# ReCall: Crowdsourcing on Basic Phones to Financially Sustain Voice Forums

**Aditya Vashistha**
University of Washington
adityav@cs.washington.edu

**Abhinav Garg**
University of Washington
aagarg@uw.edu

**Richard Anderson**
University of Washington
anderson@cs.washington.edu

## ABSTRACT

Although voice forums are widely used to enable marginalized communities to produce, consume, and share information, their financial sustainability is a key concern among HCI4D researchers and practitioners. We present *ReCall*, a crowdsourcing marketplace accessible via phone calls where low-income rural residents vocally transcribe audio files to gain free airtime to participate in voice forums as well as to earn money. We conducted a series of experimental and usability evaluations with 28 low-income people in rural India to examine the effect of phone types, channel types, and review modes on speech transcription performance. We then deployed *ReCall* for two weeks to 24 low-income rural residents who placed 5,879 phone calls, completed 29,000 micro tasks to yield transcriptions with 85% accuracy, and earned ₹20,500. Our mixed-methods analysis indicates that each minute of crowd work on *ReCall* gives users eight minutes of free airtime on another voice forum, and thus illustrates a way to address the financial sustainability of voice forums.

## CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**.

## KEYWORDS

HCI4D; voice forums; crowdsourcing; financial sustainability.

**ACM Reference Format:**
Aditya Vashistha, Abhinav Garg, and Richard Anderson. 2019. ReCall: Crowdsourcing on Basic Phones to Financially Sustain Voice Forums. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3290605.3300399

*CHI 2019, May 4–9, 2019, Glasgow, Scotland UK*
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-5970-2/19/05...$15.00
https://doi.org/10.1145/3290605.3300399

## 1 INTRODUCTION

Mainstream social computing technologies—like social media platforms, online discussion forums, or crowdsourcing marketplaces—have revolutionized how literate people with access to smartphones and the Internet participate in the information ecology and digital economy. However, these technologies are currently excluding billions of those who are illiterate, who speak low-resource languages, who live in poverty, or who do not have access to Internet-connected devices. These literacy, language, socioeconomic, and connectivity barriers result in *"utility gaps"* [24], limiting mobile phone use to making and receiving voice calls.

Recognizing these structural limitations, Human-Computer Interaction for Development (HCI4D) researchers and practitioners have leveraged the ubiquity of basic mobile phones and accessibility of voice to design *voice forums* that allow users to access, report, and share information in their local language via phone calls. Users of these services call a toll-free phone number to record voice messages, listen to messages recorded by others, and indicate their preferences (e.g., likes, dislikes). These services have amassed *millions* of low-income users, phone calls, and voice messages, and have been used in diverse HCI4D contexts, such as health information systems [29, 31], civic engagement portals [19, 26, 35], agriculture advisory services [37], and job portals [40, 46]. For example, *Mobile Kunji* [16]—a health information service in India—has disseminated over 42 million minutes of health content on phone calls to nearly half a million people.

Although voice forums have proven themselves a usable and accessible communication medium for people with literacy, language, socioeconomic, or connectivity barriers, their financial sustainability is a major concern among HCI4D researchers and practitioners. Since low-income people are unable to pay for the cost of phone calls even to services they find useful [1, 42], voice forums rely on expensive toll-free lines to make them accessible to callers. This leads to challenges in sustaining these services especially with increased usage [38] and puts them at risk of being shut down due to high operating costs. While a few services sustain themselves through advertising [15], external grants [14], and partnerships with mobile network operators (MNOs) or governments [11], these alternatives are often beyond the reach of bottom-up development-focused voice forums. There is an urgent need to explore alternatives to financially sustain voice forums.

In this paper, we examine whether low-income people in rural areas could complete useful work on their mobile phones to offset the participation costs of voice forums. Since existing crowdsourcing marketplaces such as *Mechanical Turk* [3] and *CrowdFlower* [6] are unfeasible in rural regions, we designed and built a new crowdsourcing marketplace that leverages familiarity with local language, the power of voice, and the ubiquity of basic phones to circumvent language, literacy, and connectivity barriers present in rural India.

In prior work, we created *Respeak* [44]—a smartphone-based crowdsourcing marketplace where people transcribe audio files by speaking into Android's automatic speech recognition (ASR) engine instead of typing—and evaluated it with literate smartphone users in an Indian metropolis. In this work, we draw inspiration from the voice-based design of *Respeak* to create *ReCall*, an Interactive Voice Response (IVR) based crowdsourcing marketplace accessible via ordinary phone calls where low-income rural residents vocally transcribe Hindi audio files to subsidize participation costs of voice forums as well as to supplement their income. Although *Respeak* and *ReCall* use the same underlying concept, there are fundamental differences between them with respect to the type of devices users use to complete tasks, the type of channel the applications use and the resultant quality of audio files submitted to ASR engine, the mode to review crowd work, and the demography of target users.

We conducted three controlled experiments with 28 low-income rural residents to examine how adaptations of *Respeak* to design and build *ReCall* affect crowd workers' performance on three key activities they perform—listening to an audio segment, re-speaking the content into an ASR engine, and reviewing the correctness of ASR-generated transcript—to complete a speech transcription task. In particular, we examined how phone types, channel types, and review modes affect task completion time, number of trials, and accuracy. We also conducted a usability study comparing *ReCall* and *Respeak* to examine the cumulative effect of these adaptations on usability perceptions, user experience, and transcription performance.

To examine the feasibility and acceptability of *ReCall*, we conducted a two-week field deployment with 24 low-income people in rural India. *ReCall* users collectively placed 5,879 phone calls to complete about 29,000 speech transcription tasks with an average accuracy of 73.3%, and earned ₹20,500 (USD 310)[1] for completing tasks. *ReCall* used multiple string alignment and a majority voting process to reduce random speech transcription errors. It aligned transcripts generated by 11 users to produce transcription with 85% accuracy and at a cost of USD 1.86 per minute of audio content, almost one-third of the market cost of Hindi transcription.

Our findings demonstrated that low-income rural residents can complete useful work on their mobile phones and generate enough profits to subsidize their participation costs of voice forums. Our analysis indicated that each minute spent in completing crowd work on *ReCall* could provide about eight minutes of free airtime on voice forums while also enabling *ReCall* users to earn money at an hourly rate comparable to the average hourly wage rate in India. In addition to examining the feasibility and acceptability of *ReCall*, we also conducted a usability study with ten low-income rural residents to examine users' willingness to complete tasks on *ReCall* to subsidize their phone calls to *Connect*, a social media voice forum. We found that switching between these two services did not affect the usability and user experience of participants on both *ReCall* and *Connect*.

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 discuss the lessons learned from the experimental evaluations and field deployment, as well as provide suggestions to improve *ReCall* and to integrate it into large-scale voice forums.

## 2 RELATED WORK

Although voice forums have demonstrated their potential to empower people who experience a complex array of literacy, language, socioeconomic, and connectivity barriers [19, 29, 35, 39, 43], there is a growing concern about their high operating costs since these services are often offered as toll-free lines to invite participation from low-income people [38, 42]. Most voice forums rely on external funding to subsidize the cost of voice calls, however, the unreliable nature of grants and awards makes it an unsustainable approach. For example, the founder of *CGNet Swara*—a popular service that enable rural communities in India to report and listen to locally relevant news and grievances—expressed frustrations on how limited funding to subsidize phone calls may cause them to *"shut down completely"* [1]. Some voice forums have conducted experiments to examine users' willingness to bear the cost of voice calls, however, the outcome of these experiments is discouraging at best [1, 42].

A few voice forums such as *Kan Khajura Tesan* [15]—an on-demand entertainment service in India from a consumer goods

[1] In this paper, we use an exchange rate of USD 1 = ₹66 (INR 66).

company with USD 5 billion revenue—and *Gram Vaani* [2]—a voice-based social media service with over 1.5 million users in central and north-eastern India—have used advertising revenues to subsidize the cost of voice calls. 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* [26]—a service to connect citizens with government officials in Somaliland—and *3-2-1 service* [11]—a phone call-based search engine in Africa—have partnered with government agencies and MNOs to subsidize the cost of voice calls. However, building and maintaining such partnerships is seldom possible due to mismatch in goals, expectations, and values [26]. Limitations in these existing approaches to subsidize the cost of phone calls to voice forums motivated us to address the financial sustainability challenge. Our work contributes a novel solution in which profits from crowd work by low-income callers on basic phones is used to subsidize the call costs of voice forums.

Mainstream crowdsourcing marketplaces such as *Mechanical Turk* and *CrowdFlower* require access to the Internet, computers, and English language skills, making them unfeasible and unusable for people in low-resource environments. Several HCI4D researchers have designed new crowdsourcing marketplaces to circumvent these literacy, language, and connectivity barriers. For example, *Samasource* [10] establishes outsourcing centers in low-resource regions where people living in poverty are trained in image annotation and other services. *mClerk* [27] and *MobileWorks* [33, 36] incentivizes low-income people to transcribe images sent to their phones via SMS and a mobile web-based application, respectively. *TxtEagle* and *mSurvey* incentivizes low-income people to answer SMS-based surveys. A major limitation of these systems is that they expect crowd workers to have reading and typing skills. *Respeak* and *BSpeak* [44, 45] overcomes literacy barriers by enabling smartphone users to complete micro tasks by leveraging their speaking and listening skills. *Jana* [9] provides airtime to users to watch videos, listen to songs, and download new smartphone applications. However, these systems require users to have access to a smartphone, making them unfeasible for people who own basic or feature phones. *ReCall* extends this literature by demonstrating the feasibility, acceptability, and usability of a new crowdsourcing marketplace accessible via ordinary phone calls from any phone where low-income rural residents engage in crowd work by using their listening and speaking skills.

## 3 RECALL: SYSTEM AND APPLICATION DESIGN

*ReCall* system has two main components: *ReCall* engine and *ReCall* application. The *ReCall* engine engages in three main activities to transcribe an audio file:

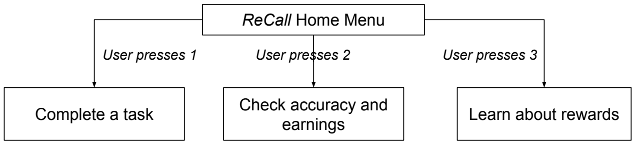

**Figure 1: High-level call flow of the *ReCall* application.**

- **Segmentation:** It segments the audio file, based on the speaking rate and occurrence of natural pauses, to yield audio segments that are typically 3–6 seconds long.
- **Distribution:** It distributes these segments to multiple *ReCall* application users who produce transcripts.
- **Merging:** For each segment, it combines transcripts from multiple users by using multiple string alignment (MSA) and a majority voting process to generate a best estimation transcript. The engine compares individual transcripts submitted by users to the best estimation transcript for determining users' reward. It concatenates the best estimation transcripts for all segments to yield the final transcript of the audio file.

To transcribe segments, users call the phone number of the *ReCall* application. Figure 1 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 application in a quiet environment. Once users re-speak the content, the *ReCall* application 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.
- **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 [22, 23, 37, 42] to make the *ReCall* application 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.

**Table 1: Key differences between *ReCall* and *Respeak***

| | ***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 |

Although *ReCall* and *Respeak* use the same underlying engine for segmenting audio files, distributing micro tasks, and merging transcripts, Table 1 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 application (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 metropolitan residents.

## 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 examine 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.

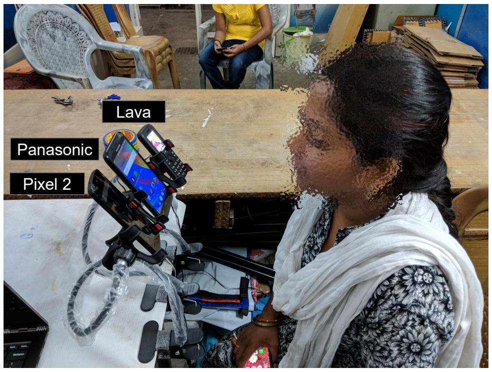

**Figure 2: A participant speaking sentences simultaneously in Pixel 2, Panasonic P100, and Lava Captain N1.**

### 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 ($90 Panasonic P100) next to each other (see Figure 2). 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.

**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 [21].

### Recruitment and Demographic Details

We partnered with NYST, a grassroots organization that has active projects on community health and education in rural India. Leveraging the network of their employees, 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.

### 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 × 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 3 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 segments 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 downsampling 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.

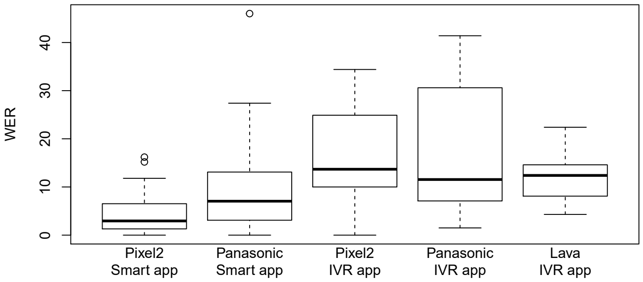

**Figure 3: Distribution of WERs for different combinations of phone types and channel types.**

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*. Compared to *Respeak*, participants took more time to complete tasks on *ReCall* 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.

**Table 2: Median scores of different usability parameters on a ten-point scale (1–low, 10–high) for *ReCall* and *Respeak*.**

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

Participants experienced higher mental demand, effort, and frustration, and lower performance on *ReCall* than on *Respeak*. Table 2 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).

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 application. 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.

## 5 FIELD DEPLOYMENT IN RURAL INDIA

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?

### Methods

Out of the 28 participants, 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 *Connect*, a social media voice forum, for 15 minutes. On calling *Connect*, participants could record audio messages and listen to messages recorded by others. We seeded *Connect* with 50 poems, jokes, and songs from *Sangeet Swara* [42], and 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 *Connect* only after completing the tasks. We asked participants questions on how integration of *ReCall* and *Connect* affected their usability and user experience on *Connect*.

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 [21]. The field deployment and usability study was approved by our institution's IRB.

### 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. The *ReCall* 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 reward amount for each audio task was assigned as ₹0.2 per second. If a *ReCall* user transcribed a segment with 80% or above accuracy, the user received the maximum reward amount for the task. If a user transcribed a segment with an accuracy between 50% and 80%, the user received the proportionate amount of the maximum reward. A user received no reward for transcribing a segment with an accuracy below 50%. To compute transcription accuracy, we compared the ASR-generated transcripts with pre-computed ground truth. We avoided using the best estimation transcripts to determine task accuracy since we could not predict how soon and with what accuracy these transcripts will be generated. 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.

### Deployment 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 *ReCall* 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 shows that users enthusiastically used *ReCall* 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 1, 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 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 3: Comparison of *ReCall*'s use by low-income rural residents and *Respeak*'s use by low-income metropolitan residents.**

| | 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 |

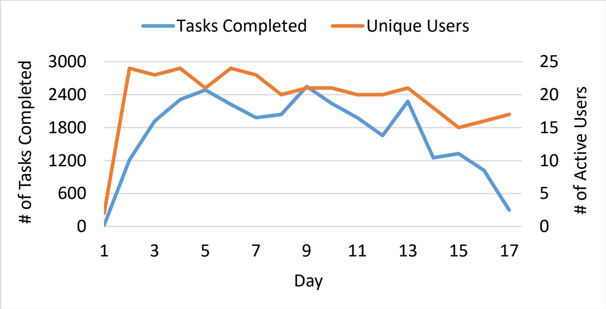

**Figure 4: The number of tasks completed and active *ReCall* users for the deployment duration.**

Table 3 compares the use of *ReCall* by low-income rural residents to the use of *Respeak* by low-income metropolitan residents [42]. 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. In the following sections, we 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 *ReCall* 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 *ReCall* 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). 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 $K$.

We also found value in asking *ReCall* users to re-speak audio segments instead of directly submitting raw audio segments to the ASR engine. The average accuracy of transcripts obtained by submitting raw segments directly to the Google Cloud Speech API [5] was 53%, compared to 73.6% accuracy when users re-spoke these segments into the ASR engine. Two reasons contributed to this significant difference. First, *ReCall* users re-spoke the segments in a quiet environment. As a result, *ReCall* submitted audio files with lower background noise to the ASR engine. Second, *ReCall* users were able to understand even those dialects and diction in raw segments that were difficult for the ASR engine to recognize.

*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 *ReCall* 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.

**Table 4: WERs obtained after aligning transcripts generated by *K* users for each content type.**

| 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** |

## 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 [7]. 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 [41]. 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 [8, 28, 30]. Following suit, all MNOs now offer STVs that provide more affordable or even free voice calls in India [4, 12, 13, 18]. As a result, the average cost of voice calls has reduced from ₹0.49 per minute to ₹0.16 per minute since March 2016 [20].

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* application. 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 [20]) 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 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 5 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 *Connect* revealed promising results. All users completed at least two tasks on *ReCall* to use *Connect* after their free credits expired. While a few participants complained about the context switch between *Connect* and *ReCall*, the majority (N=7) were comfortable in switching between the two services to earn free airtime for using *Connect*. 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 *Connect* 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.

## 6 DISCUSSION AND CONCLUSION

In this paper, we examined if profits generated from crowd work by rural residents can be used to financially sustain voice forums. We employed assets-based approach [34] 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 experimental evaluations, usability studies, and a field deployment to rigorously examine the prospects of crowd work by rural residents to subsidize participation costs of voice forums. Our findings revealed three key results with respect to *ReCall*'s feasibility, usability, and acceptability. First, we found that low-income rural residents enthusiastically transcribed Hindi audio content vocally with a satisfactory accuracy and at an optimal cost. Second, low-income rural residents supplemented their earnings at a rate comparable to 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 that are designed to include low-resource communities in the information ecology.

We opted to deploy *ReCall* for two weeks in rural India rather than for a longer duration because of two reasons. First, we wanted to minimize the impact of participation in our research on users' other responsibilities (e.g., students' coursework, farmers' harvesting activities). Second, we believed that users will regularly use *ReCall* for two weeks only if they find it valuable or engaging. Most *ReCall* users transcribed audio segments regularly and uniformly during the deployment duration, indicating that they found *ReCall* engaging and valuable. Our immediate next step is to conduct a large-scale and long-term deployment of *ReCall* by integrating it in popular voice forums in India to examine its potential to financially sustain these services.

**Table 5: *ReCall*'s cost of transcription (in USD per minute) for different values of $K$ and voice call rates ($call_{cost}$ in ₹ per minute).**

| $K$ | $cost_{rewards}$ (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 |

*ReCall* can be used to provide additional earning opportunities to people, to subsidize their cost of calls to voice forums they use, or both. In its current form, *ReCall* disburse a portion of its profits as earnings to users and another to provide free airtime credits to them. As a result, users receive eight minutes of free airtime for each minute of crowd work they do while earning money at an hourly wage rate of ₹36. If *ReCall* is used only to supplement income of rural residents, all profits can be disbursed to users as earnings at an hourly wage rate of ₹111. Similarly, if *ReCall* is deployed only to subsidize participation costs of voice forums, all profits can be disbursed as free airtime credits to users, enabling them to receive nearly 12 minutes of free airtime for each minute of crowd work they do. Since we compared *ReCall*'s use by rural residents to *Respeak*'s use by metropolitan residents, we used the reward structure that put *ReCall* users on equal footing with *Respeak* users. For future deployments, we encourage tweaking the reward structure to ensure that the amount *ReCall* users earn from an hour of crowd work is much more than the average hourly wage rate in India.

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.

Although our work is a promising first step to demonstrate the feasibility, usability, and acceptability of a crowdsourcing marketplace designed for people in low-resource environments, much more is needed to examine whether *ReCall* 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 [32]? 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? In future work, we plan to investigate these questions as well as examine how *ReCall* can fulfill the criteria suggested by the Fairwork Foundation to create fair digital work opportunities [25].

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.

The transcripts generated by *ReCall* users had some systematic errors as well; some words were unfamiliar to most users; parts of some segments were equally unclear to everyone; speech recognition could not recognize the pronunciation of some words for most users. Future work could use another layer of crowdsourcing where experienced users could use a smartphone app to listen to high quality version of these segments and correct remaining errors by typing. Although introducing this layer could increase the reward costs, it may decrease the airtime costs if unclear segments, decided based on the differences in transcripts generated by a predefined number of users, are sent to the smartphone app users rather than more *ReCall* users. Future work could examine the effects of introducing this layer on the transcription cost and accuracy.

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 *Connect*. 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 [17]) and structural limitations similar to India.

## 7 ACKNOWLEDGMENTS

We are grateful to NYST for helping us recruit rural residents. We thank Rashmi Kanthi, Rajendra Sahu, Shiji Srivastava, Ranjana Tiwari, Jennifer Webster, and the anonymous reviewers for their support and feedback.

## REFERENCES

[1] 2016. Amid fund crunch, CGNet Swara eyes shift to Bluetooth radio tech. https://www.livemint.com/Politics/UcrYsrB8fIAGTDiIoC452N/Amid-fund-crunch-CGNet-Swara-eyes-shift-to-Bluetooth-radio.html.

[2] 2016. Gramvaani | Community-Powered-Technology. http://www.gramvaani.org/.

[3] 2017. Amazon Mechanical Turk. https://www.mturk.com/mturk/welcome.

[4] 2017. BSNL's Rs 8 and Rs 19 plans offer voice calls at 15 paisa per minute. http://indianexpress.com/article/technology/tech-news-technology/bsnls-rs-8-and-rs-19-plans-offer-voice-calls-at-15-paisa-per-minute-4832774/.

[5] 2017. Cloud Speech API: Speech to text conversion powered by machine learning. https://cloud.google.com/speech/.

[6] 2017. CrowdFlower: AI for your business. https://www.crowdflower.com/.

[7] 2017. India Average Daily Wage Rate Forecast 2016-2020. http://www.tradingeconomics.com/india/wages/forecast.

[8] 2017. India's Mobile-Phone Price War Seen Spurring Consolidation. *Bloomberg.com* (Jan. 2017). https://www.bloomberg.com/news/articles/2017-01-26/india-s-mobile-phone-price-war-seen-dialing-up-consolidation.

[9] 2017. Jana. https://www.jana.com/.

[10] 2017. Samasource. http://www.samasource.org/.

[11] 2018. 3-2-1 – On-Demand Messaging for Development. http://hni.org/what-we-do/3-2-1-service/.

[12] 2018. BSNL introduces Rs 19 prepaid plan with affordable voice calling rate for 54 days. https://indianexpress.com/article/technology/tech-news-technology/bsnl-introduces-new-rs-19-prepaid-plans-brings-down-voice-calling-rates-5265617/.

[13] 2018. BSNL's New Rs. 319 Prepaid Plan Offers Unlimited Voice Calls for 90 Days. https://gadgets.ndtv.com/mobiles/news/jio-effect-bsnl-rs-319-99-unlimited-voice-calls-prepaid-plans-caller-tune-validity-1845861.

[14] 2018. CGNet Swara. http://cgnetswara.org/.

[15] 2018. Kan Khajura Tesan. http://www.kankhajuratesan.com/.

[16] 2018. Mobile Kunji | Shaping Demand and Practices. https://www.rethink1000days.org/programme-outputs/mobile-kunji/.

[17] 2018. Uniform call rate from today. https://www.thedailystar.net/news/business/telecom/btrc-minimum-call-rate-tk-045-be-effective-midnight-early-hours-tuesday-bangladesh-1620196.

[18] 2018. Vodafone's new Rs 99 prepaid recharge offer comes with unlimited calling. 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/.

[19] Sheetal K. Agarwal, Arun Kumar, Amit Anil Nanavati, and Nitendra Rajput. 2009. 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)*. IEEE Press, Piscataway, NJ, USA, 56–65. http://dl.acm.org/citation.cfm?id=1812530.1812537

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

[21] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. http://dx.doi.org/10.1191/1478088706qp063oa

[22] Dipanjan Chakraborty, Indrani Medhi, Edward Cutrell, and William Thies. 2013. 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)*. ACM, New York, NY, USA, 7:1–7:9. https://doi.org/10.1145/2442882.2442891

[23] Sebastien Cuendet, Indrani Medhi, Kalika Bali, and Edward Cutrell. 2013. 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)*. ACM, New York, NY, USA, 2833–2842. https://doi.org/10.1145/2470654.2481392

[24] Leslie L. Dodson, S. Revi Sterling, and John K. Bennett. 2013. 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)*. ACM, New York, NY, USA, 79–88. https://doi.org/10.1145/2516604.2516626

[25] Mark Graham and Jamie Woodcock. 2018. Towards a Fairer Platform Economy: Introducing the Fairwork Foundation. *Alternate Routes: A Journal of Critical Social Research* 29, 0 (2018). http://www.alternateroutes.ca/index.php/ar/article/view/22455

[26] Mohamed Gulaid and Aditya Vashistha. 2013. 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)*. ACM, New York, NY, USA, 41–44. https://doi.org/10.1145/2517899.2517947

[27] Aakar Gupta, William Thies, Edward Cutrell, and Ravin Balakrishnan. 2012. mClerk: Enabling Mobile Crowdsourcing in Developing Regions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1843–1852. https://doi.org/10.1145/2207676.2208320

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

[29] 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. 2014. 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)*. ACM, New York, NY, USA, 1595–1604. https://doi.org/10.1145/2556288.2557236

[30] Shilpa Kannan. 2016. Jio: Telecom giant Reliance sparks India price war. *BBC News* (Sept. 2016). http://www.bbc.com/news/business-37273073.

[31] 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. 2016. A Real-Time IVR Platform for Community Radio. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 343–354. https://doi.org/10.1145/2858036.2858585

[32] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 1301–1318. https://doi.org/10.1145/2441776.2441923

[33] Anand Kulkarni, Philipp Gutheim, Prayag Narula, Dave Rolnitzky, Tapan Parikh, and Bjorn Hartmann. 2012. MobileWorks: Designing for Quality in a Managed Crowdsourcing Architecture. *IEEE Internet Computing* 16, 5 (Sept. 2012), 28–35. https://doi.org/10.1109/MIC.2012.72

[34] Alison Mathie and Gord Cunningham. 2003. From Clients to Citizens: Asset-Based Community Development as a Strategy for Community-Driven Development. *Development in Practice* 13, 5 (2003), 474–486. https://www.jstor.org/stable/4029934

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

[36] Prayag Narula, Philipp Gutheim, David Rolnitzky, Anand Kulkarni, and Bjoern Hartmann. 2011. 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)*. AAAI Press, 121–123. http://dl.acm.org/citation.cfm?id=2908698.2908723

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

[38] Agha Ali Raza, Mansoor Pervaiz, Christina Milo, Samia Razaq, Guy Alster, Jahanzeb Sherwani, Umar Saif, and Roni Rosenfeld. 2012. Viral Entertainment As a Vehicle for Disseminating Speech-based Services to Low-literate Users. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development (ICTD '12)*. ACM, New York, NY, USA, 350–359. https://doi.org/10.1145/2160673.2160715

[39] Agha Ali Raza, Bilal Saleem, Shan Randhawa, Zain Tariq, Awais Athar, Umar Saif, and Roni Rosenfeld. 2018. 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)*. ACM, New York, NY, USA, 643:1–643:12. https://doi.org/10.1145/3173574.3174217

[40] Agha Ali Raza, Farhan Ul Haq, Zain Tariq, Mansoor Pervaiz, Samia Razaq, Umar Saif, and Roni Rosenfeld. 2013. 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)*. ACM, New York, NY, USA, 2803–2812. https://doi.org/10.1145/2470654.2481389

[41] Saumya Tewari. 2015. 75 percent of rural India survives on Rs 33 per day. https://www.indiatoday.in/india/story/india-rural-household-650-millions-live-on-rs-33-per-day-282195-2015-07-13.

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

[43] Aditya Vashistha, Edward Cutrell, Nicola Dell, and Richard Anderson. 2015. Social Media Platforms for Low-Income Blind People in India. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '15)*. ACM, New York, NY, USA, 259–272. https://doi.org/10.1145/2700648.2809858

[44] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2017. Respeak: A Voice-based, Crowd-powered Speech Transcription System. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1855–1866. https://doi.org/10.1145/3025453.3025640

[45] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2018. BSpeak: An Accessible Voice-based Crowdsourcing Marketplace for Low-Income Blind People. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 57:1–57:13. https://doi.org/10.1145/3173574.3173631

[46] Jerome White, Mayuri Duggirala, Krishna Kummamuru, and Saurabh Srivastava. 2012. 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)*. ACM, New York, NY, USA, 367–373. https://doi.org/10.1145/2160673.2160717