered and appropriated the platform without any outreach on our part. We believe that the ability to be an equal participant (and moderator) of a voice forum was a uniquely empowering experience for rural residents, tribal people, and visually impaired communities, who are often marginalized by their societies. Our study shows that a community of untrained callers, most of whom without any experience of Internet-based services, can accurately perform their own moderation tasks including categorizing and rating posts, thereby mitigating the bottleneck of a dedicated moderation team.

Although our work demonstrates the feasibility, acceptability, usability, and efficacy of community moderation, several aspects of it remain untested. For example, Sangeet Swara focused on the domain of entertainment, where the content is relatively uncontroversial. Extending to domains such as politics and citizen journalism will require sensitivity to stronger disagreements between callers, which could impact their ratings as well as their flagging of posts for deletion. Similarly, the community moderation algorithm can be improved by making it more sensitive to *who* is voting and *when* they are voting. For example, discounting votes of users who acted too soon or those who abused in prior posts could reduce randomness in moderation. Similarly, assigning higher weights to votes of users who call consistently and record high-quality posts could improve the quality of moderation. Future work should identify features to predict abusive behavior or casual voting by community

members. We also found that an overwhelming majority (94%) of Sangeet Swara users were male. It will be important to understand how to build a community that is inclusive and inviting towards women. We explore this subject in more detail in Chapter 6.

For community-moderated voice forums to scale further, they also require financial sustainability, which was not achieved by either Sangeet Swara or Talent Hunt. Given that Sangeet Swara users were passionate about using the system, their reluctance to pay for the phone calls is almost akin to an "impossibility proof": for users in this demographic, it is very difficult for a voice forum to deliver sufficient benefits for users to consider paying for the calls themselves. Conversely, in the case of Talent Hunt, we believe that users were able to pay, but had lesser interest in the service, since they had access to online social media platforms. More work is needed to explore the types of voice forums that can bring value to people who can afford the voice call costs.

To reduce costs in the future, one promising approach is to transfer audio content via mobile data connections (as they become readily available and more affordable) instead of voice calls [75]. It may also be possible for callers to perform more general audio micro-tasks, similar to Mechanical Turk but administered over IVR. The revenue generated could help offset the costs of calls. We explore this approach to financial sustainability in more detail in Chapter 4.

There are rich opportunities to broaden the scope of voice forums. Users of Sangeet Swara desired additional interactions, such as sending personal messages and listening to all posts by a given person. Generalizing a voice forum in this way could lead to a flexible social networking platform over IVR, leading to even greater uptake and engagement by rural users.

Chapter 4

CROWD WORK FOR FINANCIALLY SUSTAINING VOICE FORUMS

Voice forums have emerged as an alternative to the Internet in low-resource environments with poor connectivity, low literacy, and poverty. They have proven themselves as a usable and accessible communication medium that can be effectively used to meet the information and instrumental needs of marginalized people in low-resource settings. For example, Chapter 3 demonstrates how Sangeet Swara (a social media voice forum) connected people in low-resource environments and brought digital equity to marginalized communities, including rural residents and people with visual impairments.

However, a key bottleneck in scaling and sustaining these impactful services is the high cost of voice calls that service providers have to pay to make these services free for low-income callers. While a few services sustain themselves through advertising [34], external grants [30], and partnerships with mobile network operators (MNOs) or governments [27], these alternatives are often beyond the reach of most voice forums.

In the previous chapter, we explored whether low-income users can pay for voice call costs themselves if they derive significant value from a voice forum. Unsurprisingly, we found that most users could not afford the cost of voice calls, primarily because they were struggling to meet their basic needs. We also found that voice forums generally offer limited value to people who can afford to pay for voice calls, limiting the possibilities of cross-subsidizing participation of low-income users from profits from paying users.

This chapter examines whether voice forum users—generally low-income students, visually im-

paired people, and rural residents—could complete useful work on their mobile phones to offset the participation costs of voice forums. Since existing crowdsourcing marketplaces such as Amazon (MTurk) [17] and CrowdFlower [21] are unfeasible in low-resource environments, we design and build a new crowdsourcing marketplace that leverages familiarity with the local language, the power of voice, and the ubiquity of basic phones to circumvent literacy, language, and connectivity barriers.

To create this crowdsourcing marketplace, we focus our attention on speech transcription, a thriving industry where transcription work can be divided into small, manageable, and meaningful units. Speech transcription–including general, medical and legal transcription–fuels a massive industry; medical transcription alone is nearly worth USD 60 billion globally [15]. Transcription of recorded audio is demanded for a wide variety of content, including public speeches, movies, songs, television programs, advertisements, news, interviews, recorded lectures, online videos, and telephone calls. The demand of speech transcription is rapidly rising for languages and accents popular in developing regions. However, existing services produce the transcription of audio files containing low-resource languages and accents with poor accuracy and at high cost.

To enable low-income voice forum users with literacy, language, and connectivity barriers to complete useful work on their mobile phones, we designed and built Respeak: a voice-based, crowd-powered speech transcription system that pays users to transcribe audio files vocally. This chapter presents the design, deployment, and evaluation of Respeak. The Respeak system has two components: the engine and the user application (app). To transcribe an audio file, the engine segments the audio file into short utterances that are easier for users to remember. It then sends small audio segments to multiple users and pays them via mobile airtime when they submit transcripts by using the app. Instead of typing the transcript on a phone's keyboard with constrained physical space, users re-speak (i.e., repeat) audio content into an off-the-shelf speech recognition engine and submit the speech recognition output as a transcript. Once multiple users submit transcripts for a particular segment, Respeak combines the transcripts using sequence alignment algorithms to reduce random speech recognition errors.

While designing a new system, a large number of design parameters need to be investigated in a systematic manner. To design Respeak, we examined how audio files should be partitioned, what should be the length of segments, and how these segments should be presented to make it easier for users to complete transcription tasks. We also investigated how phone types, channel types, and modes to review transcripts affect task accuracy and completion time.

To examine the feasibility, acceptability, and usability of our approach in addressing the financial sustainability challenge, we iteratively built and deployed three user apps:

- **Respeak smartphone app**: To examine the feasibility of a system where users perform tasks vocally, we first built the engine and a smartphone-based user app. We then deployed the app to 28 low-income students—arguably the most tech-savvy user group among voice forum users—for a month to examine its strength and weaknesses [166].

- **BSpeak smartphone app**: Since voice forums are very popular among people with visual impairments [76, 140, 161], we examined whether the app is usable to low-income visually impaired people. To do so, we built an accessible version of the smartphone app and deployed it to 24 low-income blind people for two weeks [167].

- **ReCall IVR app**: After carefully investigating the deployments with low-income students and blind people, we adapted the smartphone app to build an IVR-based app—our end goal—and deployed it to 28 rural residents in low-income environments. Finally, to examine whether voice forum users can vocally transcribe audio files to subsidize their call costs to voice forums, we integrated the ReCall app into Sangeet Swara where users could do tasks on ReCall to get free airtime to use Sangeet Swara [163].

Our findings demonstrate that low-income students, blind people, and rural residents can transcribe audio files vocally to produce high-accuracy speech transcription at a cost lower than the industry standard. During the deployments, 73 low-income people transcribed 70 hours of audio data by

completing 50,000 micro tasks with an average accuracy of 70% and earned ₹31,000 at an hourly rate that exceeds the average hourly wage in India. The engine merged transcripts from multiple users to produce speech transcription with over 90% accuracy at nearly one-fourth of the market rate, generating sufficient profit to subsidize participation costs of other voice-based services.

We found that low-income rural residents could complete useful work on their mobile phones by using ReCall and generate enough profits to subsidize their costs to use Sangeet Swara. Our analysis indicated that each minute spent in completing crowd work on ReCall could provide about eight minutes of free airtime on voice forums. Also, switching between these two services—ReCall to complete crowd work and Sangeet Swara to use free credits—did not affect the usability and user experience of participants on both services.

Our work makes two significant contributions to HCI4D research. First, it demonstrates the feasibility, usability, and acceptability of a crowdsourcing marketplace that is accessible via ordinary phone calls from even the most basic phone. ReCall is the first crowdsourcing marketplace deployed to low-income rural residents where users earn money by vocally transcribing audio segments on phone calls. Second, our work addresses the financial sustainability challenge of voice forums by allowing user-earned profits from crowd work to provide free airtime on voice forums.

We describe the related work on speech transcription solutions and crowdsourcing in low-resource environments in the next section. We then present the design of Respeak, BSpeak, and ReCall, and describe the cognitive experiments and other evaluations we conducted to gain key design insights. We then report the deployment details. Finally, we discuss the lessons learned from these deployments and suggestions to integrate ReCall to financially sustain large-scale voice forums.

4.1 Background and Related Work

There is a large body of research examining approaches to improve speech transcription. Similarly, there is a growing interest within the HCI4D community to build and evaluate systems to provide

additional earning opportunities to people in low-resource environments. We center our discussion of related work on existing speech transcription solutions and crowdsourcing marketplaces in low-resource environments. We also discuss how our work contributes to research at the intersection of crowdsourcing and accessibility.

4.1.1 Speech Transcription Solutions

Manual transcription, while efficient, is an expensive process with a high turnaround time. Manual transcribers are trained to type faster, understand different accents, tune out ambient noise, and differentiate speakers, making manual transcription a specialized and expensive service. The cost of manual transcription service varies from USD 1–4 per minute based on several parameters, including the language, quality of speech, audio length, ambient noise, number of speakers and their accent, requested turnaround time, and verbatim versus non-verbatim transcription.

The advent of crowdsourcing has greatly impacted speech transcription industry. Several online transcription services—like SpeechPad [46], CastingWords [40], TranscribeMe [48], Rev [44], CrowdSurf [42], and Tigerfish [47]—use manual transcription from a crowdsourced labor. However, most of these services support only popular accents of English, excluding local languages and dialects spoken in developing countries. Also, their cost varies from USD 1–6 per minute based on several parameters noted earlier. Workers also transcribe files that can be up to several hours long, making transcription a high cognitive load exercise. Although several online non-crowdsourcing portals [43, 45, 49] transcribe content in languages spoken in developing regions (like Hindi, Marathi, Urdu and Indian English) using a fleet of transcribers, the cost of transcription averages at nearly USD 5 per minute.

Many researchers have used automatic speech recognition (ASR) and crowdsourcing for speech transcription [101, 152, 174]. Several others have used crowdsourcing to improve aspects of ASR like expanding language corpora and identifying prosody annotations [71, 79, 102, 104, 113, 114]. Most relevant to us is work by Parent and Eskenazi [128] and Lee and Glass [105]. They used a

two-stage, crowd-powered speech transcription process, where audio files were broken into short segments to reduce cognitive load on workers. While Lee and Glass requested workers to type transcripts for short audio segments, Parent and Eskenazi asked workers to correct ASR generated transcripts. Similarly, Lasecki et al. [103] designed a real-time captioning system where non-expert crowd workers transcribed overlapping segments of audio by typing; these segments were merged in real-time by using multiple string alignment and majority voting [122]. Respeak draws on existing research by using a two-stage process that segments a large audio file into smaller utterances and then merge generated transcripts using multiple string alignment and majority voting. However, unlike other systems, Respeak users speak the content into a standard built-in speech recognition engine rather than typing it. Using speaking skills rather than typing skills makes Respeak easy and natural to use, especially for people with no or low typing skills.

Prior research exploring re-speaking [89, 134, 154] requires significant data to generate speaker dependent acoustic models and domain dependent language models, making these solutions expensive and untenable at scale. Respeak, on the other hand, uses an off-the-shelf generic ASR system and combines transcripts generated by multiple users to reduce ASR errors. Rather than relying on high-skilled re-speakers that have undergone an intensive training of several months and capable of handling multiple hours of captioning without break in a controlled environment [134], Respeak rely on multiple unskilled crowd workers to perform micro re-speaking tasks in their everyday environment. Lastly, Respeak provides transcription for low-resource languages and accents that yield much lower ASR accuracy than the well-represented languages and accents, such as English and Japanese, used in prior works.

4.1.2 Crowdsourcing Marketplaces in Low-Resource Environments

Crowd-powered online transcription services, such as CastingWords [40] and TranscribeMe [48], provide additional earning opportunities to low-income people. However, several inclusion criteria—such as a minimum typing speed of 40 words per minute (WPM) [46], an active PayPal

account connected to a banking institution [40, 44, 46, 47, 48], and access to an Internet-connected computer [40, 44, 46, 47, 48]—makes it difficult for many in developing regions to use these platforms. Mainstream crowdsourcing marketplaces such as MTurk and CrowdFlower also require access to the Internet, computers, and English language skills, making them unfeasible for people in low-resource environments. In addition, these platforms have many usability and accessibility barriers, making them unusable for people with low literacy and people with visual impairments. For example, Khanna et al. found that low-literate people in India struggled to navigate MTurk's user interface and understand task instructions [96]. Similarly, Zyskowski et al. found that several accessibility barriers, including the inability to create an account because of CAPTCHA, poor ratings due to unfinished inaccessible tasks, and a lack of a filter to select accessible tasks, limited the earning potential of MTurk workers with visual impairments [176].

To overcome many of these literacy, language, and connectivity barriers, several HCI4D researchers have designed new crowdsourcing marketplaces to supplement income of marginalized communities in low-resource environments. For example, txtEagle [77] and mSurvey [24] send SMS-based survey to low-income people and pay them for each completed survey. Similarly, mClerk [88] and MobileWorks [100, 123] pay low-income people to transcribe images sent to their phones containing Kannada and English words, respectively. While mClerk uses SMS for sending images and receiving responses, MobileWorks uses a mobile-based web application.

Many other portals rely on the availability of Internet-connected phones or computers. For example, Samasource [9] engages in impact sourcing and establishes outsourcing centers in low-resource regions where people living in poverty are trained in image annotation and other services. Karya pays cash to rural residents in India to digitize hand-written Devanagari script documents sent to their smartphones [70]. mCent partners with MNOs to provide free Internet and airtime credits to its users for web browsing through its browser [7].

A major limitation of all these systems is that they expect crowd workers to have reading and typing skills. As Table 4.1 shows, our work overcomes this limitation by using speaking and listening skills

	Text	Voice
Internet-connected computer	MTurk CrowdFlower Samasource	–
Internet-connected phone	MobileWorks mCent Karya	Respeak BSpeak
Any mobile phone	mClerk txtEagle mSurvey	ReCall

Table 4.1: Categorization of microtasking platforms based on requirements and modailty of performing tasks.

of users for crowdwork. Although Respeak and BSpeak let users earn money by transcribing audio files vocally, they require users to have access to a smartphone—like Karya and mCent—making them unfeasible for people with basic or feature phones. ReCall overcomes this limitation by enabling users to perform speech transcription tasks vocally via ordinary phone calls from any types of phones.

4.1.3 Accessibility and Crowdsourcing

There is a large body of research at the intersection of accessibility and crowdsourcing. Most systems are designed to meet information needs of people with visual impairments. For example, VizWiz enables blind people to ask visual questions to their social media friends or MTurk workers in real time [60, 64]. Visual Answers, an extension of VizWiz, enables blind people to ask visual questions from Facebook friends of volunteers [63]. RegionSpeak, another extension of VizWiz, helps blind people receive labels from MTurk workers for all relevant objects contained in a stitched image [175].

Similarly, Chrous:View helps blind people understand their environment by facilitating a real-time conversation between a blind user and crowd workers about a video stream from the user's phone. Be My Eyes, a similar smartphone app with over two million users in 150 countries, establishes a live video connection to let blind users ask questions from sighted volunteers [19]. Note that in all these systems—VizWiz, Visual Answers, RegionSpeak, Chorus:View—blind people are *consumers* of crowdwork [176]. In contrast, BSpeak allows blind people to be *producers* of crowdwork; they transcribe audio files vocally anytime anywhere to supplement their income.

We now present the design and implementation details of the Respeak engine as well as the Respeak, BSpeak, and ReCall user apps.

4.2 Respeak Engine and User Applications

Respeak enables users to vocally transcribe audio files by using a five-step process, as shown in Figure 4.1. The Respeak system has two main components: the Respeak engine and a user app.

4.2.1 Respeak Engine

The Respeak engine engages in four main activities to transcribe an audio file.

- **Segmentation:** It segments a large audio file for transcription into short utterances that are easier for the app users to remember. The segmentation process uses the occurrence of natural pauses to yield utterances that are typically 3–6 seconds long.
- **Distribution to Crowd Workers:** It distributes these segments to multiple users who produce transcripts vocally by using the app.
- **First-stage Merging:** For each segment, it combines multiple users' output transcripts into one best estimation transcript using multiple string alignment (MSA) and majority voting.

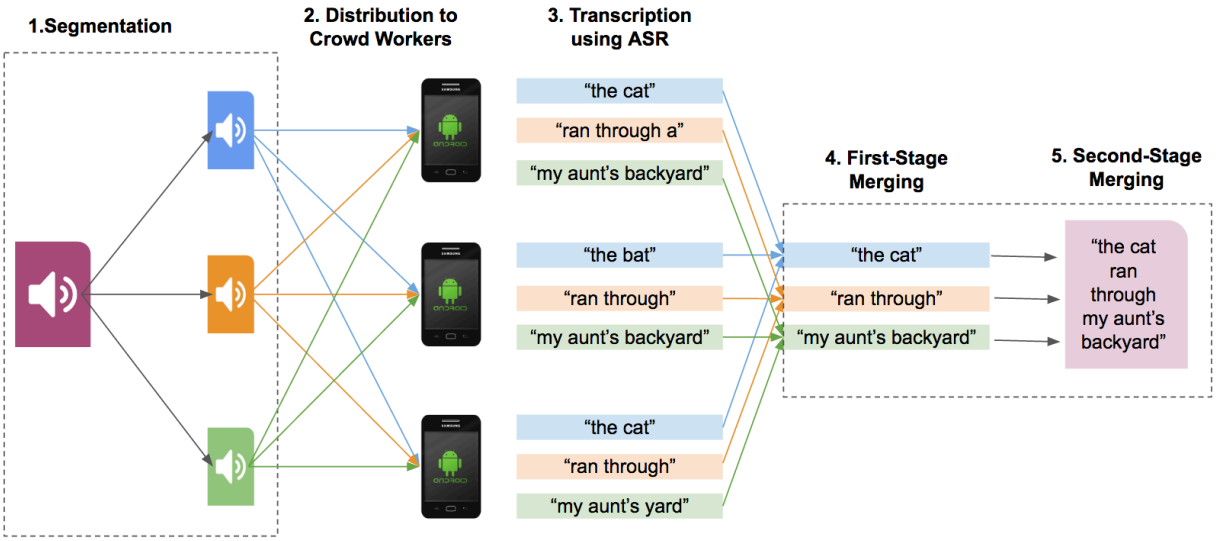

Figure 4.1: A high-level illustration of Respeak's design. Areas inside dotted lines represent the processes of the engine.

The transcript sent by the user is compared to the best estimation transcript to determine the reward. Once the cumulative reward amount earned by a user reaches ₹10, it sends mobile airtime credit or mobile payment of equivalent value to the user. We discuss the workings of first-stage merging process in more detail below.

- **Second-stage Merging:** It concatenates all best estimation transcripts from first-stage merging into one large file to yield the final transcription.

Speech transcription efficiency of the Respeak engine depends on the performance of the multiple string alignment (MSA) algorithm, and the majority voting process that is the core of first-stage merging. For each segment, the Respeak engine combines the transcripts submitted by users to produce a best estimation of actual word sequence in the segment. The MSA algorithm in our implementation uses word as the individual atomic unit rather than character or phoneme. We

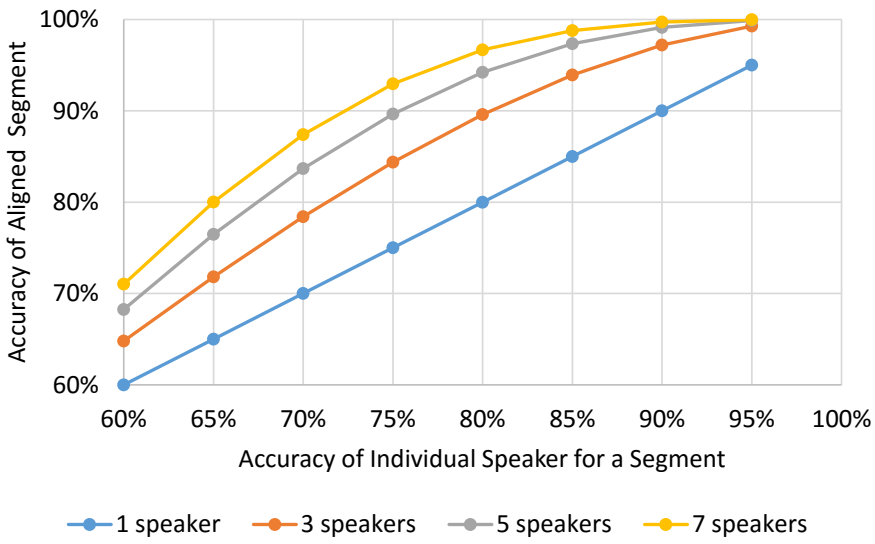

Figure 4.2: Improvement in accuracy by using MSA and majority voting.

adapted and implemented the multiple sequence alignment algorithm proposed by Naim et al. [122] that uses the A* search algorithm to reduce the search space of multi-dimensional lattice. Any ties during majority voting are broken randomly. Let us assume that first-stage merging takes as input transcripts generated by K users. If the ASR errors are uncorrelated across users, then the word error rate (WER) of the hypothesized word sequence should decrease as K increases. Let P be the average accuracy of speech recognition for individual users. The WER then is $1-P$. Assuming that the errors are randomly distributed across users, the accuracy of the alignment of segments (P_{final}) for N users computed using majority voting is:

$$1 - \binom{N}{N}(1-P)^N - \binom{N}{N-1}(1-P)^{N-1}P..... - \binom{N}{K}(1-P)^K P^{N-K}, K \geq N/2 \tag{4.1}$$

Figure 4.2 depicts the estimated improvement in accuracy achievable by aligning the transcripts generated by one, three, five and seven users for several values of P.

4.2.2 Respeak Smartphone Application

To transcribe an audio segment sent by the engine, the Respeak smartphone app users listen to the segment and repeat it into the app in a quiet environment. The app uses the built-in Android ASR engine to obtain a transcript for the spoken segment and displays it to the user. The transcript thus produced is expected to have a high WER. The user submits the transcript for the current segment and then receives a new segment that requires transcription. The user could also check the task completion accuracy, amount earned, and more details on how payments are processed.

4.2.3 BSpeak Smartphone Application

The BSpeak smartphone app is designed for people with visual impairments. It is the accessible version of the Respeak smartphone app and it uses the same underlying engine for segmentation, distribution, and merging processes. Blind users navigate the BSpeak app by using TalkBack—Android's built-in screen reader software—that reads aloud screen content on touch and swipe gestures. As illustrated in Figure 4.3, we changed several components of the Respeak app, by using Android's accessibility guidelines [16], to create the BSpeak app.

- **Labeling UI elements:** We labeled all UI elements, such as buttons and images, with appropriate descriptions to enable TalkBack to read them aloud. Without such labels, TalkBack would make generic announcements such as "image button" or "text area" which provides no information about a UI element's actual functionality.

- **Large touch targets:** Since it is difficult for blind users to navigate small touch targets on a phone's screen, we minimized empty space, increased button and font sizes, and made the touchable areas more than 48dp x 48dp.

- **Explicit instructions:** We modified verbal instructions that complement visual cues accessible only to sighted users. For example, the instructions for recording audio on the Respeak

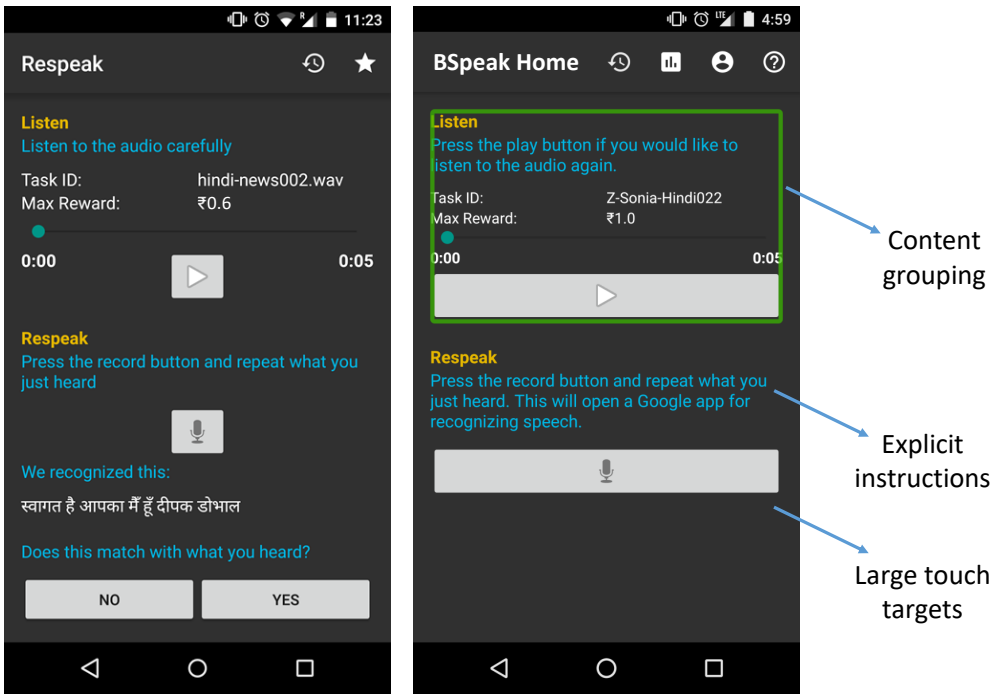

Figure 4.3: A screenshot of the Respeak app home (left) and the BSpeak app home (right).

home screen were *"Press the record button and repeat what you just heard."* While sighted users could see a separate Google app opening up for speech recognition, blind users needed to be explicitly told to expect it. Thus, in BSpeak, we changed the verbal instruction to *"Press the record button and repeat what you just heard. This will open a Google app for recognizing speech."*

- **Content grouping:** We grouped multiple UI elements into single announcements to treat them as one focusable container. Thus, as the user presses any single element within the group, the entire content of the container is announced out loud by TalkBack, which makes it easy to access logically related elements all-at-once. Without grouping, a blind user would have to individually touch text labels or swipe many times to read the elements on the screen.

We also added new features to BSpeak such as access to help page and the option to select the lan-

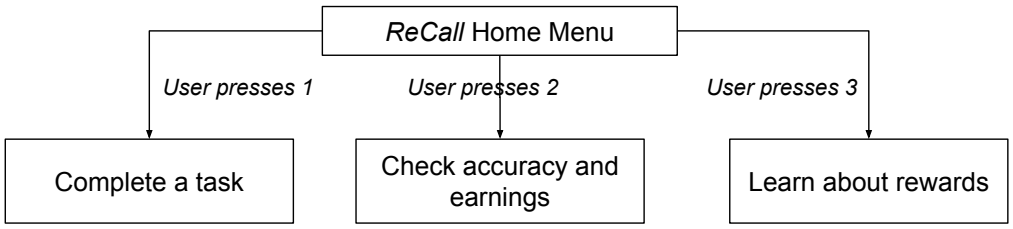

Figure 4.4: High-level call flow of the ReCall app.

guage of audio tasks. Another key difference between Respeak and BSpeak is how users accessed the output transcript generated by the ASR system. While Respeak users could read the transcript displayed on the screen, TalkBack read it aloud for BSpeak users.

4.2.4 ReCall IVR Application

The ReCall IVR app is designed for rural residents without access to smartphones. Functionally, it is similar to the Respeak smartphone app, and uses the same underlying engine for segmentation, distribution, and merging processes. To transcribe segments, users call the phone number associated with the ReCall IVR app. Figure 4.4 illustrates the high-level call flow. Once the call is connected, users select one of the three options by pressing the relevant key on their phone keypad.

- **Complete a task:** A prompt announces the task reward, and requests users to listen to an audio segment carefully and re-speak it into the app in a quiet environment. Once users re-speak the content, the ReCall app submits the re-spoken audio file to an off-the-shelf ASR engine and sends the ASR-generated transcript to a text-to-speech (TTS) engine. The audio transcript generated by the TTS engine is played back to the users. If the audio transcript is similar to the audio segment, the users presses 1 to submit the transcript for the current segment and receives a new task. The transcript is expected to have some errors since users may not fully understand the segment or TTS output, may make a mistake while re-speaking content, or the ASR engine could incorrectly recognize some words.

	ReCall	**Respeak**
Phone type	Any phone	Smartphone
Application type	IVR app	Android app
Channel used	Voice	Data
Audio quality	8kHz	44kHz
Review mode	Listening	Reading

Table 4.2: Key differences between ReCall and Respeak

- **Check accuracy and earnings:** A prompt announces the average accuracy with which the caller has completed prior tasks and the total amount they earned.

- **Learn about rewards:** A prompt explains how ReCall calculates users' earnings when they complete tasks.

We followed best practices outlined in the literature [69, 73, 131, 161] to make the ReCall app usable for low-income rural residents. For example, prompts were recorded in the local language and accent, and had clear pronunciation, colloquial diction, and proper explanations. Similarly, all key presses were single digit inputs and invalid key presses yielded informative error messages.

Although ReCall and Respeak apps use the same underlying engine, Table 4.2 outlines how ReCall and Respeak differ fundamentally in several ways. For example, Respeak users need a smartphone to complete crowd work, whereas ReCall is an IVR app accessible via ordinary phone calls from any phone. While the Respeak smartphone app download tasks on a data channel preserving the 44kHz sampling frequency of the segments, the ReCall app uses the voice channel that degrades the quality of tasks and re-spoken audio segments to 8kHz sampling frequency, making them harder for users to listen carefully and ASR engine to recognize. Similarly, Respeak users review ASR-generated transcripts by reading them. In contrast, ReCall users review tasks by listening to transcripts in a synthetic voice of TTS system, making it difficult for them to catch errors. The two systems also

differ in demographic of their target users. While ReCall is designed for low-income rural residents, Respeak was deployed to low-income students living in a metropolis.

In the following section, we discuss the cognitive experiments we conducted to understand key questions that affect the user interface design of Respeak system.

4.3 Cognitive Experiments for Interface Design

We considered several issues when designing the Respeak interface. One key issue pertains to the process of partitioning large audio file into small segments that are easier to retain and re-speak. A simple algorithm could segment a file based on the occurrence of natural pauses in speech. Because such pauses are natural transition points, the segments so obtained might be easier to remember. However, these segments could be long, making them difficult to retain for re-speaking. Moreover, detecting natural pauses in audio files with high ambient noise or music is difficult. Another segmentation approach could split the file into short, fixed-length segments. Though shorter segments would be easier to retain, their abrupt beginnings and endings could impose a high cognitive load for retention. Another main design issue involves identifying how segment length and order of micro-task presentation affects retention. Finally, evaluating the benefits and limitations of producing transcripts by re-speaking versus typing significantly affects design choices. Thus, we conducted three cognitive experiments to evaluate:

1. How audio segment length affects content retention and cognitive load experienced by users?

2. How segment presentation order (sequential vs. random) affects content retention and cognitive load?

3. Whether speaking is indeed more efficient and usable output medium for transcription than typing?

4.3.1 Methodology for Cognitive Experiments

We conducted a within-subjects design study to evaluate the first experiment. We randomly selected 14 audio segments from a televised English news broadcast in India. Two segments each were selected with a length of 1–7 seconds. The average speaking rate in the segments was 160 WPM. Participants performed 14 tasks; in each task, they played a randomly selected segment multiple times on a laptop and re-spoke the content once they memorized it.

We conducted a between-subjects design study to evaluate the second experiment. We randomly selected a one-minute segment from a televised Indian English news broadcast with a speaking rate of 137 WPM. We used a fixed-length segmentation scheme to obtain 15 segments, each of which was four-second long. Participants were randomly partitioned into two groups. The first group listened to the segments in a random order, while the second listened to the segments sequentially. Participants performed 15 tasks, one for each segment. They played the selected segment multiple times and then re-spoke the content once they memorized it.

We conducted a within-subjects study to evaluate the third experiment. We randomly selected a 100-word English news article from a newspaper in India. Participants had to do three tasks: type the article on their computer, type the article on their phone, and read the article out loud. We chose a written article than recorded material since we believed that listening and then typing/re-speaking would also test retention skills in addition to typing/speaking skills. We randomized and balanced the order in which participants completed the tasks.

We recorded and manually transcribed the content re-spoken by participants for each task in all experiments. We measured the WER of re-spoken content and the task completion time. For the first and second experiment, we also measured the number of times participants listened to the segment. We conducted semi-structured interviews after participants finished tasks in all three experiments. The interviews were recorded, transcribed, and analyzed using open coding. These evaluations were approved by our institution's IRB.

	Very good	Good	Average	Bad	Very bad
English speaking	4	14	6	0	0
English typing	4	13	7	0	0
Hindi speaking	4	11	9	0	0
Hindi typing	0	0	5	3	16

Table 4.3: Self-assessment of participants' language skills.

4.3.2 Cognitive Experiments Participants' Demographics

We used a campus-wide email list from a university in India (IIT Bombay) to invite participation and randomly selected 24 respondents. Seventeen participants were male, and seven were female. The average age of participants was 24.4 years. Eight participants were summer interns at the university; five were hired as project staff; four were pursuing a bachelor's, four a master's, and three a Ph.D. degree. Twenty participants were from the engineering disciplines and four were from the humanities. All but one participants owned a smartphone with Internet access. The average daily phone and computer usage was reported to be around 5.5 hours and 10 hours, respectively. Nine participants knew about crowdsourcing platforms, but only two had used them previously. As Table 4.3 shows, the majority of participants assessed their Hindi typing skills as being very bad.

4.3.3 Findings of Cognitive Experiments

While WER predicts the performance of content retention, task completion time and number of listens predict the cognitive load experienced by participants.

Experiment 1: Impact of Segment Length on Retention

Figure 4.5 compares the WER, time taken to retain and re-speak segments, and number of times segments were listened in the first experiment. A repeated measures ANOVA with a Greenhouse-

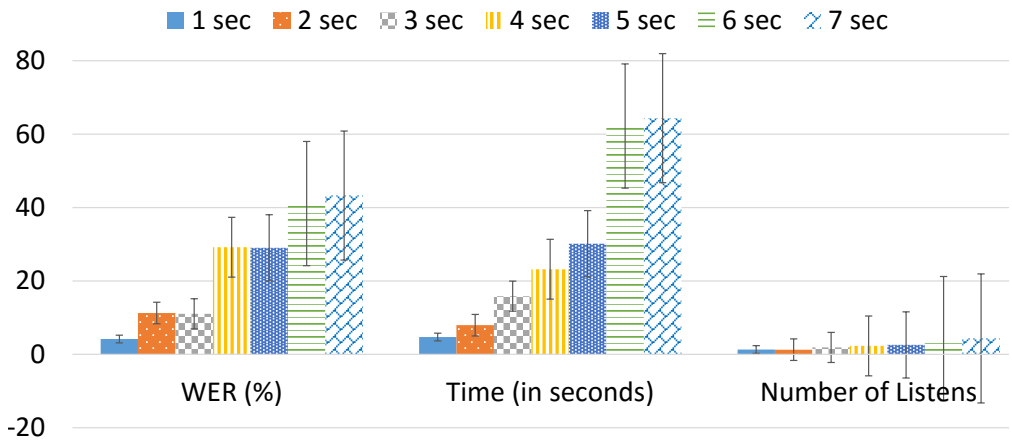

Figure 4.5: Comparison of varying length segments on several parameters.

Geisser correction determined a statistically significant difference (at $p<.001$) in the three parameters for the segments that were 1–7 seconds long. Table 4.4 highlights the parameters that significantly differ (at $p<.05$) on a pairwise comparison of the segments of different lengths. The WER and time taken were much higher for segments exceeding five seconds. Our interviews also revealed that several participants could not retain such segments because of complicated sentence constructions and the excessive number of concepts to remember. Thus, using such segments in Respeak could result in a poor accuracy speech transcription and put significant cognitive load on users. Eleven participants used synonyms or missed articles while re-speaking content. Three participants found it difficult to retain segments containing an incoherent word. Another three participants found it challenging to retain unfamiliar proper nouns. Four participants found content retention to depend on their familiarity with subject matter rather than on duration. One of them stated:

> *If you present segments on cricket to a cricket enthusiast, he will easily remember the content irrespective of how long it is. But if the same person has to remember content related to military strategies, they may not remember it.*

Twelve participants found it difficult to retain segments containing partial sentences. Abrupt cuts

	1s	**2s**	**3s**	**4s**	**5s**	**6s**	**7s**
1s	-	T	TL	WTL	WTL	WTL	WTL
2s	T	-	TL	WTL	WTL	WTL	WTL
3s	TL	TL	-	WT	WTL	WTL	WTL
4s	WTL	WTL	WT	-	T	WTL	WTL
5s	WTL	WTL	WTL	T	-	WTL	WTL
6s	WTL	WTL	WTL	WTL	WTL	-	-
7s	WTL	WTL	WTL	WTL	WTL	-	-

Table 4.4: Significant difference in WER (W), completion time (T) and number of listens (L) on pairwise comparison of varied length segments.

resulting in an incomplete or incoherent word made it substantially more difficult to retain the segment. A participant stated:

> *The segments that started or ended with a clipped word were very distracting. My mind got stuck on the clipped words, making it impossible for me to retain the content.*

Nine participants suggested using natural pauses rather than abrupt cuts to split a long sentence in multiple segments. Eleven participants found a 3–4 second length optimal for content retention. These findings prompted us to design a segmentation scheme that splits an audio file based on the occurrence of natural pauses. If the individual segments so obtained exceeded a predefined length, the segments were recursively divided into smaller chunks of the desired length.

Experiment 2: Impact of Segment Ordering on Retention

We conducted independent samples t-test to analyze the effect of segment ordering on content retention. We found a significant difference in the WER when segments were played sequentially (M = 16.59, SD = 3.85) rather than randomly (M = 32.27, SD = 13.18); $t(22) = 3.96$, p=.001. We did not

	Time Taken (seconds)			WER (%)		
	CT	PT	S	CT	PT	S
M	211.7	370.3	37.8	4.5	5.0	3.1
SD	44.8	285.9	5.5	4.4	4.6	4.1

Table 4.5: Mean (M) and standard deviation (SD) for computer typing (CT), phone typing (PT) and speaking (S) tasks.

find a difference in task completion time or the number of times participants listened to segments. These results suggest that content retention is much higher when segments are presented sequentially. In the interviews, five participants specifically mentioned that sequential ordering increased their understanding of context, making it easier to estimate incoherent and clipped words. One of them stated:

> *When I hear the second segment after the first, I am able to connect it and even predict some of the words. Hearing segments in contiguous order makes cognition very easy.*

Experiment 3: Speaking versus Typing

Table 4.5 shows descriptive statistics for the tasks in the third experiment. The average speed of computer typing, phone typing, and speaking was 29.5, 19.3, and 161 WPM, respectively. A repeated measures ANOVA with a Greenhouse-Geisser correction determined a statistically significant difference in task completion time, $F(1.03, 23.74) = 25.41$, $p<.001$. Post hoc tests using the Bonferroni correction also revealed a significant difference ($p<.05$) in task completion time, even for pairwise comparisons of all three tasks. Though the average WER for speaking was lower than for typing, we did not find any statistical evidence to substantiate this.

We also requested participants to rate the three tasks on a ten-point scale for NASA TLX parameters to assess subjective workload. As seen in Figure 4.6, participants found that speaking caused

the least mental demand, physical demand, effort, and frustration. Moreover, they perceived their performance for the speaking task to be higher than for the typing tasks. A participant explained the ease of speaking vs. typing content:

> *Speaking is better as it comes naturally to us. It does not require any gadgets. Typing is something external.*

A repeated measures ANOVA with a Greenhouse-Geisser correction determined a statistically significant difference in mental demand ($p<.001$), physical demand ($p<.001$), performance ($p=.001$), effort ($p<.001$) and frustration ($p<.001$) for the three tasks. A pairwise comparison of the three tasks revealed a statistically significant difference ($p<.05$) in all five parameters for computer typing vs. phone typing, and phone typing vs. speaking. We also found a statistically significant difference ($p<.05$) in mental and physical demand for computer typing vs. speaking. These results suggest that speaking is a more efficient and easier output medium than phone or computer typing. We believe these results would be more significant and extreme if the participants were non-engineering students.

In summary, our cognitive experiments revealed that audio files should be partitioned by detecting natural pauses to yield segments of less than six seconds in length. These segments should be presented sequentially to ensure higher retention and less cognitive load on users. The users should complete micro-transcription tasks by speaking rather than typing.

In the following section, we examine how adaptations of Respeak to design and build ReCall affect crowd workers' performance on key activities they perform to complete a speech transcription task. We also conduct a usability study comparing ReCall and Respeak to examine the cumulative effect of these adaptations on usability perceptions, user experience, and transcription performance.

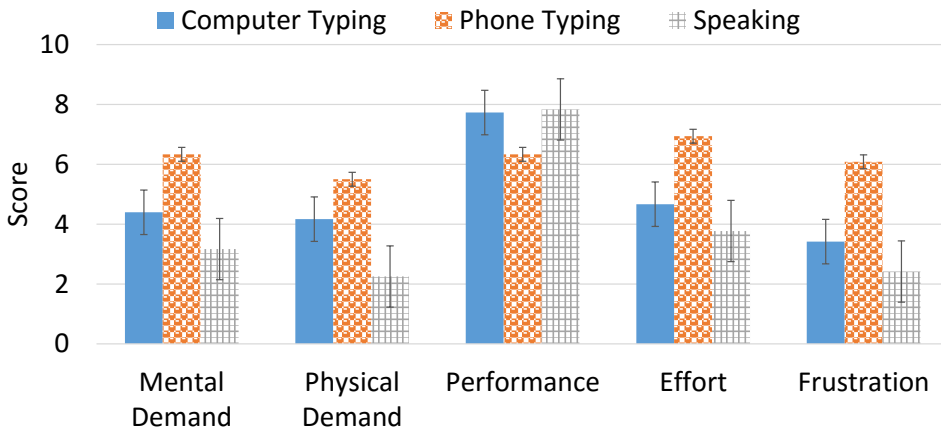

Figure 4.6: Evaluation of output modes on NASA TLX parameters.

4.4 Experimental and Usability Evaluations

We conducted three controlled experiments with low-income rural residents to examine how key differences between ReCall and Respeak affect their performance on three key activities required to complete a task: listening to an audio segment, re-speaking the segment into an ASR engine, and verifying the correctness of the ASR-generated transcript. We evaluated:

1. How phone types and channel types affect accuracy, time taken, and trials taken to listen to segments.

2. How phone types and channel types affect speech recognition accuracy when users re-speak segments.

3. How the modes to review transcripts affect accuracy, time taken, and trials taken to review transcripts.

In addition to investigating the isolated effect of phone types, channel types, and the modes to review transcripts, we also conducted a usability evaluation comparing ReCall and Respeak to exam-

ine the cumulative effect of these factors on usability, user experience, and task performance. The experimental and usability evaluations were approved by our institution's IRB.

4.4.1 Experimental Setup and Methods

Experiment 1: To examine how phone types affect listening performance, we conducted a within-subjects design experiment in which participants completed four listening tasks using a USD 600 smartphone (Pixel 2) and another four tasks using a USD 10 basic phone (Lava Captain N1). In each task, participants listened to a short segment stored in the phone's storage, read a text transcript, and verified the correctness of the transcript. Both conditions had two tasks with correct transcripts and two with erroneous transcripts. We kept the quality of segments (44kHz sampling rate) and the mode to review transcripts (reading) the same in both conditions.

To examine how channel types affect listening performance, we used the same experimental setup. Participants completed four listening tasks by calling an IVR app that uses the voice channel and another four tasks by using a smartphone app that uses the data channel. The quality of audio files varied based on the channel type used by the apps to play segments (8kHz in the IVR app vs. 44kHz in the smartphone app). We kept the phone type (Pixel 2) and the review mode (reading) the same in both conditions. We randomized and balanced the order in which participants completed tasks, and measured task completion time, trials, and accuracy.

Experiment 2: To examine how phone types and channel types affect re-speaking performance, we used desk stands to set up the basic phone (Lava), the high-end smartphone (Pixel 2) as well as an an entry-level smartphone (USD 90 Panasonic P100) next to each other (see Figure 4.7). We asked participants to speak five short Hindi segments into three phones. All phones used an IVR app, and the two smartphones also used an Android app for recording the segments simultaneously to avoid variations in the speaker's speech, tone, and diction. We submitted these segments to an off-the-shelf ASR engine and computed ASR accuracy for phone types and channel types.

Figure 4.7: A participant speaking sentences simultaneously in Pixel 2, Panasonic P100, and Lava Captain N1.

Experiment 3: To examine how the modes to review transcripts affect users' performance, we conducted a within-subjects design experiment with two conditions. In the first condition, participants completed four reviewing tasks by listening to an audio segment and then reading a text transcript. In the second condition, participants completed another four reviewing tasks by listening to an audio segment and then listening to an audio version of the transcript using a Hindi TTS system. For each task, we asked participants to verify if the transcript matched the content in the audio segment. Both conditions had two tasks with correct transcripts and two with erroneous transcripts. The type of phone (Pixel 2) and the quality of audio files (44kHz sampling rate) were kept the same in both conditions. We randomized and balanced the order in which participants completed tasks, and then measured task completion time, trials, and review accuracy.

Usability Evaluation: We provided a brief description about the ReCall and Respeak apps to participants. While we did not offer any demonstration of the apps upfront, we did provide verbal assistance when participants requested it. For each system, we requested that participants complete two randomly selected speech transcription tasks. To complete a task, participants had to listen to

a short audio segment, re-speak it into the app, and verify the correctness of ASR-generated transcript. Participants used the same phone to access the Respeak smartphone app and the ReCall IVR app. We randomized the order in which participants used the two apps, and measured task completion time, trials, and accuracy. We also requested participants to score both apps on usability parameters such as mental demand, performance, effort, and frustration.

At the end of each experiment, we asked open-ended questions to gather qualitative insights. We recorded and transcribed these responses, and subjected them to thematic analysis [65].

4.4.2 Recruitment and Demographic Details

We partnered with NYST, a grassroots organization that has active projects on community health and education in rural India. Leveraging their employees' network, we used snowball sampling to recruit 28 low-income rural residents.

Our sample had 18 female and 10 male participants. On average, participants were 22 years old. The majority (68%) had completed or were pursuing a bachelor's degree, three participants had completed a master's degree, three had completed high school, and those remaining had dropped out after middle school. About 93% of participants were unemployed and the remaining (N=2) were engaged in a temporary part-time employment. The median monthly family income for a family size of five people was USD 182, meaning that half of the participants were surviving on USD 1.21 per day. Fifteen participants (53%) came from families engaged in blue-collar work (e.g., farmers, laborers) while the remaining were from families of white-collar workers (e.g., shop owners, private jobs, teachers). All participants were native speakers of a dialect of Hindi and most of them had limited understanding of English.

Fifteen participants had a smartphone, eight had a basic phone, three had a feature phone, and two borrowed a basic phone from their family members. Most participants were new users of mobile phones; the median phone ownership time was 1.5 years. Twelve participants used special tariff

vouchers (STVs) offered by MNOs to access unlimited voice calls and capped data bundles. They often borrowed phones from family members to use the Internet. Twenty-one participants used WhatsApp and 15 participants used Facebook. Only two participants had previously used IVR systems.

4.4.3 Findings of Experimental and Usability Evaluations

Experiment 1: The majority of participants (75%) found it harder to listen to segments on the basic phone due to *"lack of clarity"* and *"buzzing sound"* because of clipping. As a result, participants listened to segments significantly more times on the basic phone (M=5.5, SD=1.1) than on the smartphone (M=4.5, SD=0.7), t(23)=4.44, p<.001). They also took significantly more time to complete listening tasks on the basic phone (M=81s, SD=13s) than on the smartphone (M=74s, SD=11s), t(23)=2.48, p=.02. Since participants could perform a task multiple times until they were satisfied with their performance, we did not find any significant difference in listening accuracy on the basic phone and the smartphone.

Many participants took more time to complete tasks on the IVR app because of *"lower volume and less clarity"* of segments and prompts. Our analysis revealed a significant difference between the task completion time on the IVR app (M=83s, SD=12s) and the smartphone app (M=76s, SD=10s), t(23)=2.24, p=.03. Although many participants took more trials to listen to segments and completed listening tasks with lower accuracy on the IVR app, we did not find significant differences between the listening trials on the IVR app (M=7.5, SD=2.8) and the smartphone app (M=6.3, SD=2.7), as well as between listening accuracy on the IVR app (M=58%, SD=24%) and the smartphone app (M=66%, SD=24%).

Experiment 2: A two-way repeated measures ANOVA (phone types $\times$ channel types) revealed a significant main effect of channel types, F(1,85)=14.38, p<.001, and no effect of phone types on ASR accuracy. Figure 4.8 shows the distribution of ASR word error rates (WER) for different combinations of phone types and channel types. ASR WERs were lowest (M=5%, SD=5%) for seg-

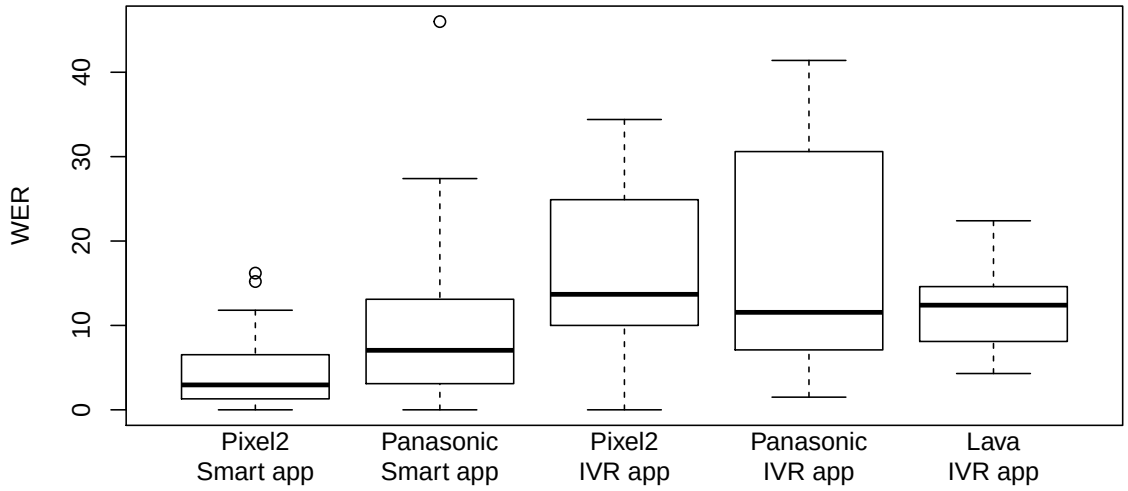

Figure 4.8: Distribution of WERs for different combinations of phone types and channel types.

ments recorded on the smartphone app on Pixel 2. The WERs increased significantly for segments recorded on the IVR app on the same phone (M=16%, SD=10%), t(17)=4.99, p<.001, due to down-sampling of the segments by the voice channel. For the same reason, we also found a significant difference between the WERs for segments recorded on the smartphone app (M=11%, SD=11%) and the IVR app (M=17%, SD=13%) on Panasonic P100, t(17)=2.60, p=.01. These results indicate that the segments spoken by ReCall users may yield higher WERs than the Respeak users.

We found a significant difference in the WERs between Pixel 2 and Panasonic P100 when participants spoke segments into the smartphone app, t(17)=2.62, p=.01, perhaps due to differences in the number of microphones in these devices and their positioning; Pixel 2 has two microphones (one at the top and other at the bottom) compared to one microphone in the Panasonic (at the bottom). However, when the segments were recorded on the IVR app, we did not find a significant difference between any combinations of the three types of phones. These results indicate that phone types may affect ASR accuracy for users of the Respeak smartphone app. However, phone types should not significantly affect ASR accuracy for users of the ReCall IVR app.

Experiment 3: The majority of participants (66%) found it easier and faster to read text transcripts

rather than listen to audio version of the text transcripts. Participants shared several reasons for their preference for reading transcripts. Many participants found it difficult to remember the content in audio transcripts because of the *"weird accent"* and *"mechanical delivery"* of TTS system. Some participants experienced a high cognitive load in remembering the audio segment as well as the audio transcript. A few participants noted that they could review text transcripts at their own pace and spot errors easily in them. Our statistical analysis supported these observations. We found a significant difference between the review accuracy for text transcripts (M=80%, SD=20%) and audio transcripts (M=48%, SD=21%), $t(23)=5.46$, $p<.001$. Several participants were worried that listening to transcripts may require more time and more trials, especially in noisy environments. Although we found no difference between the trials taken to complete review tasks, we found a significant difference between the time taken by participants to read transcripts (M=93s, SD=18s) and listen to transcripts (M=106s, SD=21s), $t(23)=2.23$, $p=.03$. These results indicate that ReCall users may take more time to review transcripts and may make more reviewing mistakes than Respeak users.

Usability Evaluation: Participants successfully completed all tasks and took comparable number of listening and re-speaking trials on both ReCall and Respeak. However, participants took more time to complete tasks on ReCall than on Respeak and produced transcripts with higher WER. We found significant differences in the task completion time on Respeak (M=173s, SD=108s) and ReCall (M=230s, SD=90s), $t(21)=2.48$, $p=.02$, as well as in the transcription WERs on Respeak (M=18%, SD=16%) and ReCall (M=25%, SD=24%), $t(21)=1.99$, $p=.05$. These results indicate that ReCall users may produce transcripts in 33% more time and with 8% lower accuracy than Respeak users.

Participants experienced higher mental demand, effort, and frustration, and lower performance on ReCall than on Respeak. Table 4.6 shows the median scores for the two systems on four usability parameters. A Wilcoxon signed-rank test indicated significant differences between ReCall and Respeak on mental demand ($W=14$, $Z=3.28$, $p<.001$), performance ($W=65$, $Z=2.26$, $p=.02$), and frustration ($W=0$, $Z=2.80$, $p<.01$).

	Mental Demand	Performance	Effort	Frustration
ReCall	5	7	3.5	1
Respeak	2	8	2.5	1

Table 4.6: Median scores of different usability parameters on a ten-point scale (1–low, 10–high) for ReCall and Respeak.

Five participants expressed difficulties in listening to segments on ReCall because of downsampling by the voice channel. Three participants found ReCall more mentally demanding than Respeak because of additional attention they paid to listen to audio prompts. Two participants found ReCall slower, perhaps due to additional time ReCall took to convert ASR-generated text transcripts into TTS-generated audio transcripts. Several participants also struggled while using Respeak. For example, six participants were confused when to repeat audio segments despite a beep sound that served as a cue to start speaking. Four participants were unsure about how to interact with the touch interface and two participants found it overwhelming to operate a smartphone. Participants with prior smartphone experience (N=14) preferred Respeak while many new smartphone users and non-smartphone users (N=8) preferred ReCall. Participants mentioned ease of listening to audio files and reviewing crowd work by reading transcripts as reasons for their preference for Respeak. On the other hand, participants preferred ReCall for its inclusive and accessible design. A participant stated:

> *There is no dependency on the Internet. Anyone can do the work even on basic phones as well.*

The usability evaluation also helped us discover and address usability barriers in ReCall. For example, participants were prompted to press pound key after re-speaking segments to signal the end of recording to the app. Since five participants forgot to press the key after recording segments, we implemented a feature that sends the signal automatically after detecting silence for two seconds.

To summarize, the experimental evaluations investigated how adaptations of Respeak to ReCall affect users' performance on three key activities they do to complete transcription tasks. The usability evaluation improved the usability of ReCall, examined the cumulative effect of different factors on transcription performance, validated the findings of the experimental evaluations, and provided enriching insights about participants' preferences and perceptions.

In the coming sections, we describe the field deployments of Respeak, BSpeak, and ReCall in India, which happened sequentially.

4.5 Respeak Field Deployment

To examine the feasibility, acceptability, and usability of transcribing audio files vocally, we first deployed Respeak smartphone app to low-income student. We hoped that the deployment learning will be a stepping stone to ReCall deployment (i.e., our eventual goal of creating a voice-based marketplace accessible via ordinary phone calls). Since many university students have smartphones connected to the Internet and also have financial constraints that might motivate them to earn mobile airtime, we sent an email inviting IIT Bombay students and interns to participate in our controlled deployment. We randomly selected 25 respondents as users and conducted a face-to-face orientation session with them to install the Respeak app on their personal smartphones, show them how to use the app, and collect demographic information.

4.5.1 Tasks

We submitted thirteen Hindi and eight English audio files to the Respeak engine for transcription. To stress test Respeak, we selected audio files that had ambient noise and heavily localized Hindi or English accents. The files contained varied content, including public speeches, telephone conversations, news, television advertisements, songs, interviews, YouTube content, and online lectures. The total duration of the Hindi and English files was 43 minutes and 12 minutes, respectively. The Respeak engine partitioned Hindi files into 499 segments and English files into 257 segments to yield

	Interview	Song	TV ad	News	Public speech	Phone call	YouTube video	Lecture
English	177	-	10	10	15	-	35	9
Hindi	-	77	-	17	313	37	54	-

Table 4.7: Number of tasks for each category of transcribed content by language.

756 unique micro-transcription tasks (see Table 4.7). The threshold length for the segmentation scheme was based on the speaking rate in the audio file: the length for public speeches and songs was 5–6 seconds, news and YouTube videos was 4 seconds, and interviews and phone calls was 3 seconds. The collective download size of all tasks was 85 MB and the cost of downloading them was roughly ₹20 (USD 0.30) on a 3G connection. Each task could be performed by a maximum of ten users who could see a high-level overview of their transcription accuracy, amount earned, payment processed, and completed tasks.

4.5.2 Reward Scale and Payment

The Respeak reward structure was designed to keep the cost of transcription below USD 1 per minute. Each transcription task was assigned a reward equal to ₹0.2 multiplied by segment length in seconds. We hypothesized that for each segment, if we aligned the transcripts generated by five users and if all of them performed the task with a high accuracy, the maximum transcription cost would still be USD 0.92 per minute. Each time a user submitted a transcript for a segment, we compared their output with the pre-computed ground truth. If the transcript's accuracy was $\geq$ 80%, we added the entire task reward to the user' earning. If the accuracy was $\geq$ 50%, we added a proportionate percentage of the task reward to the user' earning. A user received no reward if the accuracy was $<$ 50%. This reward structure gave users the incentives to produce speech transcription with more than 80% accuracy, gave proportionate returns to average performers, and penalized poor performers. Once the cumulative earnings of a user reached ₹10, we processed a mobile airtime credit of the same value to them. The maximum amount a Respeak user could earn by doing Hindi tasks was

₹514 (USD 7.80) and by doing English tasks was ₹152 (USD 2.30). The reward structure could also be designed differently to satisfy other optimization goals.

Ideally, the transcripts submitted by Respeak users should be evaluated by comparing them to the best estimation transcript generated in first-stage merging. However, at the time of deployment we were unsure about how and when people would use the app, forcing us to use the pre-computed ground truth for comparison. We ran the comparison module every 15 minutes to balance the desire of users to receive immediate feedback and the need to simulate a delay that would occur awaiting transcripts generated by others for MSA and majority voting process.

4.5.3 Methodology to Evaluate Deployment

We conducted quantitative analyses of transcription performance, cost, and turnaround time to evaluate Respeak. We also conducted in-depth, semi-structured, face-to-face interviews with 20 Respeak users at the end of deployment. Each interview lasted around 40 minutes and covered several themes, including information on the general technology use, user experience and usability, conception of Respeak, and benefits and limitations of the Respeak app. We recorded, transcribed, and analyzed the interview using open coding.

4.5.4 Respeak Users Demographics

Fifteen Respeak users were male and ten were female. Fourteen were students, six were contractual staff, and five were summer interns. Twenty users were from varied engineering departments, and five from the humanities. Eighteen users had or were pursuing a bachelor's degree, six had or were pursuing a master's degree, and one was pursuing a Ph.D. Fifteen users did not have any scholarship, stipend, or salary and were supported by their families. The average monthly income of employed users was USD 293, and their average monthly family income was USD 1557.

All users owned an Android smartphone, had cellular Internet access, and used their phones for an

average of 5 hours a day. Despite heavy and ubiquitous phone usage, 17 participants had a shoestring budget for mobile airtime and data, and relied on the free WiFi provided by the university. Like participants in the cognitive experiments, all users rated their English language skills and Hindi speaking skills highly. However, 22 users reported their Hindi typing skills to be bad. Sixteen of them did not even know how to type in Hindi.

4.5.5 Findings

The Respeak app was deployed for a month with 25 users. Figure 4.9 depicts the time series analysis of the number of tasks completed by active Respeak users. The low activity between August 10–20 corresponded to an intermittent campus-wide Internet outage at the university where we deployed Respeak. Though the deployment ended on August 31, some users continued using the app for another 20 days. 756 audio segments were presented as 5464 micro-tasks to the users, who transcribed the segments successfully with an individual average WER of 23.7%. On average, Respeak users listened to segments 2.7 times and re-spoke them 2.1 times before moving on to the next task. The median time for task completion was 36 seconds, and the cost of transcription was USD 0.83 per minute. Collectively, Respeak users spent 39.8 hours using the system and earned ₹3036 (USD 46). The expected payout for an hour of their time was ₹76 (USD 1.16), one-fourth of the average daily wage rate in India [14]. The Respeak engine combined the transcripts generated by five users for each segment, reducing the average WER to 10.6%. The best alignment yielded a WER of 6.8%.

Efficiency of Speech Transcription

The average WER of the transcripts generated by ASR engine for individual Respeak users was 23.7%. We performed a series of experiments to measure the improvement in transcription using MSA and majority voting. For each segment, we conducted ten runs of experiments. In each run, the transcripts generated by three randomly selected Respeak users were used for MSA and majority voting in first-stage merging. The WERs obtained in each of the ten runs were averaged

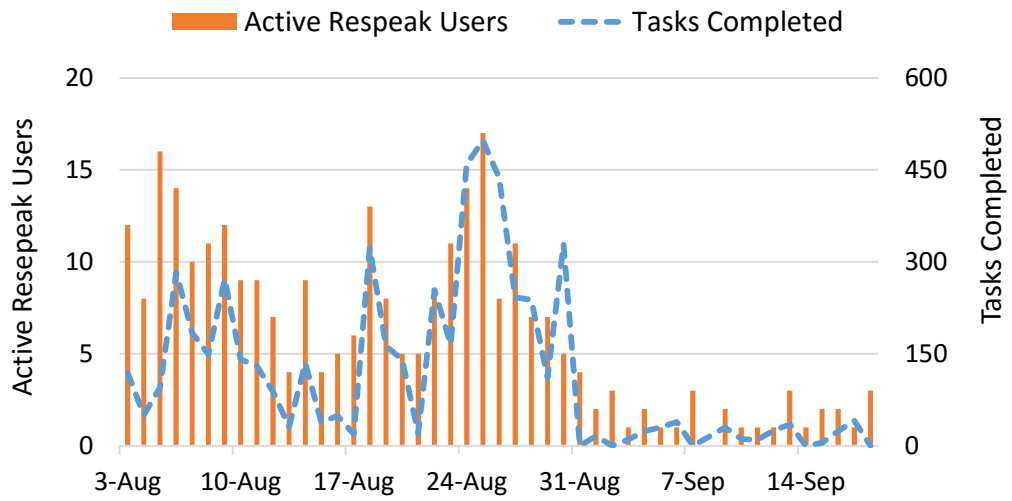

Figure 4.9: Time series analysis of active users and tasks completed.

for evaluation. By aligning the transcripts generated by 3 speakers, the WER dropped to 15.1% (an improvement of 36.3%). We used the same setup to align transcripts generated for each segment by 5 randomly selected Respeak users, and the WER dropped further to 13.2% (an improvement of 44.3%).

A closer inspection of users' transcripts and the ground truth revealed interesting cases that were registered as errors by the comparison module but were semantically correct. The app's Google ASR engine transcribed several words in English and Hindi differently for different speakers. In English, the words were often contracted or abbreviated (e.g., it is vs. it's; Doctor vs. Dr.), and the numbers were transcribed either in numeric or textual format (e.g., 3 vs. three) for different speakers. In Hindi, multiple spellings with minor variations were output for the same word based on the stress, intonation and nasality used by speakers (see Figure 4.8).

The manual correction of such corner cases in Respeak user transcripts lowered the average WER for individual users from 23.7% to 21.9%. We recomputed the set of experiments where transcripts generated by multiple users were aligned, and the WER dropped to 12.5% and 10.6% when transcripts generated by 3 and 5 randomly selected users were aligned, respectively. Thus, the alignment

Actual Word	Spelling 1	Spelling 2	Spelling 3	Spelling 4
उन्होंने (unhone)	उन्होंने (unhone)	उन्होने (un-hone)	उंन्होने (unnahone)	उन्न्होंने (unnhonne)
भाईयों (bhaeeyon)	भाईयों (bhaeeyon)	भाईयोँ (bhaeeyoun)	भाइयो (bhaiyo)	भाइयों (bhaiyon)

Table 4.8: Different spellings generated by Google ASR engine for words in Hindi. Equivalent spelling in Latin script is in brackets.

of transcripts generated by five randomly selected users reduced the average WER by 55.3%. In future deployments, we resolved these corner cases automatically in the comparison module.

To evaluate the effect on WER and cost as more transcripts are used for alignment and majority voting, we randomly selected 50% of 391 tasks that were each completed by ten Respeak users. We conducted ten runs of experiments; in each run, we used the transcripts generated by K randomly selected Respeak users. We varied the value of K from 1–9 and averaged the WER obtained for each value of K over ten runs of experiments. The cost of transcription was calculated using the rate of ₹0.2 per second of transcription per user, an overestimate that assumed that users would receive the entire reward amount promised for each task. As depicted in Figure 4.10, the WER decreased as K increased, and the cost of speech transcription linearly increased with K.

To compare Respeak with a state-of-the-art speech recognition engine, we submitted the original audio files to the Google Cloud Speech API [41] after noise reduction. The API yielded transcription with the overall WER of 50%, 4.72 times higher than the WER obtained by Respeak. The WER for Hindi audio files was 54%, 6.3 times higher than the WER obtained by Respeak. These results suggest that Respeak capitalized on the benefits of re-speaking and crowdsourcing to outperform transcription generated by the state-of-the-art speech recognition engine alone.

Tables 4.9 and 4.10 report the WER of transcription obtained for different languages and content

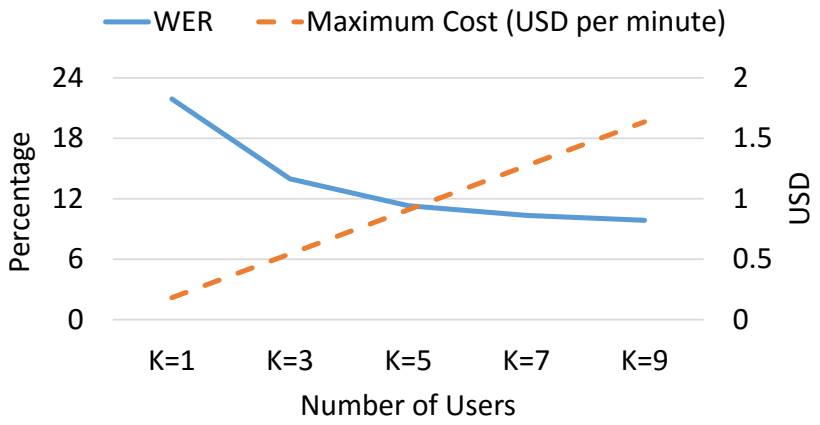

Figure 4.10: Effect of number of users on WER and cost.

Language	WER (%) before correction			WER (%) after correction		
	K=1	K=3	K=5	K=1	K=3	K=5
English	26.9	19.8	16.7	26.2	18.1	15.2
Hindi	19.9	13	11.7	17.1	10	8.6
Both	23.7	15.1	13.2	21.9	12.5	10.6

Table 4.9: WER obtained by Respeak for English and Hindi languages.

types by aligning transcripts generated by K users during first-stage merging. Our interviews indicated that six users found it easier to do Hindi tasks, and four found it easier to do English tasks. The language preference existed either because of better language skills or faster recognition from the ASR engine in their preferred language.

Seven users found it easiest to re-speak song segments while others found interviews (N=3), speeches (N=2), news (N=1), lectures (N=1) and poems (N=1) the easiest. Six users found it very difficult to understand the segments containing an interview of a former president of India, three found it hardest to retain the advertisement segments because of the audio's unclear accent, and the

Content Type	WER (%) before correction			WER (%) post correction		
	K=1	K=3	K=5	K=1	K=3	K=5
Interview	27.8	21.2	18	27.2	19.1	16.4
Song	22.9	13.2	10.3	20.2	10.9	7.8
TV ad	31.2	26	24.3	29.1	23.8	19.7
News	23.2	14	9.8	20.6	10.7	8.3
Public speech	20.1	13.2	12	17.4	10.3	8.8
Phone call	25.9	18.8	17.4	22.8	15.2	12.8
YouTube video	16.9	11.2	10.2	14.9	8.9	7.8
Online Lecture	17.4	13.2	10.7	16.5	11.3	9.8

Table 4.10: WER obtained by Respeak for different content categories.

other three found it difficult to re-speak Bollywood song segments because of the *"cheesy"* lyrics. The remaining users found no difference in the difficulty level of tasks based on content type. Three users sang the segments containing songs rather than merely re-speaking them. A user explained how he had to remain aware of his surroundings while re-speaking song segments:

> *Singing songs was difficult as I had to speak cheesy lines like, "My heart is beating for you". My parents overheard me re-speaking this and asked me, 'Who are you talking to; what is going on?' It was awkward to explain.*

Five users found it useful that the segments of an audio file were presented in a sequential order as tasks. However, three users found it monotonous to do tasks continuously for the same audio file. They suggested an alternate scheme where small blocks from different files could be randomly presented, where each block could have segments from the same audio file presented sequentially. Four users found it challenging to do tasks with clipped words either at the beginning or end; they

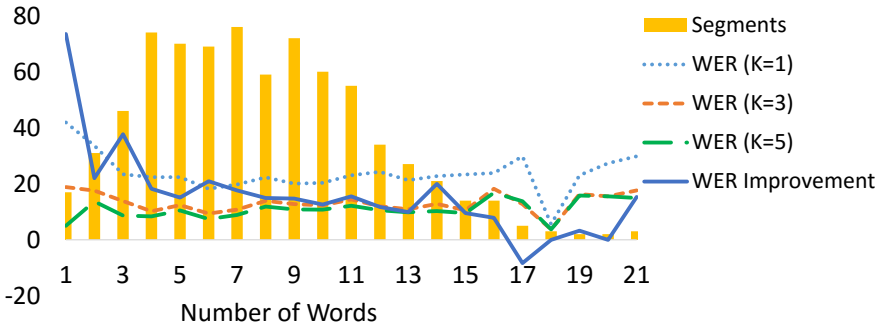

Figure 4.11: WER for segments of varying word length.

were unsure whether to re-speak or ignore such words.

Figure 4.11 plots the average WER for segments of varying word lengths. Surprisingly, the WER for individual users did not vary significantly as the number of words in segments increased. However, the improvement in WER by aligning transcripts from 5 users rather than 3 users decreased as the number of words in a segment increased. Though the randomness in errors increased with the increase in number of words in a segment, the errors were sparsely distributed reducing the performance improvement gained by MSA and majority vote.

Payments

We processed mobile airtime of ₹3040 (USD 46) to 24 users. The top 3 users earned 44% of the total payments, while the top 20% and 50% users earned 60% and 87%, respectively (see Table 4.11). Ten users earned more than their monthly phone expense. Several users reported that using the app for ten minutes daily was sufficient to subsidize their phone expenses. Respeak became a portal to transfer mobile airtime to their phones. A user reported:

> *I exhausted my phone balance while chatting with a friend. I did not have money to refill my phone online. I quickly did some tasks on Respeak using free WIFI, received a top-up, and then called him.*

Amount Earned (₹) $\leq$	Respeak Users
100	15
200	5
300	1
400	0
500	3

Table 4.11: Amount earned by Respeak users.

All but one user were happy to receive the amount earned as mobile airtime. Some users suggested payments in the form of food coupons (N=6), Amazon gift coupons (N=5), and top-ups of higher value that results in the equivalent mobile airtime [1] (N=3). Two users emphasized the need to process a ₹10 mobile airtime for immediate gratification. A user stated:

> *There is not much you get for ₹10 in market other than mobile airtime. If the amount when payment is processed is higher, many people may stop using the app, even before they reach that number.*

Eight users found that their efforts using Respeak were commensurate with the amount they earned. Six users found that the money they earned exceeded their efforts, while six others felt otherwise.

Instrumental Benefits

Seven Respeak users reported receiving instrumental benefits from the app use. Three found that Respeak improved their language and oral skills. While re-speaking audio segments, they focused on pronouncing the words correctly for faster recognition by the ASR engine. Often, they searched

[1] A top-up of ₹10 gave ₹7.8 in airtime. The lowest top-up that gave full mobile airtime is around ₹100 for different providers.

online for the meaning and pronunciation of unfamiliar words, thereby expanding their vocabulary. Respeak provided them with the opportunity to speak English aloud *"without being judged by others."* One user reported a new-found interest in the content he transcribed, viz., an old Bollywood song on YouTube. Another found Respeak to be a challenging yet fun exercise that improved his cognitive abilities. Two users reported acquiring new knowledge while doing the tasks and found some of the speeches inspiring. One of them stated:

> *Receiving a mobile recharge was good. However, listening to speeches and interviews increased my general knowledge. Most importantly, the app improved my pronunciation as I was focusing to pronounce words better so that they get recognized.*

Feedback on Respeak

Figure 4.12 presents average user ratings for NASA TLX parameters on a ten-point scale. Users enjoyed Respeak for a wide variety of reasons, including earning mobile airtime (N=8), excitement to see their speech recognized (N=6), ability to track their accuracy (N=4), easy-to-use interface (N=4), listening to interesting content (N=1), the opportunity to practice speaking English out loud (N=1), and the chance to compare their accuracy to others (N=1). Even before our interviews, we received user emails describing their enthusiasm for Respeak. One such enthusiastic user wrote:*"Respeak is cool. Got a little excited with the top-up I just received."*

Nine users found Internet usage to be a barrier to using the app. Seven users found it difficult to get their speech recognized and four faced challenges in getting ASR engine to recognize people's names. Though users voluntarily signed-up to participate in the deployment, four reported time constraints that limited their app use. Eight users suggested gamification to make the app more entertaining. Five wanted functionality to skip tasks for unclear or difficult-to-retain segments. Two suggested including a feature that let users type to edit the transcript generated by the ASR engine after multiple unsuccessful re-speaking trials. Three users wanted the ability to filter tasks

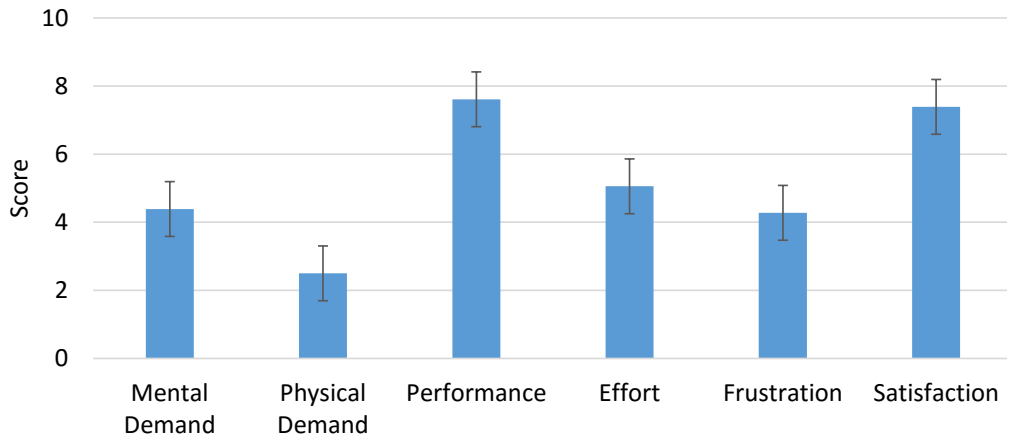

Figure 4.12: Average ratings by Respeak users for several parameters.

by language. One each suggested incorporating graphs to track improvement in user accuracy, a leaderboard, and a feature to regulate playback speed.

The users considered the ideal Respeak demographic to include: students (N=15), unemployed people (N=4), home makers (N=2), people spending long hours commuting (N=2), and those interested in learning oral skills (N=1). After the deployment, eleven participants expressed an interest in using the app daily, primarily to earn mobile airtime and improve their language skills; six stated that they would use it sparingly when they need mobile airtime; three stated that the lack of time would inhibit their app use in the future.

In the next section, we describe the details of the field deployment of BSpeak app—an accessible version of the Respeak app—with low-income people with visual impairments.

4.6 BSpeak Field Deployment

Through the network of blind trainees, staff, and alumni of Enable India, we used snowball sampling to recruit 24 low-income blind Android users to participate in a two-week deployment of BSpeak. During a face-to-face orientation session, we installed the app on users' phones, gave them a brief

demonstration of BSpeak, and collected demographic information. To set the right expectations, we informed all participants that BSpeak is a research prototype and our goal is to investigate its accessibility and feasibility, and that there are no immediate plans to release the app widely. At the end of the deployment, we conducted a web survey to evaluate their user experience. The survey had 13 subjective questions that spanned several themes including benefits and limitations of the BSpeak app and its potential to supplement their income.

4.6.1 BSpeak Users Demographics

Eighteen users were totally blind and six were partially blind. Users were 27 years old, on average. Nineteen users were male and five were female. Seventeen users had completed a bachelor's degree, four had completed high school, two had finished middle school, and one had earned a master's degree. Employed users (N=13) had an average monthly income of USD 313. Unemployed users (N=11) were dependent on their family with average monthly income of USD 169 for a family of size four, and thus were living below the poverty line of USD 1.90 per day [25]. Even though all users resided in Bangalore at the time of the study, most of them moved to Bangalore in the last three years from rural and peri-urban regions of nine Indian states. Half of them had family members in rural regions who were either dependent on them (N=3), or were working as farmers and laborers (N=9).

All users had access to mobile Internet, and used TalkBack as the primary screen reader on their phone. Their average monthly phone expense was ₹300 (USD 4.5). None of the users had prior experience with speech transcription and crowdsourcing marketplaces. Users were native speakers of Kannada (N=13), Hindi (N=4), Telugu (N=3), Marathi (N=1), Konkani (N=1), Malayalam (N=1), and Assamese (N=1). Several of them had poor Hindi and English language skills. While the self-reported scores, on a ten-point scale, for local language listening and speaking skills averaged to 9.8, the average scores for English listening, English speaking, Hindi listening, and Hindi speaking skills were 7.2, 7, 6, and 5.9, respectively. During the orientation session, many users made frequent

grammatical errors while speaking English and Hindi.

4.6.2 Speech Transcription Tasks

We selected 27 audio files in Indian English and Hindi for transcription by BSpeak users. Out of these, 21 audio files were the same as those used in Respeak's deployment because we aimed to compare the accuracy of crowdwork by blind users to sighted users. Since a majority of the BSpeak deployment participants expressed discomfort with Hindi, we selected more transcription tasks in English than in Hindi. Although the participants were more comfortable in other local languages like Kannada and Telugu, we could not provide tasks in these languages since they are not yet supported by Android's built-in ASR engine. The combined duration of the audio files was 2.75 hours, and it comprised of a wide variety of content including interviews, lectures, news, phone calls, public speeches, songs, TED talks, TV advertisements, and YouTube programs. The engine segmented these files to yield 2,560 segments (i.e., micro transcription tasks) that varied from three to six seconds in length. We outline the number of audio files, tasks, and total duration for different content types in Table 4.12 and languages in Table 4.13.

4.6.3 Reward Scale and Payment

To ensure that the earning potential of BSpeak users equaled that of Respeak users, we employed the same reward structure used in Respeak's deployment. The maximum amount a user could earn by completing all Hindi and English tasks was ₹514 (USD 7.8) and ₹1,470 (USD 22.3), respectively. Users could access the reward structure, amount earned, accuracy, list of all completed tasks, and date of last transfer on the BSpeak app.

Content type	Audio files	Tasks	Duration
Interview	3	275	13.1
Lecture	1	9	0.4
News	2	27	1.4
Phone call	3	37	1.8
Public speech	8	1102	88.7
Song	3	77	7.2
TED talk	3	931	46.5
TV ad	1	10	0.7
YouTube program	3	89	5.4

Table 4.12: Number of audio files, tasks, and total duration (in minutes) for each content type.

Language	Audio files	Tasks	Duration
English	14	2060	123 mins
Hindi	13	500	43 mins

Table 4.13: Number of audio files, tasks, and total duration for each language.

4.6.4 Methodology to Evaluate Deployment

We used a mixed-methods approach spanning quantitative analysis of word error rate, cost, and performance, as well as a qualitative analysis of surveys to evaluate BSpeak. All but one users completed the survey at the end of the deployment. We subjected their responses to thematic analysis as outlined in [65]. We also compared BSpeak's use by blind people to Respeak's use by sighted people on tasks completed, WER, performance, and transcription cost.

4.6.5 Findings

BSpeak users used the app for over 208 hours and completed 16,000 tasks, repeating 2,560 segments with an average accuracy of 61.6% to earn ₹7,310 (USD 110). Twelve users completed Hindi tasks 3,133 times, and 24 users completed English tasks 12,872 times. On average, users listened to a segment three times and repeated the content into the ASR system 1.7 times. The median task completion time was 49 seconds. The expected payout per hour of the app use was ₹36, which is comparable to the average hourly wage rate in India [14]. The engine combined the transcripts generated by eleven users to yield a transcription with 87.1% accuracy and USD 1.20 per minute of transcription cost.

Figure 4.13 shows the time series analysis of tasks completed and unique active users. Though the deployment ended after two weeks, some users continued using the app for a month. BSpeak became very popular among the social network of our users; we received over 20 requests from blind friends of users to access the app. The popularity of BSpeak prompted Enable India's management to request a disclaimer from us:

> *We understand that the BSpeak app has become a super hit among our trainees and staff. Most people we have asked love it and also have been making a lot of money through the tasks. We request you to send us a statement that you chose people on your own for deployment based on your criteria, and they are making money through the app based on their skills, and that the organization has shown no favoritism in this.*

Speech Transcription Accuracy and Cost

The average WER for transcripts generated by individual BSpeak users was 38.4%. We ran a series of experiments to reduce the WER by aligning transcripts generated by multiple users. For each segment, we conducted ten runs of the experiment. In each run of the experiment, we randomly se-

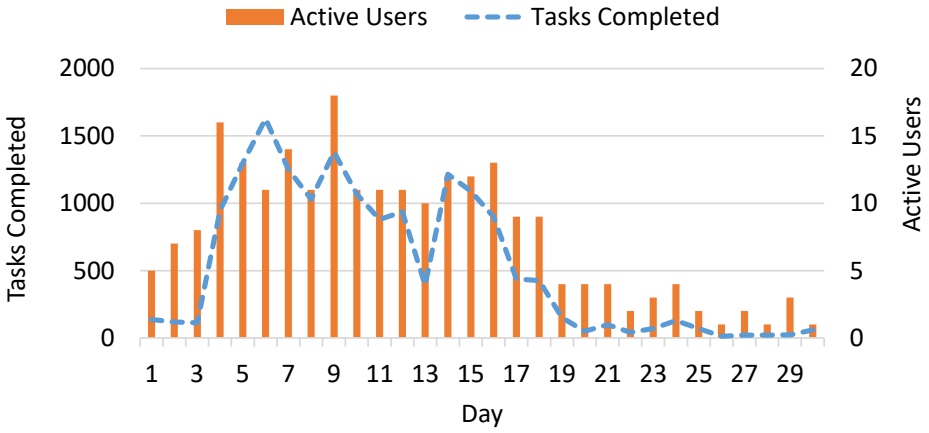

Figure 4.13: Time series analysis of active users and tasks completed.

Ground truth:	is	not	as	rosy
Transcript 1:	*its*	not	*at*	rosy
Transcript 2:	is	not	as	*rose*
Transcript 3:	is	—	as	rosy
Majority voting:	is	not	as	rosy

Table 4.14: Alignment of transcripts obtained from three speakers. Missing words are marked as — and incorrect are *italicized*.

lected transcripts generated by K users and aligned them using MSA and a majority voting process. We averaged the WER of the best estimation transcript generated in each run. Table 4.14 shows how WER of an English segment was reduced by aligning transcripts generated by three users.

We used the experimental setup to align transcripts generated by three, five, seven, nine, and eleven users by varying the value of K. Table 4.15 shows the number of tasks done by K or more users, WER obtained, and transcription cost. To compute the cost, we multiplied transcription rate (i.e., ₹0.2 per second per user) with expected payout (i.e., 61.6% of the transcription rate). Our analysis indicated that the overall WER decreased as the value of K increased. The cost linearly varied based on the number of transcripts used in the alignment process.

K	Tasks done by $\geq$ K users	WER (%)			Cost (USD/min)
		Overall	English	Hindi	
1	2,560	38.4	40.7	30.1	0.1
3	2,509	30.7	33.0	22.6	0.3
5	1,904	26.8	29.9	19.0	0.6
7	833	19.9	20.9	12.1	0.8
9	708	17.6	17.9	11.3	1.0
11	89	12.9	13.0	8.3	1.2

Table 4.15: WERs and transcription cost obtained after aligning transcripts generated by K users.

Table 4.15 also shows the WER obtained for English and Hindi tasks after aligning transcripts generated by K users. The average WER for English tasks was higher than Hindi tasks because all users were non-native English speakers and had no choice but to complete tasks in English. In contrast, only those users who were confident in Hindi opted to complete Hindi tasks. In our survey, several users reported struggling with English tasks because of *"unfamiliar vocabulary," "fast pace,"* or *"poor recognition by ASR engine."*

Figure 4.14 plots the WER obtained after aligning the transcripts generated by K users for different content types. The WER for content created for mass consumption like news, lectures, songs, and YouTube programs were low. The WER for public speeches and telephone calls was high because users struggled with ambient noise. The WER for interviews was high due to the prevalence of heavy accents and the overlapping of speakers. Although TED talks and TV advertisements are also created for public consumption, several users struggled with unfamiliar accents and technical terminologies, and yielded poor performance compared to other content types. The analysis of survey responses also indicated that users found speeches (N=11), songs (N=6), and YouTube programs (N=3) easy to remember and re-speak, and faced difficulties in transcribing interviews (N=4), telephone calls (N=4), and TED talks (N=4).

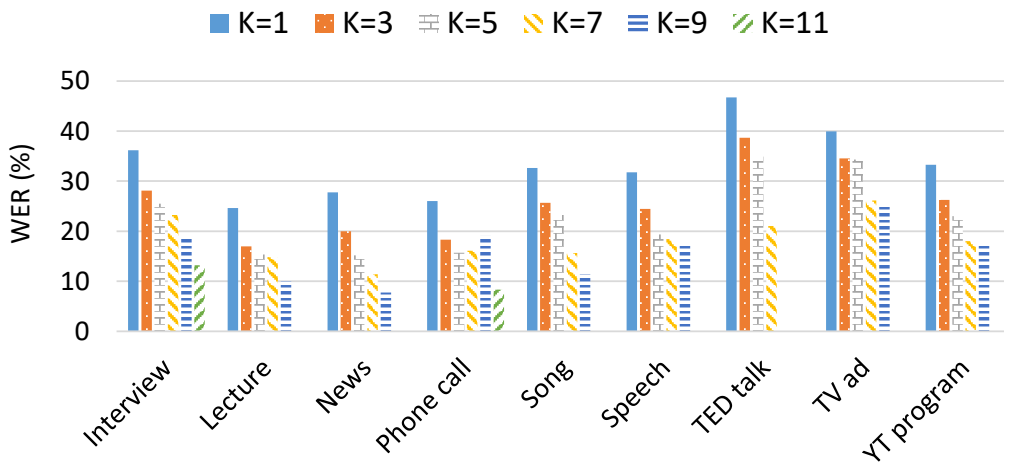

Figure 4.14: WER obtained after alignment of transcripts generated by K users for different content types. A missing bar indicate that less than K speakers completed tasks.

To examine the benefit of requesting blind users to re-speak content into the ASR engine instead of inputting audio files directly to it, we submitted the segments to the Google Cloud Speech API [41]. The average WER of the transcripts thus obtained was 52% (35% more than the average WER obtained by BSpeak users). Since blind users repeated audio content into a high-quality microphone of their device in a quiet environment, BSpeak's WER was lower than the WER yielded by the ASR engine on original audio segments.

Financial and Instrumental Benefits

BSpeak users collectively earned ₹7,310. The average amount earned by users was ₹304 (USD 4.60) and the maximum amount earned was ₹1,050 (USD 16). Nine users earned more than their monthly phone expense in just two weeks. Table 4.16 shows the distribution of the amount earned by BSpeak users. All users agreed that BSpeak has a strong potential to supplement their earnings. They valued it as a tool to *"utilize their free time in earning money."* Several users earned money for the first time in their life. One such user stated:

> *I am grateful to you for creating the app. I earned money for the first time and learned the value of each rupee.*

While 19 users found the amount earned commensurate to the time spent in completing tasks, five users suggested increasing the reward amount to at least ₹1 per task. One of them asserted:

> *Sometimes I feel the reward amount is just okay. You have to increase the amount for each task. ₹0.60 is not acceptable for a low-income person. We also have to consume Internet to use the app.*

While some participants perceived Internet costs to be significant, the total download size of BSpeak tasks was 189 MB and the cost was under ₹50; less than 2.5% of the total amount BSpeak users could earn by doing tasks. Several users self-reported receiving instrumental benefits such as improved listening skills (N=13), pronunciation (N=11), and concentration (N=3). They found BSpeak a *"tool for speech therapy"* and appreciated its ability to introduce them to new accents. Two users reported improvements in their vocabulary by finding the meaning of unfamiliar words they encountered. Another two users indicated that BSpeak had increased their knowledge about *"current affairs and new subjects."* Some users stumbled on topics they never explored earlier such as technology and rural inventions, and prison reform movements. Three users appreciated that they could listen to useful content in their free time while earning money. One of them stated:

> *The app improves listening skills, concentration, and pronunciation of difficult words. I listened to old interviews of great personalities like Kiran Bedi, Modi, and Kalam, all while I was earning during my leisure time.*

Amount earned (₹) $\leq$	BSpeak users
300	16
600	2
900	4
1200	2

Table 4.16: Distribution of the amount earned by BSpeak users.

User Feedback and Preferences

Sixteen users preferred to use the BSpeak app at home, four during their commute, and two during office breaks. Most users were active during the evening and at night. Similarly, the usage was much higher on weekends. We also asked users to report constraints that impeded their app use. Five users struggled with the availability of the Internet, five others had unexpected responsibilities at work, two had to travel, and one had to attend a family wedding.

All users commended the accessibility features of BSpeak, and found it easy to use and navigate. Eighteen users reported the ability to transcribe files through speech as BSpeak's key strength since this feature was easy to use, saved time, and yielded transcripts without spelling errors. Seven users liked that the tasks were not timed and that they could listen to audio files more than once. Five users appreciated the ability to use the app whenever they had time. Five other users found BSpeak entertaining to use. One of them exclaimed:

> *I felt like I was playing a game! The freedom to listen to the tasks multiple times helped me to understand fast speech and unclear words. The instant payment also motivated me to complete more tasks.*

Some users faced challenges in using BSpeak. Six users found it difficult to get their speech recognized on multiple occasions, especially with homophones. For example, a user stated that even

after five attempts, 'phase' was transcribed as 'face'. Eight users found it difficult to remember long sentences that began with a clipped word or contained unfamiliar words.

Users had several recommendations to improve BSpeak. Eight users suggested providing a function to edit a transcript through typing after a predefined number of unsuccessful speaking trials. Users also suggested providing a detailed review of the mistakes they had made during previous tasks. They recommended introducing new playback features such as rewind, forward, and manipulate playback speed to improve segment retention. They proposed to include the ability to select tasks based on content type. Some audio segments were blank, had excess noise, or had people clapping. Users were unsure what to re-speak in such scenarios. For example, users recorded *"clapping," "claps," "applause," "noise,"* and *"nothing"* for a segment containing applause. Users recommended adding a feature to report such task anomalies.

Comparison of Blind and Sighted Users

The BSpeak app is an accessible version of the Respeak app with the same underlying system components and ASR engine. Since a subset of BSpeak tasks were the ones used in Respeak's deployment, we compared the performance of blind users on the tasks that were completed by sighted users in Respeak's deployment.

Table 4.17 compares BSpeak's and Respeak's deployments. Although the duration of BSpeak's deployment was half that of Respeak's deployment, blind users completed over three times more tasks, spent over five times more time on the app, and earned 2.5 times more money than the Respeak users. Though blind users were more enthusiastic in doing crowdwork, sighted users generated transcripts with a lower WER than blind users. As a result, the expected payout per hour for blind users was almost half that of the payout for sighted users.

To compare the improvements yielded from MSA and a majority voting process, we used the same experimental setup to align transcripts generated by K users for segments common to both deploy-

	Respeak	BSpeak
Duration	1 month	15 days
Total users	25	24
Unique tasks	756	2,560
Tasks done	5,464	16,005
WER	23.7%	31.2%
Amount earned	₹3,036	₹7,310
Time spent	40 hours	207 hours

Table 4.17: Comparison of Respeak and BSpeak on different deployment parameters.

ments. Table 4.18 shows that for all values of K, the transcripts generated by sighted users in both Hindi and Indian English yielded a lower WER than the blind users. The differences in WERs of sighted and blind users drastically reduced when transcripts by more users were aligned. The WERs obtained after alignment of transcripts generated by 11 BSpeak users were comparable to the WERs obtained after alignment of transcripts generated by five and seven Respeak users for Hindi and Indian English, respectively. We also analyzed performance of blind and sighted users on different content types. Both sighted and blind users performed poorly in transcribing interviews and TV advertisements. However, blind users had higher success in transcribing news, lectures, and phone calls. In contrast, Respeak users found songs and YouTube programs easiest. In both deployments, the performance of users on speeches was average.

Disparity in education, socioeconomic status, and language skills contributed to differences in the WER yielded by blind and sighted users. The analysis of demographic information of blind and sighted users revealed that the average education level of sighted users (15.7 years) was higher than the blind users (14.2 years). A Mann-Whitney U test indicated a significant difference in education level of sighted and blind users (U=173.5, Z=-3.023, p=.002). While only one blind user had a master's degree, six sighted users had a master's degree and one was pursuing a Ph.D. Moreover, 20

K	BSpeak WER (%)			Respeak WER (%)		
	Overall	English	Hindi	Overall	English	Hindi
1	31.2	33.7	30.1	21.9	26.2	18.1
3	23.7	26.3	22.6	12.5	18.1	10
5	20.4	23.7	19	10.6	15.2	8.6
7	18.7	21	12.1	10.3	12.3	7.4
9	16.4	17.1	11.3	9.9	11.8	7
11	12.9	13	8.3	9.6	11.4	6.9

Table 4.18: Comparison of BSpeak's and Respeak's WERs obtained after alignment of transcripts generated by K users.

sighted users had a professional degree compared to only five blind users. Almost half of the blind users were from rural backgrounds while all sighted users were from peri-urban or urban regions. The average monthly family income of blind users was only one-fourth of that of the sighted users. Moreover, while sighted users rated their English and Hindi speaking skills highly, a majority of blind users reported struggling with both English and Hindi.

In the next section, we describe the details of the field deployment of the ReCall app—an IVR version of the Respeak smartphone app—with low-income rural residents in India.

4.7 ReCall Field Deployment

We conducted a two-week field deployment with 24 low-income rural residents to examine three key questions regarding ReCall's feasibility and acceptability:

1. Would ReCall users produce Hindi transcripts with a decent accuracy and lesser cost than the market rate?

2. Would users gain financial benefits by using ReCall?

3. Would ReCall generate enough profits to provide free airtime to users on another voice forum?

4.7.1 Methodology to Evaluate Deployment

Out of the 28 rural residents who participated in experimental and usability evaluations comparing ReCall and Respeak, 24 (14 female and 10 male) expressed their interest in using ReCall for two weeks in their free time. We informed them that our goal is to investigate the feasibility of ReCall in providing additional earning opportunities to people in rural areas, and that we do not have any immediate plans to scale the service. During an hour-long group orientation session, we demonstrated the ReCall app to users and answered their queries. At the end of the deployment, we conducted semi-structured interviews to examine the benefits ReCall users received and challenges they encountered in transcribing audio files vocally.

We also conducted a usability study with ten randomly selected ReCall users (six female and four male). We requested them to use Sangeet Swara, a social media voice forum described in Chapter 3, for 15 minutes. On calling Sangeet Swara, participants could record audio messages and listen to messages recorded by others. We gave participants a five-minute airtime credit to use the service. When participants consumed their allotted airtime, they were served ReCall tasks and could use Sangeet Swara only after completing the tasks. We asked participants questions on how integration of ReCall and Sangeet Swara affected their usability and user experience on Sangeet Swara.

We quantitatively analyzed transcription accuracy, users' earnings, transcription cost, and prospects to financially sustain voice forums. This analysis was complemented with qualitative analysis of interviews that we conducted after the deployment and usability study. Participants' responses were subjected to thematic analysis as outlined in [65]. The field deployment and usability study was approved by our institution's IRB.

4.7.2 Tasks, Rewards, and Payments

We selected 21 Hindi files, containing nearly three hours of audio content, for the deployment. Out of these, 13 audio files were the same as those used in Respeak's deployment because we wanted to compare crowd work performance of rural ReCall users to urban Respeak users. We selected only Hindi audio files for transcription since most ReCall users had poor English language skills. The engine segmented 21 files to produce 2,063 audio micro tasks. These tasks represented a wide variety of content including news, poems, songs, speeches, telephone calls, and television programs.

To ensure that the earning potential of ReCall users equaled that of Respeak users, we used Respeak's reward structure. The maximum amount a ReCall user could earn was ₹2078 (USD 31.50). Since the majority of ReCall users (80%) did not use mobile wallets, we offered to pay their earnings via mobile airtime. However, most users preferred to receive a cash transfer at the end of the deployment.

4.7.3 Findings

Low-income rural residents enthusiastically used ReCall to vocally transcribe Hindi segments. During the two-week deployment, 24 users placed 5,879 phone calls to complete 2,063 tasks nearly 29,000 times with an average accuracy of 73.3%, and earned ₹20,500 (USD 310) by transcribing segments. The average duration of phone calls was 9.5 minutes (SD=13.7 minutes). The median task completion time was 75 seconds. The engine combined the transcripts generated by five users to yield a transcription with 82% accuracy and by eleven users to yield a transcription with 85% accuracy.

Figure 4.15 shows that users enthusiastically used the ReCall app until we turned off the service at noon of day 17. The majority of users (80%) regularly used ReCall. For example, 16 users completed more than 1000 tasks, 2 users completed more than 500 tasks, and the rest completed less than 30 tasks. With respect to the call flow shown in Figure 4.4, participants spent 2.2% of the total time on the home menu, 82.7% on the task menu, 1.3% on checking accuracy and earnings, and 0.04% in

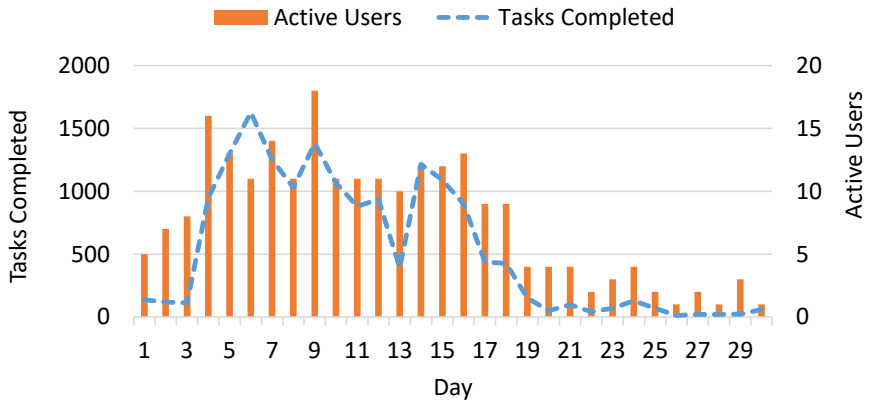

Figure 4.15: The number of tasks completed and active ReCall users for the deployment duration.

	Deployment length	Total users	Unique tasks	Tasks completed	Accuracy on common tasks	Amount earned	Median task time	Earning potential
ReCall	15 days	24	2063	28,885	71.4%	₹20,500	75s	₹36 per hour
Respeak	1 month	25	756	5,464	76.3%	₹3,036	36s	₹76 per hour

Table 4.19: Comparison of ReCall's use by low-income rural residents and Respeak's use by low-income metropolitan residents.

learning about reward calculations. The remaining time was spent in other activities like navigating between the pages and fetching segments from a remote server.

Table 4.19 compares the use of ReCall by low-income rural residents to the use of Respeak by low-income metropolitan residents. Compared to Respeak users, ReCall users completed five times more tasks and earned about seven times more money in just half of the deployment duration. However, ReCall users produced transcripts in double the time and with 7% lower accuracy than Respeak users. As a result, the expected payout per hour for ReCall users was almost half of the payout for Respeak users.

Although ReCall and Respeak users were comparable in age, they had several demographic differences. For example, ReCall users were poorer and lesser educated than Respeak users. While

all Respeak users owned a smartphone, the smartphone penetration among ReCall users was 54%. ReCall users were living in remote rural areas, whereas Respeak users were metropolitan residents. Despite these demographic differences, when we compared the two systems, we found that people in rural areas enthusiastically used ReCall even when they scored it lower on task performance and found it less usable than Respeak. These results indicate a strong appetite for crowd work and additional earning opportunities in rural areas. We now address the three questions outlined previously to examine ReCall's feasibility and acceptability to financially sustain voice forums.

Speech Transcription Accuracy

ReCall users produced transcripts with an average individual WER of 26.7%. To reduce random speech recognition errors in transcripts, the engine used multiple string alignment (MSA) and a majority voting process to merge transcripts produced by multiple users. We ran a series of experiments to examine how using more transcripts (K) in the merging process affect transcription WERs. For each value of K, we conducted five runs of the experiment. In an experimental run, for each segment, we randomly selected K transcripts and merged them to obtain a best estimation transcript. We computed the WER of the best estimation transcript by comparing it to the ground truth. We averaged the WERs obtained in five runs of the experiment for each segment. We then computed a weighted WER for a value of K by using the averaged WER for each segment. We used this experimental setup to align transcripts generated by 3, 5, 7, 9, and 11 users.

The engine aligned transcripts generated by 11 users to produce transcripts with an accuracy of 85%, indicating large improvements in accuracy via crowdsourcing (see Table 4.20). Although transcription accuracy increased with an increase in the value of K, the comparative improvements in the accuracy were more significant for smaller values of K. Similar to Respeak and BSpeak deployments, we found that the average accuracy of transcripts obtained by submitting raw segments directly to the Google Cloud Speech API [41] was 53%, compared to 73.6% accuracy when users re-spoke these segments into the ASR engine.

Content type	Unique tasks	Length in mins	Transcription accuracy after merging					
			K=1	K=3	K=5	K=7	K=9	K=11
News	17	2	15.9	11.0	8.9	7.4	6.3	6.9
TV programs	54	12	28.1	20.4	17.2	15.7	15.5	13
Phone calls	38	4	25.5	19.3	16.1	15.9	15.4	13.5
Speeches	1,738	148	25.8	19.5	17.5	16.2	14.7	13.7
Songs	77	1	32.8	23.4	18.6	18.6	17.7	16.9
Poems	139	7	35.2	28.1	25.4	19.5	18.6	17
Overall	**2,063**	**173**	**26.7**	**20.3**	**18.1**	**17.5**	**17.1**	**15.4**

Table 4.20: WERs obtained after aligning transcripts generated by K users for each content type.

ReCall users found news segments easiest to transcribe since these segments had a clear diction and pronunciation. Users' performance on TV programs, speeches, and phone calls was relatively lower than news due to background noises in speeches, multiple speakers in TV programs, and unfamiliar accent in some phone calls. Several users faced difficulties in transcribing songs. The engine segmented songs in five-second chunks because of the challenges in detecting natural pauses due to the presence of background notes. As a result, some song segments started or ended abruptly, confusing users whether to repeat cut-off words or ignore them. We also noticed that some participants sang these segments instead of repeating the content, leading to poor detection from the ASR engine. Similarly, many users transcribed poems poorly because of the challenges in understanding formal words and diction in these segments.

Earnings and Rewards from Crowd Work

ReCall users collectively earned ₹20,500 (USD 310) by transcribing audio segments. Five users earned more than ₹1,500 and ten users earned more than ₹1,000. The maximum amount a user earned was ₹1,700 (USD 26). The expected payout per hour of using ReCall—calculated based on the expected number of tasks users could do in an hour (48 tasks) and the expected payout for each

task (₹0.74)—was ₹36, comparable to the average hourly wage rate in India [14]. This indicates that even if low-income rural people use ReCall for just an hour a day, they would earn more than 75% of rural residents in India who live on less than ₹33 per day [159]. In fact, during our deployment, ReCall users earned an average of ₹57 per day by performing crowd work in their free time.

Several participants appreciated the prospects of ReCall to supplement their income. Most of them did not see ReCall as a substitute for a full-time employment, instead they perceived it as a useful app for *"part-time work"* which they can use a few hours a day to pay for their daily expenses *"like buying clothes, mobile airtime, and fruits and vegetables."* Many users found it rewarding that their older family members could also use ReCall and potentially supplement the family income without *"toiling in the fields."* Several users also reported improving their pronunciation and gaining access to new information by using ReCall. A user shared how ReCall could benefit rural residents engaged in manual labor:

> *"Several people in our village work 9–10 hours a day to earn ₹2000–2500 per month. These laborers and rickshaw pullers can increase their income by using ReCall for 2 hours daily to easily earn ₹3,000 per month. ReCall can provide them information, exposure, independence, and confidence."*

Our findings indicate that ReCall offered sufficient financial and instrumental benefits to low-income rural residents to keep them engaged in crowd work.

Transcription Costs and Financial Sustainability

ReCall has two main cost components: the monetary rewards disbursed to users for completing tasks, and the airtime costs incurred by ReCall users for completing tasks.

Reward costs: The earnings disbursed to users for transcribing a minute of audio content is based on the expected number of tasks (i.e., segments) in one minute of audio content, the expected amount

earned by users for completing one task, and the number of transcripts used in MSA and a majority voting process (K). The expected number of segments in a minute of audio content are $\frac{60}{len}$ where len is the average segment length in seconds. The expected amount users' earn for completing a task is based on the expected accuracy with which they complete the task ($accuracy_{exp}$) and the expected value of the maximum reward amount for the task ($reward_{exp}$). The reward costs per minute of speech transcription is thus calculated as

$$cost_{rewards} = \frac{60}{len} * accuracy_{exp} * reward_{exp} * K$$

In our deployment, the average segment length was 5.03 seconds, the expected transcription accuracy was 73.6%, and the average reward amount was ₹1.01. We used transcripts from 11 users in the merging process. Based on these deployment numbers, the reward costs for transcribing one minute of audio content was USD 1.46.

Airtime costs: The last two years have seen major disruptions in India's telecom industry due to the entry of Reliance Jio, an MNO that has significantly reduced voice call rates to gain new subscribers [22, 90, 92]. Following suit, all MNOs now offer STVs that provide more affordable or even free voice calls in India [20, 28, 29, 36]. As a result, the average cost of voice calls has reduced from ₹0.49 per minute to ₹0.16 per minute since March 2016 [59].

We use two models to compute the airtime costs incurred by ReCall users. In the first model, we assume that ReCall users pay regular call rates to use the ReCall app. In the second model, we assume that ReCall users use an STV to get unlimited free voice calls. The airtime costs for transcribing a minute of audio content is based on the expected number of segments in a minute of audio content ($\frac{60}{len}$), the number of minutes users take to complete one task (N_{mins}), the per minute cost of voice calls ($cost_{call}$), and the number of transcripts used in the merging process (K). The airtime costs per minute of speech transcription is thus calculated as

$$cost_{airtime} = \frac{60}{len} * N_{mins} * cost_{call} * K$$

In our deployment, the median task completion time was 1.25 minutes and the average segment

length was 5.03 seconds. Since the regular call rates in India is ₹0.60 per minute, the airtime costs for 11 users to transcribe one minute of audio content was USD 1.49. When considering the average cost of voice calls in India (i.e.,₹0.16 per minute [59]) instead of the regular call rate, the airtime costs came out to be USD 0.40.

At the beginning of the deployment, we spent ₹1,634 to buy STVs that offer unlimited free voice calls for 16 users who were not already using these STVs. In the second model, the per minute cost of voice calls is ₹0.03 per minute, calculated by dividing the total call duration (54,600 minutes) into the total cost of buying these STVs. Thus, the airtime costs for 11 users to transcribe one minute of audio content was USD 0.07 in the second model.

Market Cost of Hindi Transcription: To gain an understanding of the existing market rates for Hindi audio transcription, we conducted a survey of 12 organizations that we found via web search queries, such as 'Hindi transcription services', 'Hindi transcription India', and 'Indian language transcription', among others. Out of these 12 organizations, eight sent us a quote, which were (in USD per minute) 7, 5.25, 5.25, 5.25, 5, 4, 3.15, 0.25, and 0.15. The two lowest quotes were from organizations that provided an interactive editor so that requesters can remove errors themselves in transcripts obtained by submitting raw audio files directly into the ASR engine. Since these organizations relied on requesters to remove a majority of transcription errors, we excluded them from our analysis, yielding the average market cost of Hindi audio transcription as USD 4.99 per minute.

Financial Sustainability: For ReCall to be financially sustainable, the reward costs and airtime costs must be less than the market cost. Based on the average call rate in India, ReCall's per minute cost of Hindi transcription was USD 1.86 per minute. Since the average market cost of Hindi transcription is USD 4.99 per minute, ReCall earned profits at the rate of USD 3.13 per minute of speech transcription. These profits when equally distributed between 11 users provide each of them with nearly ₹19 (equivalent to 117 airtime minutes) for transcribing one minute of audio content. Since ReCall users on average transcribed a minute of audio content in 15 minutes ($N_{mins} * \frac{60}{len}$), each minute of crowd work on ReCall gives them 7.8 minutes of free airtime on another voice forum. In

K	$cost_{rewards}$ (USD per min)	$cost_{airtime}$ (USD per min)			Total Cost = $cost_{rewards}$ + $cost_{airtime}$			Airtime received on another voice forum by 1 minute of crowd work on ReCall		
		$cost_{call}$=0.03	$cost_{call}$ = 0.16	$cost_{call}$ = 0.60	$cost_{call}$=0.03	$cost_{call}$ = 0.16	$cost_{call}$ = 0.60	$cost_{call}$=0.03	$cost_{call}$= 0.16	$cost_{call}$ = 0.60
1	0.13	0.01	0.04	0.14	0.14	0.17	0.27	711.3	132.6	34.6
3	0.40	0.02	0.11	0.41	0.42	0.51	0.81	223.4	41.1	10.2
5	0.67	0.03	0.18	0.68	0.70	0.85	1.34	125.9	22.8	5.3
7	0.93	0.05	0.25	0.95	0.98	1.19	1.88	84	14.9	3.3
9	1.20	0.06	0.33	1.22	1.26	1.52	2.42	60.8	10.6	2.1
11	1.47	0.07	0.40	1.49	1.54	1.86	2.96	46	7.8	1.4

Table 4.21: ReCall's cost of transcription (in USD per minute) for different values of K and voice call rates ($call_{cost}$ in ₹ per minute).

the first model when users pay a regular call rate of ₹0.60 per minute to use ReCall, each minute of crowd work on ReCall gives them 1.4 minutes of free airtime credits. In the second model when ReCall users have STVs, each minute of crowd work on ReCall gives them 46 minutes of free airtime credits. Table 4.21 shows the transcription cost of ReCall for different values of call rates and the number of transcripts used in the merging process (K).

Our usability evaluations with ten participants who completed tasks on ReCall to subsidize their participation costs on Sangeet Swara revealed promising results. All users completed at least two tasks on ReCall to use Sangeet Swara after their free credits expired. While a few participants complained about the context switch between Sangeet Swara and ReCall, the majority (N=7) were comfortable in switching between the two services to earn free airtime for using Sangeet Swara. Our participants also provided useful insights about how ReCall could be integrated with other voice forums. Five participants suggested that users should be allowed to do more tasks in one go to minimize the context switch. Similarly, three participants suggested that ReCall should announce the amount of free airtime a user has earned on Sangeet Swara by completing tasks on ReCall. Two participants suggested that users should decide how much money they will receive as earnings and how much would be used to provide them free airtime credits.

4.8 Discussion and Conclusion

In this chapter, we examined if profits generated from crowd work by rural residents can be used to financially sustain voice forums. We employed assets-based approach [112] to design a crowdsourcing marketplace for people in low-resource environments by leveraging their skills and the resources available to them. In doing so, we overcame three significant barriers to democratizing crowd work to voice forums users who experience literacy, language, socioeconomic, and connectivity barriers: (1) since most users do not have access to smartphones, we leveraged the ubiquity of basic phones, (2) since most users do not have access to the Internet, we leveraged the availability of phone calls, and (3) since most users have low literacy skills, we leveraged the power of voice, a natural and accessible communication medium.

We conducted several cognitive experiments, usability studies, experimental evaluations, and field deployments to rigorously examine the prospects of crowd work by voice forum users to subsidize their participation costs. Our findings revealed three key results with respect to ReCall's feasibility, usability, and acceptability. First, we found that low-income users enthusiastically transcribed audio content vocally with a satisfactory accuracy and at an optimal cost. Second, they supplemented their earnings at a rate higher than the average hourly wage rate in India by engaging in crowd work. Third, the profits earned by completing one minute of crowd work on ReCall provided users eight minutes of free airtime on another voice forum, addressing the financial sustainability challenge of voice forums.

Our work has implications beyond financially sustaining voice forums. For example, one of the most direct ways to empower low-income communities in resource-constrained settings is to provide them with additional earning opportunities. Our work on Respeak provides a definite step forward in realizing a smartphone- and voice-based crowdsourcing marketplace. Even technology-savvy university students found the amount they earned by using Respeak appealing. Respeak has potential to be transformative for marginalized communities—like low-literate people and people with visual impairments—due to its voice-based implementation. For example, it is well-documented

that blind people experience huge barriers in finding full- and part-time employment [66, 67, 72]. While crowdsourcing marketplaces like MTurk have provided additional earning opportunities to over half a million people across the globe [18], severe accessibility barriers impede the use of such platforms by blind people [176]. Our work on BSpeak demonstrates that a simple user interface, use of voice input, and untimed tasks could make a crowdsourcing marketplace more accessible to low-income blind people in low-resource environments. Similarly, ReCall can be used only to provide additional earning opportunities to people without smartphones and Internet connectivity. In its current form, ReCall disburse a portion of its profits as earnings to users and another to provide free airtime credits to them. If ReCall is used only to supplement income of users, all profits can be disbursed to them at a rate three times higher than the average hourly wage rate in India.

Although our work is a promising first step to demonstrate the feasibility, usability, and acceptability of crowdsourcing marketplaces designed for people in low-resource environments, much more is needed to examine whether these marketplaces provides a fair, collaborative, and sustainable experience to its users. For example, can ReCall match the standards of a crowd workplace in which we would want our children to participate [97]? Can it enable users to have the agency to protect their rights, increase their wages, or improve their working conditions? How can it encourage workers to collaborate rather than compete? Can users reject tasks that they find offensive without being penalized? Future work should investigate these questions as well as examine how ReCall can fulfill the criteria suggested by the Fairwork Foundation to create fair digital work opportunities [85].

At the very least, future work could explore ways to increase the payout to users. There are several promising directions, such as by improving the accuracy with which users complete tasks, by decreasing the time taken by them to complete tasks, and by raising the rewards offered for completing tasks. To improve the accuracy, future work could incorporate a functionality to edit a transcript after multiple unsuccessful speaking attempts. Future iterations could also include tasks in other local languages, such as Kannada and Marathi, by integrating local language speech recognition APIs like [23] for improved performance. To minimize errors due to mishearing of transcripts, the app could either automatically reduce the playback speed of audio files containing unclear speeches

or high speaking rate. To decrease the task completion time, future iterations could automatically skip tasks after a pre-defined number of unsuccessful speaking trials. Sending such difficult tasks to expert users could improve accuracy and reduce task completion time. Also, since the average industry transcription cost for audio files in local languages and accents is nearly USD 5 per minute [43, 45, 49], there is a scope to quadruple the payout while keeping the transcription cost below the industry standard. Though this calculation does not account for server, maintenance, and personnel costs, it could increase the payout up to ₹144 per hour for Respeak and BSpeak users. In addition, experimenting with ASR word lattices to reduce WER [81], and sending a segment to more users only when the transcripts generated by the initial set vary over a threshold could lead to significant reductions in transcription cost, yielding spillover benefits to users.

Future work could also explore the opportunities to further improve ReCall. For exaple, the median task completion time had an inverse impact on the reward costs and direct impact on the airtime costs. Since ReCall users spent about 45% of their time listening to IVR prompts, using shorter yet meaningful prompts for experienced ReCall users could reduce the task completion time significantly. For example, while verifying the correctness of the transcript generated by the ASR engine, experienced ReCall users could be presented with a prompt *"To submit the task, press 1. To do the task again, press 2."* instead of *"Is the audio transcript similar to the content in the audio task? If yes, to submit the task, press 1. If no, to do the task again, press 2."* Similarly, ReCall users spent 11.5 hours of airtime in checking their accuracy and earnings in 2,631 calls. Since text messages are lower priced than voice calls in India, sending the information about user's accuracy and earnings as a text message to literate ReCall users could reduce airtime costs. We observed that ReCall users re-spoke about 40% segments more than once because they were unsatisfied with the ASR-generated transcripts in their initial attempts. Interestingly, in many cases, there was no difference between the transcript generated in the penultimate attempt and the last attempt. This happened because several users struggled to understand certain words spoken by the TTS system due to its unclear diction and mechanical voice. Future work could focus on improving the diction of TTS systems for Indic languages as well as evaluating the effect of different TTS systems on ReCall users' task

accuracy and completion time.

Respeak, BSpeak, and ReCall users were primarily driven to use the app for earning mobile airtime. Some users found these apps to be monotonous and less enjoyable towards the end of the deployment. To make it more interesting, many users created informal leaderboards to compete with each other on accuracy of speech transcription and the number of tasks they completed. They conducted these discussions over face-to-face interactions, emails, and WhatsApp groups. Future work could use gamification to increase user retention and entertainment value. Several users reported receiving instrumental benefits, such as improved vocabulary and pronunciation skills, access to new information and knowledge, and a new-found interest in content. Though we did not have any quantitative measure of these indicators, future work could use language learning aspects to re-design and evaluate these systems.

Our work has some limitations as well. For example, the engine did not distinguish speakers in transcription for audio files with multiple speakers. Future versions could consider an improved segmentation scheme that is cognizant of speakers in an audio file containing multiple speakers. Further, the transcription generated by ASR engine lacked punctuation marks. Though punctuation marks could be added automatically based on the identification and length of natural pauses, a better algorithm is needed when it would be difficult to detect natural pauses due to ambient noise. One possible solution could be to send an audio segment and corresponding transcript generated by the engine to users, who are then asked to identify speakers and place punctuation marks.

Since several MNOs provide STVs that provide free voice calls, is there still a need of ReCall to subsidize participation costs of voice forums? Half of our participants did not know specific details of these STVs and nearly two-thirds did not use them. Our interviews and observations indicated several reasons for the limited use of STVs in rural areas. STVs are often offered only in selected circles and to selected consumers, and often the plan details keeps changing. As a result, people in rural areas have to visit local mobile phone shops to know offers available to them. Even phone shop owners have to make multiple calls to verify whether an STV would work on a specific phone

number, indicating variations in STVs based on SIM cards. Several participants also reported that these STVs are used by male family members, indicating that women face discrimination in using these STVs. We argue that ReCall has value both for STV non-users as well as STV users. While ReCall could provide income as well as subsidized airtime to non-STV users, STV users could receive the full portion of their profits on ReCall as earnings. An hour of crowd work on ReCall will then enable STV users to earn ₹120, more than three times the average hourly wage rate in India. If the process to discover available STVs becomes easier in future, ReCall could first use the profits to give users STVs so that they can freely access any voice forum, and then use the remaining profits entirely to supplement their earnings.

Our preliminary usability study indicated willingness of low-income rural people to complete ReCall tasks for earning free airtime to use another voice forum. We found that participants perceived context switch to be manageable when switching between ReCall and Sangeet Swara. Future work could examine how ReCall could be integrated seamlessly with voice forums, how many audio tasks ReCall users should complete in one go to subsidize their participation costs, and when and where in the call flow should tasks be presented to minimize users' cognitive load and disruptions in their user experience. ReCall also has a potential to financially sustain voice forums in other developing countries like Bangladesh that have affordable voice call rates (BDT 0.45 or ₹0.39 per minute [35]) and structural limitations similar to India.

Chapter 5

A TOOLKIT FOR REPLICATING VOICE FORUMS

In previous chapters, we described how challenges like managing local language audio content and high cost of voice calls makes it difficult to run voice forums. While these limitations significantly impede how voice forums scale and sustain, building these services is often the first-order concern of many non-profits and non-governmental organizations (NGOs) that lack necessary software development skills and expertise. As a result, it is very hard for them to build voice forums and replicate them in new contexts.

In addition, most voice forums operate in silos; for example, a voice forum may connect farmers in one county, but not in others; it may facilitate intra-community conversations, but not inter-community dialogues; it may include one stakeholder (e.g., farmers with basic phones) while excluding others (e.g., agricultural experts with Internet access). Although voice forums have succeeded in fostering communication within a community, unlike mainstream social media platforms, they are still far away from giving a global reach to local voices.

This chapter presents IVR Junction: a free and open-source toolkit that enable organizations to build and replicate voice forums [168]. IVR Junction has three main advantages over existing IVR toolkits—like Asterisk, FreedomFone, FreeSwitch—used for building voice forums:

- **Easy to build and set up**: IVR Junction makes it easier for organizations with limited technical skills to build, set up, and maintain voice forums. Using IVR Junction, anyone with basic computer literacy can use templates and configure simple options to set up a robust voice

forum as an ordinary program on a Window-based commodity machine.

- **Distributed architecture**: IVR Junction enables distributed access points, thereby connecting multiple geographically distributed communities via inexpensive local calls as well as enabling robustness to regional power outages or crackdowns by repressive regimes.

- **Global reach**: IVR Junction integrates voice forums with free Internet services and social media platforms: recordings contributed over the phone are immediately broadcast on YouTube and Facebook, and posts made on the Internet can also be listened to over the phone. Thus, IVR Junction enables anyone with a basic mobile phone to participate in global social media; low-income populations can record and listen to posts via mobile phone, while the global community can access and contribute recordings via the Internet. This capability enables remote communities to create their own repositories of highly-relevant information, while also sharing them with audiences worldwide.

In this chapter, our primary contribution is the design and implementation of IVR Junction. In the following sections, we describe the architecture of IVR Junction, and show how several governmental agencies, social enterprises, and grassroots entities used IVR Junction to connect people in low-resource settings in South Asia and Africa.

5.1 IVR Junction 's Architecture and Features

Contrary to existing offerings that rely on Asterisk or FreeSWITCH, IVR Junction is easy to configure on an off-the-shelf Windows machine. It is also scalable, as it utilizes distributed nodes synchronized via the cloud to enable international participation at low calling costs. Figure 5.1 depicts the overall architecture of the system. To host a voice forum, an organization needs to purchase only three pieces of hardware: a laptop (or desktop) computer, a GSM (or landline) modem, and a GSM SIM card (or landline connection). This hardware constitutes a telephony access point which services phone calls from users. In addition to the physical servers, the organization has to establish

accounts with free cloud-based services—like YouTube (or SoundCloud), OneDrive (or Dropbox), Facebook (or Twitter)—in order to provide storage and moderation of audio messages. These linkages are needed only if the service provider wants to integrate IVR Junction with Internet services. YouTube or SoundCloud is used as free content hosting platform and content moderation platform. Dropbox or SkyDrive is used as free cloud storage portal. Facebook or Twitter are used as social media outlets of the voice forum. Though other tools like FreedomFone, Asterisk, FreeSwtich also facilitates creation of voice forums, they are quite complex to install, configure and maintain since they are based on Linux operating system and require expertise that is usually beyond the reach of many non-profits and NGOs. Also, they require service providers to set up their own web server to host, store, and manage audio messages.

Some NGOs have geographically distributed branches. Although the branches are in different geographical units, it may be desirable that branches share their audio repository as users (such as farmers) in both regions speak same local language, share similar local constraints and geographical features, and have comparable standards of living. Thus, users would benefit from listening to audio repositories of both branches. As the branches are in different states, having a common phone number for both branches will require users in one state to dial a long-distance number. Hence, though the information is accessible to users, the long-distance call would discourage them from using the service. In IVR Junction, the NGO's branches can choose to configure automatic synchronization of their audio repositories to facilitate increased knowledge sharing.

The overall functionality of the system can be best understood via a usage scenario. Consider a Q&A forum, in which callers can record audio questions as well respond to the questions that others have posted. A user interacts with the system by calling the local Indian NGO – say, in Jaipur, Rajasthan (see Figure 5.1). The call is processed locally via the GSM modem and laptop, using a local audio repository that was previously synchronized with the cloud. When the call finishes, the server uploads the new recording (a new question) to the central server in the cloud. At this point, the question awaits approval by the moderator, who logs in via the Internet to listen to the question, categorizes it, and perhaps summarize in textual form (for the benefit of Internet-based

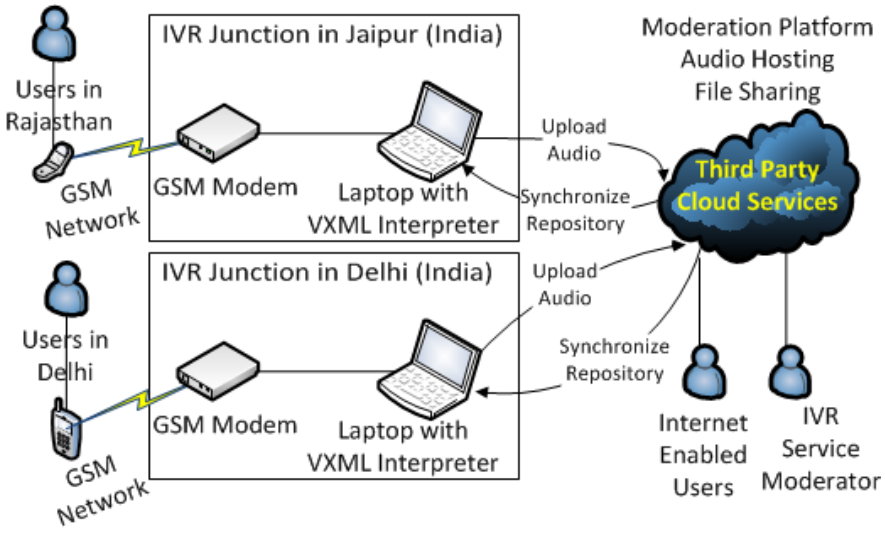

Figure 5.1: IVR Junction's system architecture.

users). Following approval by the moderator, the question becomes live on a website, making it accessible to Internet users around the world. Also, at this point in time, the question automatically becomes visible to (i.e., appears in the local audio repository of) other branches of the NGO.

For example, the server in Delhi would detect that a new question is available, download that question into its local cache, and play that question when requested by future callers in Delhi. If a caller in Delhi responds to the question, the response is later synchronized with the central server and upon moderator's approval, becomes visible to Internet users and voice forum users, and thus to the original questioner who is in Jaipur. By leveraging the distributed architecture, IVR Junction can also be used to connect people in multiple countries, something that is very difficult and expensive to execute currently. For example, in order to connect people in India, Pakistan, and Bangladesh, a voice forum provider needs to set up one IVR Junction node in each of the three countries. Callers in one country (say India) could make a local phone call to record their message and listen to messages recorded by users in all three countries. This functionality is not yet available in any other IVR toolkit.

In developing regions, many NGOs have offices and operations teams in towns which have intermittent availability of electricity and Internet connectivity. While the GSM modem requires an electrical outlet, the remainder of IVR Junction is tolerant to intermittent power and Internet outages. IVR Junction is designed to run on a low-end laptop. The battery backup of a laptop makes it tolerant to intermittent power outages, while the low cost, ease of deployment and mobility of the laptop enables non-expert individuals to set up voice forums by themselves. The modems are low-powered and can work with an external battery or solar panel in case of power outages. Finally, the system does not require continuous Internet connectivity at any node. Each local repository is periodically synchronized with the cloud in an opportunistic manner, depending on the available connectivity.

The access points of IVR Junction can be easily scaled up to support many parallel callers. GSM modems can be added incrementally, with calls forwarded between numbers that are busy. This feature, known as "call or line hunting", is commonly available in developing regions. Also, all hardware components required to run IVR Junction servers—such as laptop, modems, SIM cards, external battery—can easily fit into a backpack. This provides an excellent opportunity to change the location and set it up again in case of threat of natural disaster or physical crackdown by a repressive regime.

5.2 IVR Junction Deployments

In the last few years, IVR Junction has been used by many organizations to connect people in low-resource environments and provide them access to information, news, and governance. For example, in Somaliland, IVR Junction was used to build a voice forum that established a direct communication channel between the rural tribal population and government officials to bring transparency and trust in the political processes. Somaliland—an autonomous region of Somalia—has fragile political institutions, fragmented and polarized media, and unstable government. Parliamentarians in the capital city were unable to convey their policies and receive feedback from low-literate con-

stituents living in remote, rural, disconnected regions due to misinformation by partisan media. The solution to this intractable problem appeared simple: connect parliamentarians directly with their communities. However, Facebook and Twitter were infeasible solutions in a region with less than 5% Internet penetration and 37% adult literacy rate. To overcome these challenges, IVR Junction was used to build Ila Dhageyso, a voice forum that enabled parliamentarians and constituents to call a phone number, and record and listen to asynchronous audio messages in a discussion forum format. Ila Dhageyso also automatically posted these audio messages to a YouTube channel to engage with Somaliland's diaspora. The voice forum was supported by the Office of the Communication of the President and Telesom (largest telecommunication company in Somaliland), and was launched as a toll-free line so that people living in poverty could contribute and access audio messages. The deployment received an enthusiastic response both from the constituents and parliamentarians who recorded over 4,300 audio messages in just five months [87].

In war-torn Mali, the Broadcasting Board of Governors and Voice of America used IVR Junction to provide on-demand, reliable, and up-to-date news in the local language. People in Mali called the service to listen to three-minute news broadcast by Voice of America, thereby getting access to breaking news and health information as well as sharing theirfeedback.

In India, IVR Junction was used by women's rights activists in response to a gang rape incident in New Delhi that sparked international outrage. They built a voice petition forum where supporters from all economic backgrounds and varied literacy levels raised their voice for women's safety and empowerment. The contributions, which spanned from support for the victim, to plans for sensitizing local communities, were available not only on the voice forum but also on a YouTube channel and Facebook page.

Chapter 6

BENEFITS AND PITFALLS OF SOCIAL COMPUTING SYSTEMS

In the previous chapter, we demonstrated via deployments of IVR Junction that voice forums can address information and instrumental needs of marginalized communities in low-resource environments; a voice forum in Somaliland involved indigenous communities in governance and politics; a voice forum in Mali enabled Voice of America to obtain real-time feedback from marginalized users; a voice forum in India enabled people from all walks of life to participate in a social movement.

So far, this thesis outlined approaches to scale, sustain, and replicate voice forums. Equally important is to examine how these services are used by people in low-resource environments. For example, why do marginalized communities gravitate towards voice forums? What benefits and limitations these services offer to them? Are these services always empowering and inclusive?

This chapter aims to investigate these questions and contribute to the ongoing debate about the benefits and pitfalls of social computing. We do so by examining the case of Sangeet Swara, a social media voice forum described in Chapter 3. Sangeet Swara enabled low-income people to record, listen to, vote on, and share voice messages in local languages. In an eleven-week deployment in rural India, Sangeet Swara received over 25,000 calls and 5,000 voice messages from more than 1,500 people. The user analysis of Sangeet Swara found two unexpected usage patterns: (1) high adoption by blind people and (2) low adoption by women.

Though we did not promote Sangeet Swara on any of the channels accessible to blind community, it received impassioned usage by low-income blind people in rural India; more than 26% of its users were blind. Also, though we designed Sangeet Swara to be accessible to people in low-resource

environments, it received very little participation from low-income women; only 6% of posts were recorded by female users. In this chapter, we investigate the use and non-use of Sangeet Swara by these two user groups. In particular, we examine

1. Why participation of low-income blind people was so high on Sangeet Swara?
2. Why participation of low-income women was almost non-existent on Sangeet Swara?

Analyzing how low-income blind people and low-income women used Sangeet Swara helps us examine both its strengths as well as weaknesses. More importantly, this also helps us understand how the same service can impact two user groups differently.

In this chapter, we make the following contributions:

- We conduct the first analysis of how low-income blind people use a voice forum. We present detailed analysis of the content generated by blind Sangeet Swara users, reasons for the high adoption of the service, strengths and weaknesses of the service, and design implications for future systems [162].
- We conduct the first analysis of how men and women interacted with each other on a voice forum. In particular, we analyze what content they posted, liked, and disliked as well as what factors contributed to low-participation of women [164].
- In doing so, we also present evidence that voice forums—like most social computing technologies—may widen existing socioeconomic inequalities. They may benefit some user groups in low-resource environments while marginalizing others.

Our mixed-methods approach spanning quantitative and qualitative analyses found that blind users deeply valued their interactions with other users on the service. The analysis of call logs of 53 blind

users found that although they made up just 3.5% of all users, they contributed 25% of all posts, 24% of all playback events, 19% of all calls, and 25% of all votes. Our qualitative interviews and phone surveys indicated that blind users received several instrumental benefits, shared entertaining content, and built social capital by using Sangeet Swara.

We also found that women on Sangeet Swara faced systemic discrimination and harassment in the form of abusive, threatening, and flirty posts directed at them. Most women lacked agency to retaliate due to deep-rooted patriarchal values and most men who behaved inappropriately ganged up on those men and women who criticized their behavior. Most male users condoned unruly behavior by men and disapproved of abusive, flirty, and threatening posts less strongly than did women. These factors dissuaded women from using Sangeet Swara.

In the following sections, we describe the experience of blind people and women on Sangeet Swara in more detail. We then discuss the benefits and pitfalls of voice forums, and use an intersectional HCI lens [146, 172] to examine marginalities within marginalities in the use of voice forums. Finally, using a feminist HCI lens [56, 57], we discuss how voice forums could be redesigned to provide an equitable and inclusive platform to women.

6.1 Use of Sangeet Swara by Low-income Blind People

About 90% of the world's 285 million people with visual impairments live in low-income settings [12]. India has the largest blind population, with more than 63 million people with visual impairments [129]. The majority of them experience a wide array of barriers—like high cost of Internet-connected smartphones, difficulties in understanding the audio output of screen reader software in English, inaccessible features of existing social media platforms, and lack of training in digital skills—that impede their participation in mainstream social computing technologies [162]. This highlights the need to create new social computing technologies that are more cognizant of socioeconomic and infrastructural realities of low-income blind people.

The design elements of voice forums—e.g., reliance on toll-free calls, speech interface, and local language—make these services very appealing to people with visual impairments; over 25% of Sangeet Swara users were visually impaired [161]; nearly 68% of users of Baang—a social media voice forum deployed in Pakistan [140]—were blind people. These initial successes led to deployments of large-scale voice forums that are especially designed for people with disabilities (e.g., Namma Vaani service in India [76]). We contribute to this growing literature by presenting the first detailed account of how low-income blind people in rural and peri-urban India use voice forums. In particular, we examine:

- What content blind users produced, consumed, and shared on Sangeet Swara?
- How was their usability and accessibility experience?
- How did Sangeet Swara impact their lives?

In the following subsections, we briefly describe our methods to evaluate above questions. We then discuss how low-income blind people used Sangeet Swara, what content they produced, as well as the benefits and pitfalls of Sangeet Swara.

6.1.1 Methodology

We used a mixed-methods approach to analyze the usage of blind users. We conducted a structured phone survey that asked one pre-recorded question to callers every time they called Sangeet Swara. The survey consisted of 15 subjective questions recorded in Hindi. The questions requested participants to share demographic data, background information, community moderation experience, and benefits and limitations of Sangeet Swara. A total of 204 users completed the survey, out of which 53 (26%) voluntarily identified themselves as blind. For the analyses presented in this chapter, we only consider data contributed by these 53 respondents.

For user analysis, we studied survey responses contributed by 53 blind users. We translated and transcribed their responses in English and analyzed them using open coding. The average length of the response was 38 words. We also studied their call logs to understand usage patterns.

For content analysis, we randomly sampled 100 posts that were recorded by blind users and inspected them on several criteria like gender, content type, location of callers, and quality of the recording. We also conducted 13 semi-structured qualitative interviews with blind users to investigate user engagement as well as strengths and weaknesses of Sangeet Swara. We reviewed and analyzed data immediately after conducting each interview. The insights obtained from the data analysis added more questions for the next interview. The interviews were translated and transcribed in English, and were analyzed using open coding.

6.1.2 Analysis of Call Logs

We were surprised to see how actively the blind survey respondents used Sangeet Swara. Though these participants were only 3.5% of all users of the service, they were responsible for recording nearly 25% of all contributions. The median number of posts recorded by them was 13 (max = 170 posts). Seven of them recorded more than fifty posts each. They placed 4,784 voice calls (19% of total calls), cast 7,350 upvotes (18% of all upvotes) and 26,559 downvotes (27% of all downvotes), shared 57 posts (8% of all shared events), and listened to posts 46,090 times (24% of all playback events).

While ten participants answered the survey partially, the remaining (N = 43) answered all survey questions. We also observed that although a few blind participants did not record any posts, they were heavy listeners of Sangeet Swara. For example, two such blind listeners called the voice forum 23 times and 123 times, respectively. These listeners also recorded emphatic and verbose responses (average response length = 50 words) to the questions asked in the phone survey.

Overall, the number of calls, posts, playback events, and votes as well as their enthusiastic partici-

pation in the survey demonstrates that they deeply valued Sangeet Swara.

6.1.3 User Analysis

Blind participants in our sample were from thirteen states in India. Two-thirds of them were from rural regions. About 93% of them were male, and 7% were female. On average, they were 24 years old (min = 15 years, max = 42 years, S.D. = 8.1 years). They came from a broad range of educational backgrounds: 17% held or were pursuing a master's degree, 19% held or were pursuing a bachelor's degree, 21% were in high school, 10% were in middle school, 2% only completed primary school, 2% were uneducated, and 10% received formal education in music. Nearly 19% of the participants did not share information on their educational background.

Nearly 25% of blind participants were employed and earned an average monthly income of USD 107 (min = USD 5, max = USD 334, S.D. = USD 110). About 45% of them were students, 14% were teachers, 12% were unemployed, 9% worked either as a telephone operator or a singer. We did not have employment information for 20% of the participants.

All blind participants owned a mobile phone. Nearly 25% of them reported using SMS. Only one participant had an email account and three participants had a Facebook account. Many participants had never even heard of Facebook and often responded: "*We do not have a Facebook account, but we have an account in Bank of India.*" They associated the word 'account' with banking services rather than Internet services.

6.1.4 Content Analysis

All 100 posts were recorded by male users. In 68 posts users reported their location, in 77 posts they shared their name and in 25 posts they shared their phone number publicly with all users. Based on the location they reported, users were from nine states in India. All posts but one were high-quality recordings. The average length of posts was 47 seconds (min = 5 seconds, max = 70 seconds, S.D. =

22.2 seconds).

Nearly 40% of the posts were similar to what people generally share on mainstream social media platforms like Facebook, WhatsApp, and Twitter. This category comprised discussion on topics trending in the service, generic informative posts, posts intended for specific people, news on topics of national and regional interest, posts requesting feedback from other participants, and posts requesting or sharing phone number. We found seven flirtatious posts where participants showered special attention and adulation to female contributors. One person also recorded a post reprimanding those who were flirting with women participants. We also found four posts where participants spoke about visual impairment.

About 25% of the posts were poems. Most of them were written to express feelings on love, separation, motherhood, visual impairment, environment, women empowerment, success, and persistence. Twenty-one posts were songs, including folk songs (N = 10), Bollywood songs (N = 8), and even recordings from a playback device (N = 3). To our surprise, we saw nine posts where people shared general knowledge information with each other by asking questions or recording answers to the questions asked previously. One example of a question asked on the forum is, "*When is the World Environment Day celebrated?*"

Two posts were jokes. We also found two posts containing abusive language and one post where a participant recorded sexually explicit content. We have made available twenty-five randomly selected posts recorded by blind users at: `https://soundcloud.com/socialmediavoiceforum/sets/random25`.

6.1.5 Benefits and Limitations

Sangeet Swara received impassioned usage from blind people (see Figure 6.1 for photograph of one of our users). Although 26% of the survey respondents self-reported themselves as a blind, we believe this number is a conservative estimate of the actual percentage of blind participants who

Figure 6.1: A blind user accessing Sangeet Swara.

used Sangeet Swara. It is worth noting that the representation of blind people on our service is significantly higher than their representation on Facebook, Twitter, WhatsApp, or even among the population of India. More importantly, blind users were spread out all across India, indicating that the service found broad appeal among blind people in low-resource environments.

Benefits of Sangeet Swara

Blind users recorded strong positive sentiments about the service and shared impactful stories on how the service was playing an influential role in transforming their lives. For example, several users shared that the service connected them with blind people in other states and far-off locations. Sangeet Swara was the first introduction to a social media service for 95% of participants in our sample. They valued interactions with other blind people. One such participant stated:

> *Using this service is a great experience. I listen to people from all over India, made many new friends, and heard many creative talents of other blind people. In this fast life, no one has time to listen to jokes, songs, and one-liners. Those who have time, do not have resources. Those who have resources, they do not have time. Now a days, literate, illiterate,*

> *poor, rich everyone has a mobile phone. The service has enabled those who do not have resources to consume entertaining content anywhere, anytime, and in any quantity.*
>
> U1 (Male, Telephone operator, 31 years, Maharashtra)

Many participants perceived Sangeet Swara to be exclusively designed for and used by low-income blind people, primarily because of the sheer number of blind users and abundance of songs, poems, and discussions central to visual impairment. For example, we found three songs on the importance of Braille and a discussion on Louis Braille during the content analysis. Blind people used the service to meet new people and earn social capital. Many users also exchanged their phone numbers with each other by recording a message on the service for having longer offline conversations:

> *The service is a boon for blind people. It gives us the opportunity to show and improve our talent. Blind people who use the service are very competitive and they continue to improve their messages. We also reach out to people in far-off towns and get to know them better. We get a lot of knowledge. I also get inspiration from listening to other blind people. Blind people who want to learn and make progress share informative messages with us.*
>
> U2 (Male, High school student, Uttar Pradesh)

We were curious to understand how blind users heard about Sangeet Swara. Eleven participants reported that they were told about the service either by a friend or a teacher. Five participants spread information about the service by calling their friends. During the qualitative interviews, one participant reported receiving a phone call from a friend to convey the gratitude for introducing him to the service. His friend told him: *"You have given me a new life. The service is very good."*

All participants were excited that their posts were heard by people all across India. When asked who (according to them) listened to the posts on the service, many participants responded that *"literate,*

knowledgeable and inquisitive folks," and people of all generations listen to it. One of them stated: *"Mothers, sisters, kids, old, government workers, officers, students, farmers, everyone listens to Sangeet Swara."*

Many users shared their personal stories and accounts of life on Sangeet Swara. Some blind users were so comfortable with the service that they recorded their children singing songs or reciting poems. Most blind users regarded the service as an avenue to access entertainment, share information, and learn skills. Given Sangeet Swara's focus on songs, many of them perceived it as a service to show, judge, and share feedback on musical talent. Five participants believed that the service was developed by 'The National Academy of Music', 'Dance and Drama', or 'the Government of India' to provide opportunities to low-income blind musicians. Sangeet Swara was also used to discuss current national and regional news. For example, five participants recorded performances and news on the 2013 North India floods:

> *Whatever I say about this service will not be enough. We hear good jokes, songs, poems and even useful knowledge. We listen to the important news of India and world. We also got to know the latest situation of North India floods on the service.*
>
> U3 (Male, High school student, 18 years, Gujarat)

Many users also felt comfortable sharing their career goals, aspirations, and vision with others on Sangeet Swara. They used the service for motivating people to fight corruption and serve marginalized communities:

> *I want to become a good man and fight corruption in India. Some people are using violence against women, killing the innocents, depriving the poor of the dignity. When will this end? It will end when we decide to become righteous and law-abiding citizens. We are the future, we have to make our country successful.*

U4 (Male, Student, 15 years, Jharkhand)

Many blind users derived instrumental benefits from Sangeet Swara. For example, five participants reported learning social skills by using the service. An eighteen-year old student from a small city in the state of Madhya Pradesh reported that he *"learnt how to speak properly, how to behave, and how to respect others"* by observing Sangeet Swara posts. Three participants reported that the service improved presentation of their thoughts, refined their grammar and accent, and helped them learn new vocabulary. They attributed an increase in their self-confidence to Sangeet Swara:

> *The service has provided me a lot of self-confidence. I can learn anything from the service. I learn a lot from general knowledge questions asked on the service. It is a great way to learn and understand principles of life. No matter how much I praise, it will never be enough. We get entertainment and knowledge. We also learn how to record better messages. The service gives me a lot of pleasure and knowledge.*

U5 (Male, High school student, Orissa)

Sangeet Swara provided more accessible venues to women and young girls for accessing information and entertainment. A fifteen-year old female student from a small town in the state of Uttar Pradesh found the messages on Sangeet Swara informative and suggested that the service helped her find new friends without the need to go a cybercafé: *"It is a great knowledge tool. We get to know more people and they get to know me. It is much better than Internet, Facebook and Twitter because we can use it without spending money. We can chat, listen to messages, understand them and learn from them."*

Sangeet Swara was successful because it could be accessed via ordinary phone calls from any phone. It provided several useful features like voice chatting, voting, and content sharing. As voice is a natural and accessible medium, the service was usable by blind people with limited technology exposure. The language of the service was in Hindi and hence it was usable even for people with no English language skills. Because the service was a toll-free line, even the poorest of the poor could

also use it. Sangeet Swara enabled several uneducated and unemployed blind people to create their own India-wide social network. One such user stated: *"I come from a village where it is very difficult to get educated. I want to thank you sincerely because you enabled all blind people in India to get to know each other."*

Limitations of Sangeet Swara

We found 22 posts containing abusive content. Twelve blind participants complained about the abusive content during the phone survey. One participant stated: *"Abusive messages should not be played. It causes pain in our heart. Please note the phone number of people who record abusive content and warn those who are misusing the service. It is a true adage that one bad fish can spoil the whole pond."* These messages formed perceptions that the service is not suitable for children and women. Although we did not allow users to flag abusive posts, future work could explore how well flagging can reduce abusive, derogatory, and inappropriate content on the service.

Most blind users struggled to 'share' messages with others. Only a fraction of all events (playback, vote, share, record) were share events, primarily because the sharing of content required users to read and send SMS. We found that blind users either remembered the phone numbers of their friends or wrote it on a Braille paper. Future work could provide a functionality where users enter phone number of a friend (rather than forwarding an SMS) to share the post. Once a valid phone number is entered, the friend will receive a call and listen to the the post. Future work could also use acoustic quick response codes for sharing the call position in an IVR tree with others [133]. However, using this technique for remote generation and recognition of audio codes would require setting-up additional IVR servers.

Blind participants also shared several suggestions for improving the design of the service for future deployments. Six participants requested a feature to send personal messages to other users. Two participants requested a discussion forum where they could record replies on posts while listening to them. We plan to include these features in the future deployment. In the next section, we discuss

the experience of low-income women on Sangeet Swara.

6.2 Use of Sangeet Swara by Low-income Women

More women than men in the world are subjected to intimate partner violence, early marriage, unpaid care and domestic work, and workplace discrimination [38]. These structural limitations, lack of agency to take life decisions [37], and limited access to education, healthcare, and financial resources [31, 32, 38] perpetuate the vicious cycle of gender discrimination. The United Nations has identified gender equality as a development goal fundamental to the foundation of a peaceful, prosperous, and sustainable world, and has advocated using Information and Communication Technologies (ICTs) to promote women empowerment [38].

Unfortunately, gender inequality also manifests in adoption, access, and use of ICTs. For example, women in South Asia are 38% less likely than men to own a mobile phone [39]. Even when they own a phone, they make and receive fewer calls, send fewer text messages, and use the Internet sparingly than men. Moreover, they perceive barriers to phone ownership and usage, such as cost of devices and the Internet, security and harassment concerns, and limited digital literacy, more acutely than men [39]. These factors significantly limit their participation on mainstream social computing technologies. For example, only one-fourth of all Facebook users in India are women [39].

Unexpectedly, voice forums have also received extremely low-participation from women. For example, CGNet Swara [109] and Sangeet Swara [161] in India have only 12% and 6% female contributors, respectively. Similarly, Baang [140] and Polly [141] in Pakistan have only 10% and 11% female contributors, respectively. Ila Dhageyso [87] in Somaliland has only 15% female users.

Although prior works have raised concerns about low-participation of women on these services [119, 140, 161], and a few have provided scattered insights (e.g., how posts by women received more votes due to flirting from men [161]), no prior work has systematically examined the factors that result in the limited use of these services by women, characterized women's participation by

analyzing usage logs, and outlined design recommendations for creating inclusive and inviting social media voice forums for women. Our work presents the first in-depth account of how a social media voice forum was used by low-income women in India and examines why the participation of women is almost non-existent on a forum that is intended to be inclusive, accessible, and usable for everyone.

6.2.1 Methodology

We used a mixed-methods analysis spanning quantitative and qualitative methods to examine how women and men used Sangeet Swara.

Quantitative Analysis

We selected all 5,361 posts on Sangeet Swara to conduct an in-depth content analysis. We recruited three coders (one male and two females) to analyze these posts. The coders were familiar with local socio-cultural norms, languages, and colloquial terms. On average, the coders were 32 years old. They had at least a bachelor's degree and were from middle-income families.

We requested coders to use the following rubric to analyze audio posts. For each post, a coder noted content type, gender of the recorder, how the post is related to women, and whether the recorder is threatening other users. The coders could select content type as 'abuse', 'blank or unclear post', 'flirt', 'self-introduction', 'joke', 'news', 'poem', 'question and answer', 'song', 'a message to other users', and 'pre-recorded content'. The coders could select the gender of the recorder as 'female', 'male', 'unsure', or 'blank'. The gender was coded 'blank' when the recorder did not speak anything (e.g., in a blank post or for pre-recorded content). When a post had multiple speakers, the coders were asked to mark the gender of the person who spoke for the most time. If a recorder referred to specific female users in the post, the coders marked the post as 'directed at female users'. If a recorder referred to women generally in the post, the coders marked the post as 'directed at women in general.' If a recorder had a conversation with other male users on topics that followed prior conversations with

female users or on women, the coders marked the post as 'discussion because of women'. The coders marked posts as 'threatening' when the recorder threatened other users in the post.

Initially, we assigned 100 audio posts to each coder to fill the rubric. We then computed inter-rater agreement using Cohen's Kappa coefficient. The lowest Kappa coefficient was 0.90, indicating high agreement between coders. We then divided the remaining dataset into three non-overlapping partitions and assigned one partition to each coder. Collectively, these coders listened to nearly 67 hours of posts to generate metadata that is central to our analysis. We analyzed this data on several interesting probes, such as the number of female and male contributors, similarities and differences in content recorded by women and men, content of posts directed at women, and interactions between female and male users, among other things.

To examine how female and male users reacted to non-inclusive posts such as abuse, flirts, threats, or verbal harassment, we mapped each anonymized phone number with the gender of the person who used that phone number to record posts. We only considered a phone number if it was used by the same gendered users (male or female), and discarded if it was used by both male and female users.

Qualitative Analysis

To recruit participants for surveys, we randomly selected users who used these services more than ten times and recorded at least one post. We conducted structured telephonic surveys with 18 Sangeet Swara users. The surveys explored several aspects including demographic information, limits imposed by family members in using these services, attitudes towards flirty, threatening, and abusive posts, inclusion and safety perceptions, and suggestions to make these services more inclusive for women. The surveys were conducted in Hindi.

We transcribed audio recordings and translated transcripts to English. We subjected our data to thematic analysis as outlined by Braun and Clarke [65] and rigorously categorized our codes to identify

factors that affect women's participation on Sangeet Swara. We engaged in regular discussions and iterated on our codes. Our first-level codes were specific, such as *"women ignored abusive messages," "women did not respond to flirt,"* and *"men hesitated to recommend the service to women in their family."* After several rounds of iteration, we condensed our codes into high-level themes, such as *"lack of agency," "structural limitations,"* and *"systemic discrimination."*

Ten Sangeet Swara participants were male and eight were female. On average, participants were 24 years old. A majority of them were unmarried. About 40% of the participants had less than 10 years of education. Half of the participants were employed, and the rest were homemakers, students, or unemployed. Among employed participants, nearly 33% were farmers, 20% were teachers, and 14% each were in private jobs or government jobs. On average, the monthly family income for a family of nine people was USD 160. A majority of participants owned a basic phone.

Limitations

Our analysis has a few limitations. First, the coders assigned gender based on the masculine or feminine characteristics of the voice in the audio posts. Our analysis thus excluded non-binary gender identities. Second, since we could not determine the gender of the users who did not record any post, our analysis on how men and women reacted to audio posts is based on those users who recorded at least one post.

Ethics

Sangeet Swara users were informed in the first call that their posts will be publicly available and will be used for research purposes. The services requested users to not record any private information such as their address or gender identity or phone numbers. The data we used for analysis did not have any personal identifiable information. The phone numbers were replaced with anonymized strings. Our study also received institutional review board approval. We also anonymized names of users and participants for use in this chapter.

Gender	Total posts	Unique users	Likes	Dislikes	Shares
Male	4,764	419	21,630	58,644	189
Female	275	31	270	2,636	15

Table 6.1: Usage statistics by gender for Sangeet Swara.

6.2.2 High-Level Usage Patterns

We analyzed 5,361 posts on Sangeet Swara. An overwhelming majority of these posts (89%) were recorded by men. Only 5% posts were recorded by women. The remaining posts were either blank or unclear or contained pre-recorded content. Users recorded posts from 506 unique phone numbers. We discarded data for phone numbers that were used by both men as well as women to record posts. Assuming a one-to-one mapping between remaining phone numbers and users, Sangeet Swara had 450 unique contributors, out of which 419 were male and 31 were female.

Table 6.1 shows how men and women participated on Sangeet Swara. Men recorded 17 times more posts, and liked and disliked these posts 80 times and 22 times more than women. On average, they recorded 1.5 times more posts, and liked and disliked posts 6 times and 1.6 times more than women, indicating that the participation was dominated by men. Even the most fervent female users were far behind their male counterparts. For example, the number of posts contributed by top 25 female contributors combined were less than the number of posts recorded by the most prolific contributor among men. Figure 6.2 shows the distribution of the number of posts recorded by top 25 male and female Sangeet Swara contributors. The median number of posts recorded by these men and women were 81 and 4, respectively. A Mann-Whitney's U test indicated a significant difference between the number of posts recorded by top 25 male and female contributors, ($U = 615, Z = 5.8, p < 0.001$). We also found significant effect of gender ($p < .001$) on total calls, total likes, and total dislikes by top 25 male and female users.

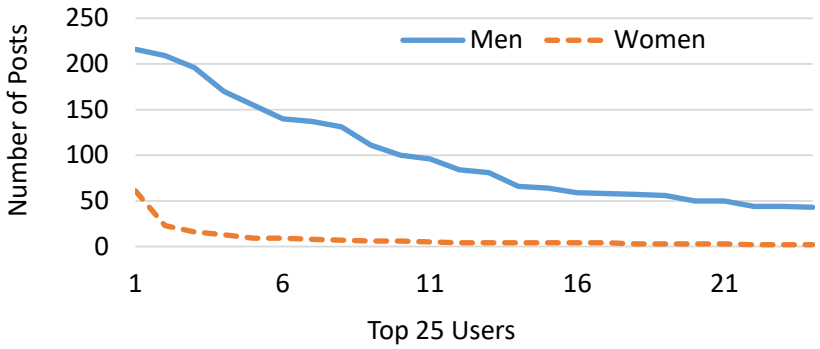

Figure 6.2: Distribution of the number of posts recorded by top 25 men and top 25 women contributors of Sangeet Swara.

Content Analysis

Figure 6.3 shows the number of posts of different types (on a log scale) recorded by male and female contributors. A Fisher's exact test indicated significant differences in the content recorded by men and women ($p < .0001$). Most posts by women contained songs (34%) and most posts by men contained messages for other users (39%). The second-most popular category was poems among female contributors and songs among male contributors. Poems and songs together accounted to 66% posts among female contributors and 39% posts among male contributors. On average, we found that men recorded more posts containing abuses, flirts, introductions, messages to other users, news, and informative general knowledge questions and answers than women. In contrast, women recorded more songs, jokes, poems, pre-recorded content, and unclear or blank posts than men.

Analysis of Votes

In general, users disliked posts more than liking them. This is probably because the top-ranked content on Sangeet Swara (chosen based on the likes and dislikes given by users) was featured as 'the best post', leading to unhealthy competition among users who disliked posts recorded by others

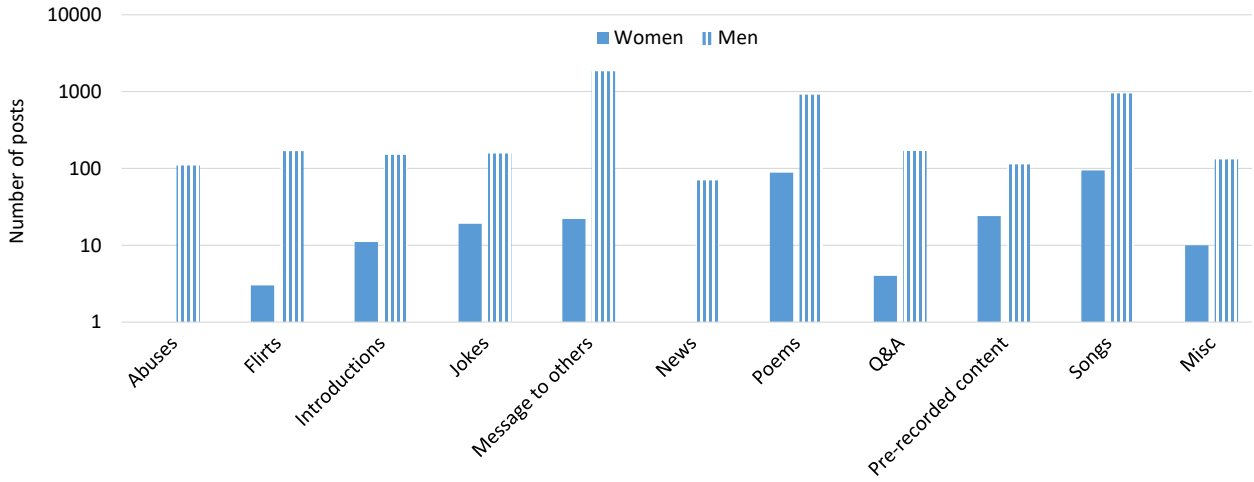

Figure 6.3: Distribution of posts (on a log scale) by content types and gender.

to improve their chances to get a higher rank.

Generally, women were more disapproving of content than were men. While men disliked 2.7 posts for each post they liked, the ratio of dislike to like was 9.8 for women. A Chi-square test indicated a significant effect of gender on the distribution of likes and dislikes ($\chi^2(1, N = 83,180) = 449.72$, $p < 0.001$).

Our content analysis also indicated that a notable number of posts were directed at women and a significant number of posts contained abusive, flirtatious, and threatening messages. In the next subsections, we analyze who recorded posts directed at women, how users were flirting with each other, and why they were threatening and abusing other users.

6.2.3 Posts Directed at Women

Our coding indicated that male and female users recorded 602 posts (11%) that were either directed at other female users or discussed their participation. We classified these posts in three categories: (1) posts that called out female users, (2) posts that referred to women generally, (3) posts that

Table 6.2: Examples of posts focusing on women or on topics that follow prior posts involving women.

Type of posts	Example post
Mentioning women user	Hello, my name is Roshan. Reshma I want to know your mobile phone number.
Discussing women generally	Why do women wear revealing clothes? Why they want to show skin? Why are they following western values? An Indian girl should feel ashamed for exposing her skin.
Spiraled from conversations on women users or women generally	Hello, some fool was just abusing in the last post. Do not abuse. Women and men from all over India listen to your messages. Do not misbehave here. (A user reprimanding another user for abusing women in a prior post)

followed topics spiraled from prior conversations centered on women. Table 6.2 shows examples of posts for these categories. The first, second, and third categories had 372, 147, and 83 posts, respectively.

Female users recorded 19 posts (3%) to appreciate other female users for recording good content or to celebrate womanhood and motherhood. For example, a woman recorded the following poem on female infanticide:

> *Daughters are our pride, they make a home happy.*
> *They are not a burden, they bring us prosperity.*

Figure 6.4 shows the distribution of posts (on a log scale) for the three categories. Male users recorded 500 posts (83%) that were directed at female users or that discussed women in general. About 95% of the posts that called out female users had abuse, flirts, and adulation. Many men

also recorded posts criticizing women in general for receiving *"more votes because of preferential treatment from other men."* They often actively encouraged other users to dislike all posts recorded by women. About one-fourth of the posts referring to women in general had abuse. Male users recorded 83 posts on topics in prior conversations centered on women. About 90% of these posts had male users fighting with each other to impress other female users or requesting troublemakers to avoid recording abusive and flirty posts.

These results indicate that although Sangeet Swara users had only 11% of all posts targeting or discussing women, most of these posts were harassment in the form of abuse, flirts, and threats. Often these posts spiraled several heated arguments among community members, creating an acrimonious environment for female users.

6.2.4 Flirty Posts

Sangeet Swara had 171 flirty posts. Men recorded 98% of these posts. A Fisher's exact test indicated a significant effect of gender on the number of flirty posts ($p = .02$, odds ratio = 0.3). We also analyzed who were the target of these flirty posts. Men flirted with women in 166 posts (97%) and with other men in two posts, and women flirted with men in three posts. A Fisher's exact test indicated a significant effect of gender of the recorder on the number of flirty posts sent to men and women ($p < .0001$).

Men flirted with women in several ways. For example, many male users inundated female contributors with adulation and incessantly requested them to record more content. They requested female users to dedicate a song or poem to them. Some men requested other users to like posts from women with whom they were flirting. For example, a man posted this messages on Sangeet Swara:

> *Sapna, your voice is so sweet. I want to be your friend. Everyone, please upvote all posts from Sapna.*

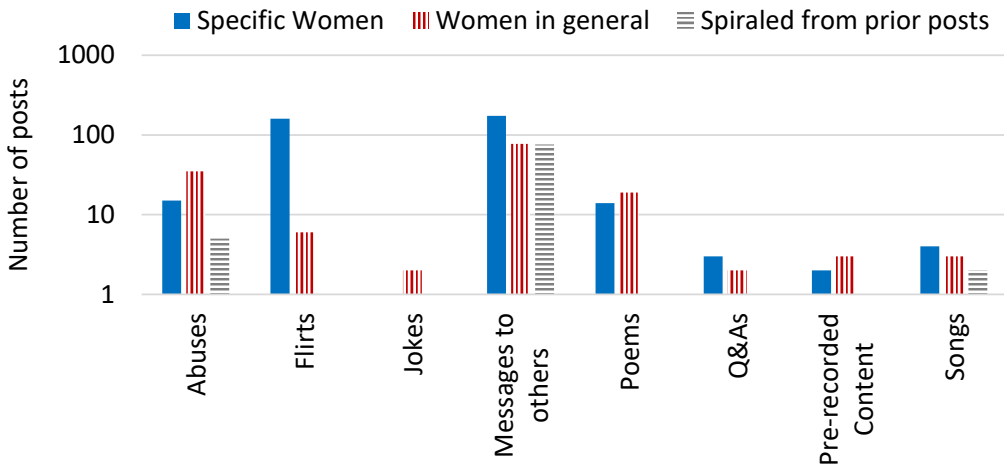

Figure 6.4: Distribution of different types of posts (on a log scale) directed at female users.

Some men were forceful in sharing their feelings with women. They often harassed female users by repeatedly professing love, proposing to them, and asking them to reciprocate their feelings. They shared their phone numbers publicly and asked women to call them. In a sample of 100 random recordings, we found that men shared their phone numbers in 21 posts. For example, a man recorded:

> *Saroj, please call me and tell me how are you. You have to become my friend. Where are you from? Where do you live? I am in love with you. Call me on xxxxx xxxxx or give me your personal number.*

We found two posts where men flirted with other men and shared their number publicly inviting them for offline conversations. We also found three posts where women flirted with men; two women recorded posts stating that they are looking to make male friends and another woman expressed excessive admiration for a male user. In our entire sample, we found only one post where a woman shared her phone number and asked male users to call her.

We also analyzed users' votes to examine their reactions to flirty posts. A Fisher's exact test indicated

a significant difference between the proportion of dislikes and likes given by male and female users to flirty posts ($p < .01$, odds ratio = 3.1), suggesting that women disliked flirty posts more than men. All female participants in our surveys reported flirting to be a cause of distress and a key reason for their low participation. In contrast, several male participants disregarded that flirty posts created an uninviting environment for female users. A male participant stated:

> *"When guys and girls are together, flirting can't be stopped and should not be stopped."*

Male and female participants gave different reasoning for why men were flirting with women. Some male participants held movies responsible for flirtatious behavior of men. One of them stated:

> *"Men see films and TV shows and think that the only way to gain attention of a woman is by teasing and pursuing her. That is what they see and do."*

On the other hand, four female participants blamed women for flirtatious behavior of men. One of them stated:

> *"Some women don't leave a good impression based on how they talk and what they say. We are also at fault. Not all fault is of men."*

To summarize, these results indicate that most of the flirty posts were directed at women. Often these posts were disturbing and some posts had a sexual undertone. Most male users condoned flirting and many female users showed stronger disapproval for these posts than male users. While men avoided responsibility for unjust behavior by blaming television soaps and movies, women engaged in victim blaming to justify flirtatious behavior of men.

6.2.5 Threatening Posts

Sangeet Swara had 104 threatening posts. All of them were recorded by men and the majority of these posts (62%) were directed at women. We found a significant difference between the proportion of threatening posts directed at women and men ($\chi^2(1, N = 5,361) = 332.3, p < 0.0001$). Most threatening posts (45%) were abusive in nature. In 56% of these posts, users called out the names of intended recipients. Some men threatened other users indirectly by singing songs and reciting poems. For example, a man recited the following content as a poem:

> *Don't pluck the flowers, you will get stung by thorns. Don't tease my girl, you will get a slap. If anyone troubles Maya, I will behead them.*

We found three main reasons why users were threatening others. First, several men who were trying to impress female users were threatening each other when others flirted with women they liked. Second, some men recorded sexually suggestive content, sparking sharp criticism from other male users. Often these arguments resulted in a series of abusive and threatening posts. Third, a few men threatened women who did not respond to their flirty posts or who condemned their behavior. A female participant shared her ordeal:

> *"A man posted offensive content, when I did not respond to his advances. It is my wish if I want to talk to him. How can he force me? When I could not tolerate the misbehavior, I left the service."*

The analysis of how users reacted to threatening posts indicated that men condoned these posts by not disliking them as much as did women; while the ratio of dislikes to likes for women was 42, the ratio was only 4.4 for men. A Fisher's exact test indicated a significant difference between the proportion of likes and dislikes given by men and women on these posts ($p < .0001$, odds ratio =

14.8). Many male participants stated that women should not record posts containing threats and abuse, instead *"they should tolerate it."*

These results indicate that women encountered substantial number of threatening posts that were either directed at them or were because of men fighting with each other to gain their attention. We also found an evidence of systemic bias present in patriarchal societies where unruly behavior by men was not only condoned but often appreciated. However, the same behavior by women received an immediate disapproval.

6.2.6 Abusive Posts

Sangeet Swara had 109 abusive posts. All of them were recorded by men. We found a significant difference between the proportion of abusive posts recorded by men and women (Fisher's exact test: $p < .01$), indicating that men recorded more abusive posts than women.

We also analyzed who were the target of these abusive posts. About 46% of these posts were directed at women. Flirting by men transpired almost half of the abusive posts. For example, some men constantly harassed women to share their phone numbers. When these women did not share their number, men felt rejected and recorded abusive posts directed at these women. Some men also abused women who thanked or appreciated other men instead of responding to their flirty posts. When other users admonished these men for abusing women, they ganged up on those who were berating their unruly behavior.

We also analyzed the votes given by users to examine whether they condemned or condoned abusive posts. The ratio of dislikes to likes given by female and male users was 15 and 3, respectively, meaning that women disliked abusive posts more strongly than men. A Fisher's exact test indicated a significant effect of the gender of users on the proportion of likes and dislikes given on the abusive posts ($p = .01$, odds ratio = 0.19). These findings indicate that women encountered substantial abusive posts that were either directed at them or were exchanged between men arguing over women.

6.2.7 Blackmailing

We found that a few women were blackmailed by men. As previously outlined, several men shared their phone numbers publicly and requested female users to call them. Some men also recorded posts suggesting that they could help women in finding jobs or that they need *"talented female singers for their orchestra*". A few women willingly gave their phone numbers to men. When the women stopped talking to men after unpleasant phone calls, they were threatened that their phone numbers will be publicly released if they do not continue the private conversations. When these women ignored the threats, men posted these women's phone numbers on the service, suggested romantic relationships with them, and assassinated their character. A woman stated:

> *"The man told me that if you stop talking to me, then I will share your number with others. When I did not pick his phone calls, he recorded a post saying that I am not a good woman and people should stay away from me."*

These incidents were also corroborated by several male participants. One of them shared:

> *"I have heard men saying to women that 'if you won't talk to me then I will share your number with everyone'. I have heard them abusing women and talking dirty stuff with them."*

Such posts and predatory behavior by men strongly discouraged women to use Sangeet Swara. Such posts prompted a few female users to assume a different identity on Sangeet Swara by using a pseudonym and a different phone number.

6.2.8 Agency

Although only some men participated in unruly behavior, the abusive, flirty, and threatening posts tremendously damaged perceptions about the inclusivity and safety of women on these services.

Some male users prohibited women in their family and social circles from using these services because of the indecorous content recorded by other male users. A male user stated:

> *"I will not allow women in my family to get exposed to these abusive messages."*

Since Sangeet Swara spread virally by the word-of-mouth, negative experiences of early female users adversely affected its adoption and use by women since most female users stopped recommending it to their female friends and relatives. A woman shared:

> *'How can I ask my family and friends to listen to these posts where men are abusing women and other men. I would be in trouble, if my family learns that the service has such posts. Family members are accepting of these behaviors if a man does it, but not when it is done by women."*

We found that women were extremely hesitant to object to abusive, threatening, and flirty posts directed at them, primarily due to deep-rooted patriarchal values that discourages women to argue and question others. Most women were worried that they will face backlash, on the service from predatory men and in real life from family members, if they record threatening responses or responded to flirty posts. They lacked the agency to retaliate unruly behavior or explore friendships with people from the opposite sex due to socio-culutral sensitivities shaped by patriarchy. Most men took the participation of women for granted and viewed them as objects of desire, reinforcing patriarchy in these digital social spaces.

In the next section, we discuss the benefits and pitfalls of Sangeet Swara. We examine its design using the feminist HCI lens and discuss several design suggestions to create more vibrant, inclusive, and equitable social media experience for women.

6.3 Discussion and Conclusion

In this chapter, we presented the detailed analysis of how low-income blind people used a social media voice forum that was not originally designed for them. The service spread rapidly among them without any marketing effort and enabled them to make new connections, showcase their talent, and learn information. The service also enabled them to derive several instrumental benefits, gain social acceptance, and access entertainment. On the other hand, the participation of low-income women was surprisingly low on the service despite its accessible design. We found that women struggled to negotiate their identity and experienced abuse driven by patriarchal norms.

There are several practical barriers in digital inclusion of women in low-resource communities. Such barriers include comparatively lower literacy and financial agency among women than men which results in lack of access to mobile phones and connectivity [39]. A large fraction of women in such communities only have access to shared mobile devices where usage is directed by male family members. Social media voice forums such as Sangeet Swara have been successful in reaching some women in such communities. Once connected, these women enjoy access to community-generated content and play an active role in creating and moderating content. These services provide them a voice, a digital social identity, and more independence. Their social inclusion leads to greater connectivity and access to entertainment, employment, education, and health opportunities on equal terms as men.

However, our work highlighted significant secondary barriers to women's digital inclusion beyond the basic hurdles of literacy, connectivity, and availability of devices. Once connected through social media voice forums, these women faced harassment, abuse, threats, and systemic marginalization. Using an intersectional HCI lens [146, 172], we found that certain groups within marginalized communities are more marginalized than others. For example, several male Sangeet Swara users abused other users based on the gender, caste, or religion. A few Sangeet Swara users exhorted the community to downvote posts of a female user belonging to a minority group in India after an argument with her. Similarly, while Sangeet Swara empowered a section of marginalized communities (e.g.,

low-income blind people), at the same time it disenfranchised the rights, voice, and liberty of women in these communities. We found that *access* is just a first step towards actual and meaningful social inclusion, and voice forums like Sangeet Swara are still a long way from providing a welcoming, vibrant, safe, and enriching environment to women.

We faced significant barriers in reaching female users for follow-up surveys and interviews. Most of our phone calls were answered by male family members. Even when women answered the call, many of them handed over the phone to a male family member as soon as they realized that there is an actual person (a female surveyor) calling them. We found that most women users were comfortable engaging in an asynchronous social interaction through the mediation of a social media service compared to actual conversations with unfamiliar men and women. Even among the women who agreed to participate in our surveys, a few did not acknowledge that they had used the service and some made an excuse to hangup to avoid the conversation. We believe that these women had a bad experience with the service and did not want to admit that they used it, or did not want their family to know about their experience.

The design of Sangeet Swara was only partially compatible with the principle of pluralism from the feminist HCI framework. Although the service was designed to be inclusive of low-income people by making it toll-free and low-literate people by enabling speech-based interactions, no special provisions were made to welcome participation from women. The prompts were recorded in a male voice, reinforcing the perception that men are the primary target users. Simple adaptations in prompts, for example, enabling users to select between prompts in male voice or female voice and explicitly inviting participation from both men and women, could lead to significant changes in perceptions about inclusivity of voice forums. Another way to encourage participation of women is by rewarding them with soft incentives for their participation. For example, a post recorded by women could be rewarded with extra virtual airtime to access the forum for free. Although it is expected that male users may come up with ways to deceive such gender recognition filters to earn additional free access, we still expect such incentives to encourage female participation. This would also convey to users that these services are not exclusively designed for men and warmly welcome

participation of women. Even changing the perception about a service may lead to an improvement in users' behavior.

Another way to promote participation of women in voice forums might be to provide optional audio filters to mask their gender identity. Such disguise could allow them anonymous access, hence alleviating their fears regarding gender-specific targeted abuse. These filters could also provide them agency to retaliate against harassment while protecting themselves against patriarchy driven social abuse. However, anonymization is a double-edged sword. Men can also use it to hide their identity and post inappropriate content targeting women. Similarly, the very fact that a post is gender-anonymized could signal vulnerability and be taken as a cue that it is recorded by women. Moreover, voice is not the only gender-cue in audio posts and the use of gender pronouns, linguistic constructions, and women-specific discourse could also reveal their identity. We believe that anonymization might not be a viable solution to mask users' gender identity, however, it may help with masking their personal identity from people who oppose their use of voice forums. Since nuanced treatment of identity and self are one of the central tenets of feminism, we also feel that taking away a woman's gender identity is not a solution to a problem that must be solved through an acceptance of her identity and rights that it entails.

From a standpoint of participation from the feminist HCI framework, Sangeet Swara enabled users to participate equally in deciding whether posts adhere to community standards. The service masked the information from users about who liked, disliked, or reported their posts, putting every user on equal footing, a decision reflecting the feminist commitment to equality. Although community moderation in Sangeet Swara showed promise since users disliked blank and unclear posts, it did not work well for abusive, flirty, and threatening posts. Most of these posts involved multiple users in a heated exchange, and many users did not objectively vote on these posts. Instead, the voting was based on the sides users picked among people involved in the argument. Community voting was also misused by some users who lobbied to downvote posts of their opponents. Despite its current limitations, we expect community moderation to play an active corrective role in voice forums. We believe that a service that enables the community to set its own rules and implement

them through community moderators, has a chance of evolving into an inclusive service for women. The service could also allow members to hold regular elections for voting on community rules and roles. The service could also have a user reputation system that is based on community votes and directly linked to concrete outcomes like additional virtual airtime. Assigning weights to votes based on the number of female and male users could also put judgments by men and women on equal footing.

Many women requested a dedicated voice forum for them. Such a gendered model matches the pattern of their daily social lives where they have women-only carriages in trains and dedicated compartments in buses. Dedicated services for women is not an alien concept even in developed countries where special interest groups around maternal health, pregnancy, and breastfeeding are often women-only forums. We believe a women-only service could encourage more meaningful participation from women in low-resource environments who are afraid to raise their voice on mixed-gender voice forums due to the fear of harassment driven by patriarchy or simply because they are shy to openly express themselves in situations where men are expected to hear and comment. A women-only service could only be successful if it blocks uninvited participation of men. Although it is possible to identify and remove male-recorded audio posts using natural language processing and community moderation techniques, preventing passive male users from listening to posts and expressing their opinions via non-verbal means (e.g., likes, dislikes) is far more challenging. A passcode-based access to the service could make it too complex for the primary target user group of low-literate women. Instead, the service could require users to announce themselves every time they access the service. The audio could be then gender-identified to allow or deny access.

Sangeet Swara lacked in values of self-disclosure and advocacy from the feminist HCI framework. It did not explain the importance of votes to its users, confusing them how posts are ranked and ordered. These limitations could be overcome by leveraging participatory design processes and integrating low-income, low-literate men and women in the design process, something that we initially neglected. From a standpoint of ecology from the feminist HCI framework, there is a need to reflect how the design of voice forums like Sangeet Swara could transfer social injustice and patriarchy

driven harassment from offline social spaces to digital social spaces.

The interface and features of current voice forums are not modelled to prevent harassment of women and to make them feel safe and included. However, such services do connect women who otherwise have no means of participating in digital social spaces. Design considerations such as the ones suggested above could create voice forums that welcome women, prevent harassment, and evolve the behavior of the connected user-base through policies and practices that originate from better values of the society itself.

Chapter 7

CONCLUSION

While social computing technologies—like social media platforms, online forums, crowdsourcing marketplaces, gig-economy platforms—have transformed how people participate in the information ecology and digital economy, these platforms have discounted the needs and wants of billions of people who experience literacy, language, socioeconomic, and connectivity barriers. To address the information and instrumental needs of these people, several HCI4D practitioners and researchers have designed voice forums that enable users to interact with others via ordinary phone calls in local languages. Although voice forums have had a demonstrated impact on marginalized communities, most forums operate at a pilot scale because of challenges in managing local language content, high costs of voice calls, and difficulties in building and deploying these services. In this thesis, we discussed several approaches to scale, sustain, and replicate voice forums by addressing limitations that significantly impede their impact.

To manage local language content on voice forums, we used community moderation by low-income, low-literate voice forum users, most of whom were first-time users of a social computing technology, and lacked digital skills and training in moderation. We demonstrated that a community of marginalized voice forum users can categorize and moderate local language content with an accuracy comparable to expert content moderators.

To ensure that low-income people can afford the cost of phone calls to voice forums instead of relying on toll-free lines that becomes very expensive to sustain as the usage scales, we built a new speech transcription marketplace that enabled low-income basic mobile phone users to transcribe

audio files vocally. We also demonstrated the feasibility, acceptability, and usability of our system to financially sustain voice forums while supplementing income of users of voice forums.

To enable practitioners, governments, and non-profit organizations with limited technical capacity to build, set up, and maintain voice forums, we designed and built a free and open source toolkit. Using services deployed on the toolkit, basic mobile phone users can record and listen to audio messages, and their voices can be heard by a global community on Facebook and YouTube. Thus far, the toolkit has been used by more than a dozen governmental agencies, social enterprises, and grassroots organizations to set up voice forums. Collectively, over 25,000 people in South Asia and Africa have spent 6,000 hours and made 100,000 phone calls to access services deployed using the toolkit.

This thesis also advances the dialogue on the benefits and pitfalls of social computing. For example, while Facebook has connected billions of people worldwide, there are rising concerns about privacy breaches and data misuse. While Uber allows millions of car owners worldwide to supplement income, there are reports about Uber's exploitation of drivers to increase revenues. Similarly, while Amazon Mechanical Turk (MTurk) lets millions of people with limited skills to earn money by doing basic micro tasks, labor exploitation in the form of unpaid work, unfair evaluations, no insurance, and sub-minimum-wage has made MTurk *"a new kind of poorly paid hell"* [147]. Voice forums, like any other social platform, come with their own benefits and pitfalls. They end up reflecting the existing sociocultural norms and values of the society, including its strengths, shortcomings, and biases. For example, while Sangeet Swara transformed lives of low-income blind users, the same service exposed women to patriarchy-driven abuse, threats, and flirty content.

We emphasize the importance of using intersectionality lens while designing technologies for people in low-resource environments who are often affected by multiple discriminations and disadvantages. Our work demonstrates that not everyone in low-resource environments are equally marginalized, some user groups are more marginalized than others. For example, female Sangeet Swara users were marginalized not only based on their income levels and literacy skills, they also

encountered gender-based discrimination. Our findings accentuate the need to design technological interventions carefully so as to minimize unintended negative consequences. If poorly designed, introducing new technologies may widen existing economic, cultural, and societal inequalities. The HCI4D literature has many examples of how technology has potential to positively impact people in low-resource environments. While it is important to share success stories, equally important is to recognize weaknesses and failures of technological interventions. We hope that our findings will encourage readers to share their success as well as failure stories in more detail. Finally, the experience of women on Sangeet Swara demonstrated that access is not equal to inclusion, and much more is required to address secondary barriers beyond the basic hurdles of literacy, connectivity, and poverty.

Both mainstream social media platforms and voice-based social media services face grand challenges when tackling misinformation, disinformation, harassment, and abuse. These platforms and services differ greatly in their scale, features, interfaces, and supported languages. Moreover, their target users have key differences in literacy and digital skills, geopolitical environments, and sociocultural values that dictate their participation on these platforms and services. As a result, solutions to tackle misinformation and harassment on Facebook might be ineffective for Sangeet Swara, and vice versa. Future research could use techniques from collaborative filtering and machine learning to reduce inappropriate remarks and misinformation on voice-based social computing services. This presents interesting research questions, such as which features could identify inappropriate content in local language audio files? How to identify interconnected networks and interrelated activities of those spreading disinformation? How to prevent misuse of 'report abuse' features to blacklist posts from competitors? How to address situations where the collective ignorance of community members eclipse collective intelligence (e.g., the community condoning bullying of women)? The HCI4D community needs to address these open challenges to make voice forums truly diverse, inclusive, and impactful.

BIBLIOGRAPHY

[1] Asterisk. https://www.asterisk.org/home.

[2] Awaaz.De | Mobile Solutions for Social Impact. https://awaaz.de/.

[3] engageSPARK. https://www.engagespark.com/.

[4] Exotel - Cloud Communication APIs for Calls, SMS, Authentication. https://exotel.com/.

[5] Freedom fone. http://freedomfone.org/.

[6] FreeSWITCH - Open Source Telecom Stack. https://freeswitch.com/.

[7] Jana. http://www.jana.com/.

[8] Ozonetel. https://www.kookoo.in/.

[9] Samasource. http://samasource.org/.

[10] Tropo. https://www.tropo.com/.

[11] Twilio - Communication APIs for SMS, Voice, Video and Authentication. https://www.twilio.com.

[12] WHO | Global data on visual impairment. http://www.who.int/blindness/publications/globaldata/en/.

[13] Amid fund crunch, CGNet Swara eyes shift to Bluetooth radio tech, September 2016. https://www.livemint.com/Politics/UcrYsrB8fIAGTDiIoC452N/Amid-fund-crunch-CGNet-Swara-eyes-shift-to-Bluetooth-radio.html.

[14] India average daily wage rate forecast 2016-2020, 2016. http://www.tradingeconomics.com/india/wages/forecast.

[15] Medical Transcription Services Market – Global Industry Analysis, Size, Share, Growth, Trends and Forecast, 2013 – 2019. Technical report, Transparency Market Research, 2016.

[16] Accessibility | Android Developers, 2017. https://developer.android.com/guide/topics/ui/accessibility/index.html.

[17] Amazon Mechanical Turk, 2017. https://www.mturk.com/mturk/welcome.

[18] AWS Developer Forums: MTurk CENSUS: About how many workers were on mechanical turk in 2010?, 2017. https://forums.aws.amazon.com/thread.jspa?threadID=58891.

[19] Be My Eyes ~ Lend Your Eyes to the Blind, 2017. http://bemyeyes.com.

[20] BSNL's Rs 8 and Rs 19 plans offer voice calls at 15 paisa per minute, September 2017. http://indianexpress.com/article/technology/tech-news-technology/bsnls-rs-8-and-rs-19-plans-offer-voice-calls-at-15-paisa-per-minute-4832774/.

[21] Crowdflower: AI for your business, 2017. https://www.crowdflower.com/.

[22] India's Mobile-Phone Price War Seen Spurring Consolidation. *Bloomberg.com*, January 2017. https://www.bloomberg.com/news/articles/2017-01-26/india-s-mobile-phone-price-war-seen-dialing-up-consolidation.

[23] Liv.ai, 2017. https://liv.ai/.

[24] mSurvey, 2017. https://www.msurvey.co/.

[25] Poverty & Equity Data | India | The World Bank, 2017. http://povertydata.worldbank.org/poverty/country/IND, urldate = 2016-09-30.

[26] 2018 Global and Regional ICT Estimates. Technical report, International Telecommunications Union, December 2018.

[27] 3-2-1 – On-Demand Messaging for Development, 2018. http://hni.org/what-we-do/3-2-1-service/.

[28] BSNL introduces Rs 19 prepaid plan with affordable voice calling rate for 54 days, July 2018. https://indianexpress.com/article/technology/tech-news-technology/bsnl-introduces-new-rs-19-prepaid-plans-brings-down-voice-calling-rates-5265617/.

[29] BSNL's New Rs. 319 Prepaid Plan Offers Unlimited Voice Calls for 90 Days, 2018. https://gadgets.ndtv.com/mobiles/news/jio-effect-bsnl-rs-319-99-unlimited-voice-calls-prepaid-plans-caller-tune-validity-1845861.

[30] CGNet Swara, 2018. http://cgnetswara.org/.

[31] Empowering Women, Developing Society: Female Education in the Middle East and North Africa – Population Reference Bureau, 2018. https://goo.gl/5oZzTq.

[32] Facts and Figures: Economic Empowerment, 2018. http://www.unwomen.org/en/what-we-do/economic-empowerment/facts-and-figures.

[33] Gram Vaani, 2018. http://www.gramvaani.org/.

[34] Kan Khajura Tesan, 2018. http://www.kankhajuratesan.com/.

[35] Uniform call rate from today, August 2018. https://www.thedailystar.net/news/business/telecom/btrc-minimum-call-rate-tk-045-be-effective-midnight-early-hours-tuesday-bangladesh-1620196.

[36] Vodafone's new Rs 99 prepaid recharge offer comes with unlimited calling, August 2018. https://indianexpress.com/article/technology/tech-news-technology/vodafone-launches-new-rs-99-pre-paid-tariff-plan-to-take-on-reliance-jio-and-airtel-5305588/.

[37] Voice and Agency: Empowering Women and Girls for Shared Prosperity, 2018. http://www.worldbank.org/en/topic/gender/publication/voice-and-agency-empowering-women-and-girls-for-shared-prosperity.

[38] Spotlight on Sustainable Development Goal 5: Achieve gender equality and empower all women and girls, 2018-02-01.

[39] Bridging the gender gap: Mobile access and usage in low- and middle-income countries, 2018-02-03.

[40] Castingwords, 2019. https://castingwords.com/.

[41] Cloud speech api: Speech to text conversion powered by machine learning, 2019. https://cloud.google.com/speech/.

[42] CrowdSurf, 2019. http://crowdsurfwork.com/.

[43] Quick transcription service, 2019. http://www.quicktranscriptionservice.com/Hindi-Transcription.html.

[44] Rev, 2019. https://www.rev.com/.

[45] Scripts complete, 2019. http://scriptscomplete.com/Hindi-Transcription-Services.php.

[46] SpeechPad, 2019. https://www.speechpad.com/.

[47] Tigerfish, 2019. http://tigerfish.com/.

[48] TranscribeMe, 2019. http://transcribeme.com/.

[49] Transcription Services Us, 2019. http://www.transcription-services-us.com/Language-Transcription-Rates.php.

[50] Amna Abid and Suleman Shahid. Helping Pregnant Women in the Rural Areas of Pakistan Using a Low-cost Interactive System. In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*, ICTD '17, pages 42:1–42:5, New York, NY, USA, 2017. ACM.

[51] Sheetal K. Agarwal, Anupam Jain, Arun Kumar, and Nitendra Rajput. The World Wide Telecom Web Browser. In *Proceedings of the First ACM Symposium on Computing for Development*, ACM DEV '10, pages 4:1–4:9, New York, NY, USA, 2010. ACM.

[52] Sheetal K. Agarwal, Arun Kumar, Amit Anil Nanavati, and Nitendra Rajput. Content Creation and Dissemination By-and-for Users in Rural Areas. In *Proceedings of the 3rd International Conference on Information and Communication Technologies and Development*, ICTD'09, pages 56–65, Piscataway, NJ, USA, 2009. IEEE Press.

[53] Chris Anderson. The internet has created a new industrial revolution | Chris Anderson. *The Guardian*, September 2012. https://www.theguardian.com/technology/2012/sep/18/chris-anderson-internet-industrial-revolution.

[54] Andrew Arnold. Can Social Media Have A Positive Impact On Global Healthcare?

[55] Siddhartha Asthana, Pushpendra Singh, and Amarjeet Singh. A Usability Study of Adaptive Interfaces for Interactive Voice Response System. In *Proceedings of the 3rd ACM Symposium on Computing for Development*, ACM DEV '13, pages 34:1–34:2, New York, NY, USA, 2013. ACM.

[56] Shaowen Bardzell. Feminist HCI: Taking Stock and Outlining an Agenda for Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1301–1310, New York, NY, USA, 2010. ACM.

[57] Shaowen Bardzell and Jeffrey Bardzell. Towards a Feminist HCI Methodology: Social Science, Feminism, and HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 675–684, New York, NY, USA, 2011. ACM.

[58] John C. Bertot, Paul T. Jaeger, and Justin M. Grimes. Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. *Government Information Quarterly*, 27(3):264–271, July 2010.

[59] Ananya Bhattacharya. Making a phone call in India is now nearly free, 2018. `https://qz.com/india/1331946/reliance-jio-effect-phone-calls-in-india-are-now-nearly-free/`.

[60] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. VizWiz: Nearly Real-time Answers to Visual Questions. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 333–342, New York, NY, USA, 2010. ACM.

[61] Soutik Biswas. How WhatsApp helped turn a village into a mob. *BBC News*, July 2018. `https://www.bbc.com/news/world-asia-india-44856910`.

[62] John Bohannon. Mechanical Turk upends social sciences. *Science*, 352(6291):1263–1264, June 2016.

[63] Erin Brady, Meredith Ringel Morris, and Jeffrey P. Bigham. Gauging Receptiveness to Social Microvolunteering. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1055–1064, New York, NY, USA, 2015. ACM.

[64] Erin L. Brady, Yu Zhong, Meredith Ringel Morris, and Jeffrey P. Bigham. Investigating the Appropriateness of Social Network Question Asking As a Resource for Blind Users. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1225–1236, New York, NY, USA, 2013. ACM.

[65] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[66] Stacy Butler, Adele Crudden, and William Sansing. Overcoming barriers to employment: Strategies of rehabilitation providers. *Journal of Visual Impairment & Blindness (JVIB)*, 99(06), 2005.

[67] Anthony Candela and Karen Wolffe. A qualitative analysis of employers' experiences with visually impaired workers. *Journal of Visual Impairment & Blindness (JVIB)*, 96(09), 2002.

[68] Dipanjan Chakraborty, Akshay Gupta, and Aaditeshwar Seth. Experiences from a Mobile-based Behaviour Change Campaign on Maternal and Child Nutrition in Rural India. In *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development*, ICTD '19, pages 20:1–20:11, New York, NY, USA, 2019. ACM.

[69] Dipanjan Chakraborty, Indrani Medhi, Edward Cutrell, and William Thies. Man Versus Machine: Evaluating IVR Versus a Live Operator for Phone Surveys in India. In *Proceedings of the 3rd ACM Symposium on Computing for Development*, ACM DEV '13, pages 7:1–7:9, New York, NY, USA, 2013. ACM.

[70] Manu Chopra, Indrani Medhi Thies, Joyojeet Pal, Colin Scott, William Thies, and Vivek Seshadri. Exploring Crowdsourced Work in Low-Resource Settings. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 381:1–381:13, New York, NY, USA, 2019. ACM.

[71] Jennifer Cole, Timothy Mahrt, and Joseph Roy. Crowd-sourcing prosodic annotation. *Computer Speech & Language*, 45:300–325, September 2017.

[72] Adele Crudden, Lynn W. McBroom, Amy L. Skinner, and J. Elton Moore. *Comprehensive Examination of Barriers to Employment among Persons Who Are Blind or Visually Impaired*. Mississippi State University, Rehabilitation Research and Training Center on Blindness and Low Vision, P, May 1998.

[73] Sebastien Cuendet, Indrani Medhi, Kalika Bali, and Edward Cutrell. VideoKheti: Making Video Content Accessible to Low-literate and Novice Users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2833–2842, New York, NY, USA, 2013. ACM.

[74] Leslie L. Dodson, S. Revi Sterling, and John K. Bennett. Minding the Gaps: Cultural, Technical and Gender-based Barriers to Mobile Use in Oral-language Berber Communities in Morocco. In *Proceedings of the Sixth International Conference on Information and Communication Technologies and Development: Full Papers - Volume 1*, ICTD '13, pages 79–88, New York, NY, USA, 2013. ACM.

[75] Krittika D'Silva, Meghana Marathe, Aditya Vashistha, Gaetano Borriello, and William Thies. A Mobile Application for Interactive Voice Forums: Design and Pilot Deployment in Rural India. In *Proceedings of the Fifth ACM Symposium on Computing for Development*, ACM DEV-5 '14, pages 121–122, New York, NY, USA, 2014. ACM.

[76] Karn Dubey, Palash Gupta, Rachna Shriwas, Gayatri Gulvady, and Amit Sharma. Learnings

from Deploying a Voice-based Social Platform for People with Disability. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS '19, New York, NY, USA, 2019. ACM.

[77] Nathan Eagle. txteagle: Mobile Crowdsourcing. In *Internationalization, Design and Global Development*, pages 447–456, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[78] Daniel A. Epstein, Bradley H. Jacobson, Elizabeth Bales, David W. McDonald, and Sean A. Munson. From "Nobody Cares" to "Way to Go!": A Design Framework for Social Sharing in Personal Informatics. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1622–1636, New York, NY, USA, 2015. ACM.

[79] Keelan Evanini and Klaus Zechner. Using Crowdsourcing to Provide Prosodic Annotations for Non-Native Speech. In *Proceedings of 12th Annual Conference of the International Speech Communication Association*, INTERSPEECH '11, pages 3069–3072, 2011.

[80] Jason Gainous and Kevin M. Wagner. *Tweeting to Power: The Social Media Revolution in American Politics*. Oxford University Press, November 2013. Google-Books-ID: cc48BAAAQBAJ.

[81] Yashesh Gaur, Florian Metze, and Jeffrey P. Bigham. Manipulating Word Lattices to Incorporate Human Corrections. In *Proceedings of 17th Annual Conference of the International Speech Communication Association*, INTERSPEECH '16, pages 3062–3065, September 2016.

[82] Vindu Goel, Suhasini Raj, and Priyadarshini Ravichandran. How WhatsApp Leads Mobs to Murder in India. *The New York Times*, July 2018. `https://www.nytimes.com/interactive/2018/07/18/technology/whatsapp-india-killings.html`.

[83] David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. Using Collaborative Filtering to Weave an Information Tapestry. *Commun. ACM*, 35(12):61–70, December 1992.

[84] Joseph K. Goodman, Cynthia E. Cryder, and Amar Cheema. Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*, 26(3):213–224, 2013.

[85] Mark Graham and Jamie Woodcock. Towards a Fairer Platform Economy: Introducing the Fairwork Foundation. *Alternate Routes: A Journal of Critical Social Research*, 29(0), 2018.

[86] Aditi Sharma Grover, Karen Calteaux, Etienne Barnard, and Gerhard van Huyssteen. A Voice Service for User Feedback on School Meals. In *Proceedings of the 2nd ACM Symposium on Computing for Development*, ACM DEV '12, pages 13:1–13:10, New York, NY, USA, 2012. ACM.

[87] Mohamed Gulaid and Aditya Vashistha. Ila Dhageyso: An Interactive Voice Forum to Foster Transparent Governance in Somaliland. In *Proceedings of the Sixth International Conference on Information and Communications Technologies and Development: Notes - Volume 2*, ICTD '13, pages 41–44, New York, NY, USA, 2013. ACM.

[88] Aakar Gupta, William Thies, Edward Cutrell, and Ravin Balakrishnan. mClerk: Enabling Mobile Crowdsourcing in Developing Regions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1843–1852, New York, NY, USA, 2012. ACM.

[89] Toru Imai, Atsushi Matsui, Shinichi Homma, Takeshi Kobayakawa, Kazuo Onoe, Shoei Sato, and Akio Ando. Speech Recognition with a Re-speak Method for Subtitling Live Broadcasts. In *Proceedings of 7th International Conference on Spoken Language Processing*, ICSLP '02 - INTERSPEECH '02, Denver, Colorado, USA, September 2002.

[90] Rishi Iyengar. India's mobile price war just claimed another victim, February 2018. `https://money.cnn.com/2018/02/28/technology/aircel-bankruptcy-india-mobile-price-war/index.html`.

[91] Anirudha Joshi, Mandar Rane, Debjani Roy, Nagraj Emmadi, Padma Srinivasan, N. Kumarasamy, Sanjay Pujari, Davidson Solomon, Rashmi Rodrigues, D.G. Saple, Kamalika Sen, Els Veldeman, and Romain Rutten. Supporting Treatment of People Living with HIV / AIDS in Resource Limited Settings with IVRs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 1595–1604, New York, NY, USA, 2014. ACM.

[92] Shilpa Kannan. Jio: Telecom giant Reliance sparks India price war. *BBC News*, September 2016. `http://www.bbc.com/news/business-37273073`.

[93] J. Katz, M. Barris, and A. Jain. *The Social Media President: Barack Obama and the Politics of Digital Engagement*. Springer, December 2013. Google-Books-ID: 1jTFAgAAQBAJ.

[94] Micha Kaufman. The Internet Revolution is the New Industrial Revolution. *Forbes*, October 2012. `https://www.forbes.com/sites/michakaufman/2012/10/05/the-internet-revolution-is-the-new-industrial-revolution/`.

[95] Konstantinos Kazakos, Siddhartha Asthana, Madeline Balaam, Mona Duggal, Amey Holden, Limalemla Jamir, Nanda Kishore Kannuri, Saurabh Kumar, Amarendar Reddy Manindla, Subhashini Arcot Manikam, GVS Murthy, Papreen Nahar, Peter Phillimore, Shreyaswi Sathyanath, Pushpendra Singh, Meenu Singh, Pete Wright, Deepika Yadav, and Patrick Olivier. A Real-Time IVR Platform for Community Radio. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 343–354, New York, NY, USA, 2016. ACM.

[96] Shashank Khanna, Aishwarya Ratan, James Davis, and William Thies. Evaluating and Improving the Usability of Mechanical Turk for Low-income Workers in India. In *Proceedings of the First ACM Symposium on Computing for Development*, ACM DEV '10, pages 12:1–12:10, New York, NY, USA, 2010. ACM.

[97] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John

Zimmerman, Matt Lease, and John Horton. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1301–1318, New York, NY, USA, 2013. ACM.

[98] Zahir Koradia, Piyush Aggarwal, Aaditeshwar Seth, and Gaurav Luthra. Gurgaon Idol: A Singing Competition over Community Radio and IVRS. In *Proceedings of the 3rd ACM Symposium on Computing for Development*, ACM DEV '13, pages 6:1–6:10, New York, NY, USA, 2013. ACM.

[99] Zahir Koradia and Aaditeshwar Seth. PhonePeti: Exploring the Role of an Answering Machine System in a Community Radio Station in India. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, ICTD '12, pages 278–288, New York, NY, USA, 2012. ACM.

[100] Anand Kulkarni, Philipp Gutheim, Prayag Narula, Dave Rolnitzky, Tapan Parikh, and Bjorn Hartmann. MobileWorks: Designing for Quality in a Managed Crowdsourcing Architecture. *IEEE Internet Computing*, 16(5):28–35, September 2012.

[101] Jennifer Lai and John Vergo. MedSpeak: Report Creation with Continuous Speech Recognition. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI '97, pages 431–438, New York, NY, USA, 1997. ACM.

[102] Ian Lane, Alex Waibel, Matthias Eck, and Kay Rottmann. Tools for Collecting Speech Corpora via Mechanical-Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 184–187, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[103] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. Real-time Captioning by Groups of Non-experts. In

Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology, UIST '12, pages 23–34, New York, NY, USA, 2012. ACM.

[104] Jonathan Ledlie, Billy Odero, Einat Minkov, Imre Kiss, and Joseph Polifroni. Crowd Translator: On Building Localized Speech Recognizers Through Micropayments. *ACM SIGOPS Operating Systems Review*, 43(4):84–89, January 2010.

[105] Chia-ying Lee and James Glass. A Transcription Task for Crowdsourcing with Automatic Quality Control. In *Proceedings of 12th Annual Conference of the International Speech Communication Association*, INTERSPEECH '11, pages 3041–3044, Florence, Italy, August 2011.

[106] Adam Lerer, Molly Ward, and Saman Amarasinghe. Evaluation of IVR Data Collection UIs for Untrained Rural Users. In *Proceedings of the First ACM Symposium on Computing for Development*, ACM DEV '10, pages 2:1–2:8, New York, NY, USA, 2010. ACM.

[107] Dennis Linders. From e-government to we-government: Defining a typology for citizen coproduction in the age of social media. *Government Information Quarterly*, 29(4):446–454, October 2012.

[108] Michael A Madaio, Vikram Kamath, Evelyn Yarzebinski, Shelby Zasacky, Fabrice Tanoh, Joelle Hannon-Cropp, Justine Cassell, Kaja Jasinska, and Amy Ogan. "You Give a Little of Yourself": Family Support for Children's Use of an IVR Literacy System. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS '19, page 13, New York, NY, USA. ACM.

[109] Meghana Marathe, Jacki O'Neill, Paromita Pain, and William Thies. Revisiting CGNet Swara and Its Impact in Rural India. In *Proceedings of the Seventh International Conference on Information and Communication Technologies and Development*, ICTD '15, pages 21:1–21:10, New York, NY, USA, 2015. ACM.

[110] Meghana Marathe, Jacki O'Neill, Paromita Pain, and William Thies. ICT-Enabled Grievance Redressal in Central India: A Comparative Analysis. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*, ICTD '16, pages 4:1–4:11, New York, NY, USA, 2016. ACM.

[111] Eleanor R. Marchant. Interactive Voice Response and Radio for Peacebuilding: A Macro View of the Literature and Experiences from the Field. Technical report, Annenberg School for Communication, February 2016.

[112] Alison Mathie and Gord Cunningham. From Clients to Citizens: Asset-Based Community Development as a Strategy for Community-Driven Development. *Development in Practice*, 13(5):474–486, 2003.

[113] Ian McGraw, Alexander Gruenstein, and Andrew M. Sutherland. A self-labeling speech corpus: Collecting spoken words with an online educational game. In *Proceedings of 10th Annual Conference of the International Speech Communication Association*, INTERSPEECH '09, pages 3031–3034, January 2009.

[114] Ian McGraw, Chia-ying Lee, I Lee Hetherington, Stephanie Seneff, and James Glass. Collecting Voices from the Cloud. In *Proceedings of the International Conference on Language Resources and Evaluation*, LREC '10, pages 1576–1583, Valletta, Malta, May 2010.

[115] Timothy McLaughlin. How WhatsApp Fuels Fake News and Violence in India. *Wired*, December 2018.

[116] Roger McNamee. Opinion | A Brief History of How Your Privacy Was Stolen. *The New York Times*, June 2019.

[117] Indrani Medhi, Somani Patnaik, Emma Brunskill, S.N. Nagasena Gautama, William Thies, and Kentaro Toyama. Designing Mobile Interfaces for Novice and Low-literacy Users. *ACM Trans. Comput.-Hum. Interact.*, 18(1):2:1–2:28, May 2011.

[118] Aparna Moitra, Vishnupriya Das, Gram Vaani, Archna Kumar, and Aaditeshwar Seth. Design Lessons from Creating a Mobile-based Community Media Platform in Rural India. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*, ICTD '16, pages 14:1–14:11, New York, NY, USA, 2016. ACM.

[119] Preeti Mudliar, Jonathan Donner, and William Thies. 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, pages 159–168, New York, NY, USA, 2012. ACM.

[120] Preeti Mudliar, Jonathan Donner, and William Thies. Emergent Practices Around CGNet Swara: A Voice Forum for Citizen Journalism in Rural India. *Information Technologies & International Development*, 9(2):pp. 65–79–79, June 2013.

[121] Randall Munroe. Reddit's new comment sorting system, October 2009. `http://www.redditblog.com/2009/10/reddits-new-comment-sorting-system.html`.

[122] Iftekhar Naim, Daniel Gildea, Walter Lasecki, and Jeffrey P. Bigham. Text Alignment for Real-Time Crowd Captioning. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 201–210, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[123] Prayag Narula, Philipp Gutheim, David Rolnitzky, Anand Kulkarni, and Bjoern Hartmann. MobileWorks: A Mobile Crowdsourcing Platform for Workers at the Bottom of the Pyramid. In *Proceedings of the 11th AAAI Conference on Human Computation*, AAAIWS'11-11, pages 121–123. AAAI Press, 2011.

[124] Xavier Ochoa and Erik Duval. Quantitative analysis of user-generated content on the web. In *Proceedings of WebEvolve2008: web science workshop at WWW2008*, pages 1–8, April 2008.

[125] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. In *Eighth International AAAI Conference on Weblogs and Social Media*, May 2014.

[126] Joyojeet Pal. Banalities Turned Viral: Narendra Modi and the Political Tweet. *Television & New Media*, 16(4):378–387, May 2015.

[127] Igor Pantic. Online Social Networking and Mental Health. *Cyberpsychology, Behavior and Social Networking*, 17(10):652–657, October 2014.

[128] Gabriel Parent and Maxine Eskenazi. Toward better crowdsourced transcription: Transcription of a year of the Let's Go Bus Information System data. In *2010 IEEE Spoken Language Technology Workshop*, pages 312–317, December 2010.

[129] Donatella Pascolini and Silvio Paolo Mariotti. Global estimates of visual impairment: 2010. *British Journal of Ophthalmology*, 96(5):614–618, May 2012.

[130] Neil Patel, Sheetal Agarwal, Nitendra Rajput, Amit Nanavati, Paresh Dave, and Tapan S. Parikh. A Comparative Study of Speech and Dialed Input Voice Interfaces in Rural India. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 51–54, New York, NY, USA, 2009. ACM.

[131] Neil Patel, Deepti Chittamuru, Anupam Jain, Paresh Dave, and Tapan S. Parikh. 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, pages 733–742, New York, NY, USA, 2010. ACM.

[132] Neil Patel, Kapil Shah, Krishna Savani, Scott R. Klemmer, Paresh Dave, and Tapan S. Parikh. Power to the Peers: Authority of Source Effects for a Voice-based Agricultural Information Service in Rural India. In *Proceedings of the Fifth International Conference on Information*

and Communication Technologies and Development, ICTD '12, pages 169–178, New York, NY, USA, 2012. ACM.

[133] Jennifer Pearson, Simon Robinson, Matt Jones, Amit Nanavati, and Nitendra Rajput. ACQR: Acoustic Quick Response Codes for Content Sharing on Low End Phones with No Internet Connectivity. In *Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services*, MobileHCI '13, pages 308–317, New York, NY, USA, 2013. ACM.

[134] Ales Prazák, Zdenek Loose, Jan Trmal, Josef V. Psutka, and Josef Psutka. Novel Approach to Live Captioning Through Re-speaking: Tailoring Speech Recognition to Re-speaker's Needs. In *Proceedings of 13th Annual Conference of the International Speech Communication Association*, INTERSPEECH '12, pages 1372–1375, 2012.

[135] Alexander J. Quinn and Benjamin B. Bederson. Human Computation: A Survey and Taxonomy of a Growing Field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1403–1412, New York, NY, USA, 2011. ACM.

[136] Nimmi Rangaswamy and Edward Cutrell. Anthropology, Development and ICTs: Slums, Youth and the Mobile Internet in Urban India. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, ICTD '12, pages 85–93, New York, NY, USA, 2012. ACM.

[137] Agha Ali Raza, Awais Athar, Shan Randhawa, Zain Tariq, Muhammad Bilal Saleem, Haris Bin Zia, Umar Saif, and Roni Rosenfeld. Rapid Collection of Spontaneous Speech Corpora Using Telephonic Community Forums. In *Interspeech 2018*, pages 1021–1025. ISCA, September 2018.

[138] Agha Ali Raza, Rajat Kulshreshtha, Spandana Gella, Sean Blagsvedt, Maya Chandrasekaran, Bhiksha Raj, and Roni Rosenfeld. Viral Spread via Entertainment and Voice-Messaging

Among Telephone Users in India. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*, ICTD '16, pages 1:1–1:10, New York, NY, USA, 2016. ACM.

[139] Agha Ali Raza, Mansoor Pervaiz, Christina Milo, Samia Razaq, Guy Alster, Jahanzeb Sherwani, Umar Saif, and Roni Rosenfeld. 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, pages 350–359, New York, NY, USA, 2012. ACM.

[140] Agha Ali Raza, Bilal Saleem, Shan Randhawa, Zain Tariq, Awais Athar, Umar Saif, and Roni Rosenfeld. Baang: A Viral Speech-based Social Platform for Under-Connected Populations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 643:1–643:12, New York, NY, USA, 2018. ACM.

[141] Agha Ali Raza, Farhan Ul Haq, Zain Tariq, Mansoor Pervaiz, Samia Razaq, Umar Saif, and Roni Rosenfeld. Job Opportunities Through Entertainment: Virally Spread Speech-based Services for Low-literate Users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2803–2812, New York, NY, USA, 2013. ACM.

[142] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW '94, pages 175–186, New York, NY, USA, 1994. ACM.

[143] Waleed Riaz, Haris Durrani, Suleman Shahid, and Agha Ali Raza. ICT Intervention for Agriculture Development: Designing an IVR System for Farmers in Pakistan. In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*, ICTD '17, pages 33:1–33:5, New York, NY, USA, 2017. ACM.

[144] Kevin Roose. Can Social Media Be Saved? *The New York Times*, March 2018.

[145] Nithya Sambasivan, Ed Cutrell, Kentaro Toyama, and Bonnie Nardi. Intermediated Technology Use in Developing Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 2583–2592, New York, NY, USA, 2010. ACM.

[146] Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. Intersectional HCI: Engaging Identity Through Gender, Race, and Class. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 5412–5427, New York, NY, USA, 2017. ACM.

[147] Alana Semuels. The Internet Is Enabling a New Kind of Poorly Paid Hell. *The Atlantic*, January 2018. `https://www.theatlantic.com/business/archive/2018/01/amazon-mechanical-turk/551192/`.

[148] Upendra Shardanand and Pattie Maes. Social Information Filtering: Algorithms for Automating "Word of Mouth". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '95, pages 210–217, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.

[149] J. Sherwani, N. Ali, S. Mirza, A. Fatma, Y. Memon, M. Karim, R. Tongia, and R. Rosenfeld. Healthline: Speech-based access to health information by low-literate users. In *2007 International Conference on Information and Communication Technologies and Development*, pages 1–9, Dec 2007.

[150] Sujit Shinde, Divya Piplani, Karthik Srinivasan, Dineshkumar Singh, Rahul Sharma, and Preetam Mohnaty. mKRISHI: Simplification Of IVR Based Services For Rural Community. In *Proceedings of the India HCI 2014 Conference on Human Computer Interaction*, IndiaHCI '14, pages 154:154–154:159, New York, NY, USA, 2014. ACM.

[151] Clay Shirky. The Political Power of Social Media. January 2016. https://www.foreignaffairs.com/articles/2010-12-20/political-power-social-media.

[152] Venkatesh Sivaraman, Dongwook Yoon, and Piotr Mitros. Simplified Audio Production in Asynchronous Voice-Based Discussions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 1045–1054, New York, NY, USA, 2016. ACM.

[153] Thomas N. Smyth, Satish Kumar, Indrani Medhi, and Kentaro Toyama. Where There's a Will There's a Way: Mobile Media Sharing in Urban India. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 753–762, New York, NY, USA, 2010. ACM. event-place: Atlanta, Georgia, USA.

[154] Matthias Sperber, Graham Neubig, Christian Fügen, Satoshi Nakamura, and Alexander H. Waibel. Efficient speech transcription through respeaking. In *Proceedings of 14th Annual Conference of the International Speech Communication Association*, INTERSPEECH '13, pages 1087–1091, 2013.

[155] Kate Starbird and Leysia Palen. "Voluntweeters": Self-organizing by Digital Volunteers in Times of Crisis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1071–1080, New York, NY, USA, 2011. ACM.

[156] Kate Starbird and Leysia Palen. (How) Will the Revolution Be Retweeted?: Information Diffusion and the 2011 Egyptian Uprising. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 7–16, New York, NY, USA, 2012. ACM.

[157] Xiaoyuan Su and Taghi M. Khoshgoftaar. A Survey of Collaborative Filtering Techniques. *Adv. in Artif. Intell.*, 2009:4:2–4:2, January 2009.

[158] R. Talhouk, T. Bartindale, K. Montague, S. Mesmar, C. Akik, A. Ghassani, M. Najem, H. Ghattas, P. Olivier, and M. Balaam. Implications of Synchronous IVR Radio on Syrian Refugee Health and Community Dynamics. In *Proceedings of the 8th International Conference on Communities and Technologies*, C&T '17, pages 193–202, New York, NY, USA, 2017. ACM.

[159] Saumya Tewari. 75 percent of rural India survives on Rs 33 per day, 2015. `https://www.indiatoday.in/india/story/india-rural-household-650-millions-live-on-rs-33-per-day-282195-2015-07-13`.

[160] Aditya Vashistha, Erin Brady, William Thies, and Edward Cutrell. Educational Content Creation and Sharing by Low-Income Visually Impaired People in India. In *Proceedings of the Fifth ACM Symposium on Computing for Development*, ACM DEV-5 '14, pages 63–72, New York, NY, USA, 2014. ACM.

[161] Aditya Vashistha, Edward Cutrell, Gaetano Borriello, and William Thies. 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, pages 417–426, New York, NY, USA, 2015. ACM.

[162] Aditya Vashistha, Edward Cutrell, Nicola Dell, and Richard Anderson. Social Media Platforms for Low-Income Blind People in India. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, ASSETS '15, pages 259–272, New York, NY, USA, 2015. ACM.

[163] Aditya Vashistha, Abhinav Garg, and Richard Anderson. ReCall: Crowdsourcing on Basic Phones to Financially Sustain Voice Forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 169:1–169:13, New York, NY, USA, 2019. ACM.

[164] Aditya Vashistha, Abhinav Garg, Richard Anderson, and Agha Ali Raza. Threats, Abuses, Flirting, and Blackmail: Gender Inequity in Social Media Voice Forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 72:1–72:13, New York, NY, USA, 2019. ACM.

[165] Aditya Vashistha, Neha Kumar, Anil Mishra, and Richard Anderson. Mobile Video Dissemination for Community Health. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*, ICTD '16, pages 20:1–20:11, New York, NY, USA, 2016. ACM.

[166] Aditya Vashistha, Pooja Sethi, and Richard Anderson. Respeak: A Voice-based, Crowd-powered Speech Transcription System. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 1855–1866, New York, NY, USA, 2017. ACM.

[167] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 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*, CHI '18, pages 57:1–57:13, New York, NY, USA, 2018. ACM.

[168] Aditya Vashistha and William Thies. IVR junction: Building scalable and distributed voice forums in the developing world. In *Presented as part of the 6th USENIX/ACM Workshop on Networked Systems for Developing Regions*, Boston, MA, 2012. USENIX.

[169] Jerome White, Mayuri Duggirala, Krishna Kummamuru, and Saurabh Srivastava. Designing a Voice-based Employment Exchange for Rural India. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, ICTD '12, pages 367–373, New York, NY, USA, 2012. ACM.

[170] Derek Willis. Narendra Modi, the Social Media Politician. *The New York Times*, September

2014. https://www.nytimes.com/2014/09/26/upshot/narendra-modi-the-social-media-politician.html.

[171] Nikolas Wolfe, Juneki Hong, Agha Ali Raza, Bhiksha Raj, and Ronald Rosenfeld. Rapid development of public health education systems in low-literacy multilingual environments: combating ebola through voice messaging. In *SLaTE*, 2015.

[172] Marisol Wong-Villacres, Arkadeep Kumar, Aditya Vishwanath, Naveena Karusala, Betsy DiSalvo, and Neha Kumar. Designing for Intersections. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, pages 45–58, New York, NY, USA, 2018. ACM.

[173] Deepika Yadav, Pushpendra Singh, Kyle Montague, Vijay Kumar, Deepak Sood, Madeline Balaam, Drishti Sharma, Mona Duggal, Tom Bartindale, Delvin Varghese, and Patrick Olivier. Sangoshthi: Empowering Community Health Workers Through Peer Learning in Rural India. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 499–508, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[174] Dongwook Yoon, Nicholas Chen, François Guimbretière, and Abigail Sellen. RichReview: Blending Ink, Speech, and Gesture to Support Collaborative Document Review. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, pages 481–490, New York, NY, USA, 2014. ACM.

[175] Yu Zhong, Walter S. Lasecki, Erin Brady, and Jeffrey P. Bigham. RegionSpeak: Quick Comprehensive Spatial Descriptions of Complex Images for Blind Users. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2353–2362, New York, NY, USA, 2015. ACM.

[176] Kathryn Zyskowski, Meredith Ringel Morris, Jeffrey P. Bigham, Mary L. Gray, and Shaun K. Kane. Accessible Crowdwork?: Understanding the Value in and Challenge of Microtask Em-

ployment for People with Disabilities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1682–1693, New York, NY, USA, 2015. ACM.