Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2819–2826
Marseille, 11–16 May 2020

Crowdsourcing Speech Data for Low-Resource Languages from Low-Income Workers

Basil Abraham*, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti[†], Sunayana Sitaram, Vivek Seshadri
Microsoft Research India, *Microsoft India Development Center, [†]IIT Bombay
basil.abraham@microsoft.com, danishdgoel@gmail.com, dsiddarth1@gmail.com, kalikab@microsoft.com,
mchopra@cs.stanford.edu, monojitc@microsoft.com, pratikmjoshi123@gmail.com,
pjyothi@cse.iitb.ac.in, sunayana.sitaram@microsoft.com, visesha@microsoft.com

Abstract

Voice-based technologies are essential to cater to the hundreds of millions of new smartphone users. However, most of the languages spoken by these new users have little to no labelled speech data. Unfortunately, collecting labelled speech data in any language is an expensive and resource-intensive task. Moreover, existing platforms typically collect speech data only from urban speakers familiar with digital technology whose dialects are often very different from low-income users. In this paper, we explore the possibility of collecting labelled speech data directly from low-income workers. In addition to providing diversity to the speech dataset, we believe this approach can also provide valuable supplemental earning opportunities to these communities. To this end, we conducted a study where we collected labelled speech data in the Marathi language from three different user groups: low-income rural users, low-income urban users, and university students. Overall, we collected 109 hours of data from 36 participants. Our results show that the data collected from low-income participants is of comparable quality to the data collected from university students (who are typically employed to do this work) and that crowdsourcing speech data from low-income rural and urban workers is a viable method of gathering speech data.

Keywords: Labelled Speech Data, Low-resource Languages, Rural Communities

1. Introduction

In the next 3 years, roughly 400 million Indians are expected to get access to a smartphone for the first time (BusinessStandard, 2019). These new smartphone users speak one of India's 19,569 different languages (IndiaToday, 2019), most of which are not represented in the mainstream. Moreover, given the complexity of existing hierarchical, text-based user interfaces, voice-based technologies have proven to be the best way to reach rural users (Medhi Thies, 2015). Therefore, there is significant value in creating local-language technologies to serve these users.

Almost all Indian languages can be considered as "low-resource languages", i.e. there is little to no speech data available in these languages. Unfortunately, collecting good-quality labelled speech data in any language is an extremely expensive and resource-intensive operation. This makes speech data collection the biggest bottleneck in creating local-language voice technologies.

Prior work has looked at crowdsourcing as a way to collect speech data at low cost (Hughes et al., 2010). So far, these platforms have only worked with educated urban participants (such as university students) as they can be reached easily and are typically digitally savvy. However, hiring only urban participants leads to a lack of diversity in the collected dataset. It is important to note that even within the same language, the dialects spoken by rural Indians can be fairly distinct from dialects spoken by urban Indians. Thus, such non-diverse datasets are not sufficient to build technologies to effectively serve rural users.

Our goal in this paper is to explore the use of crowdsourcing to collect speech data from economically backward communities who are not well-versed in digital technologies. This approach allows us to get labelled speech data from a diverse set of local language speakers that are traditionally underrepresented in such datasets. In addition to this, it could also be a viable source of supplementary income for our participants. In this paper, we explore whether it is possible to collect labelled speech data from underserved communities (both rural and urban) using smartphones. It is important to note here that as of 2014, 71% of rural Indians and 86% of urban Indians are literate in at least one language. Thus, the vast majority of rural and urban Indians are literate, and are able to read and write their local languages.

For our study, we collected speech data in Marathi. Despite being one of the 22 scheduled languages of India, the amount of linguistic resources available in Marathi are few. The Linguistic Data Consortium for Indian Languages (one of the leading open-source datasets) only has 168 hours of spoken data in Marathi (Shrishrimal et al., 2012).

We collected Marathi speech data from three different user groups: 1) low-income rural Marathi speakers in Amale, a tribal village in rural Maharashtra, 2) low-income urban Marathi speakers in the slums of Kolhapur, a small city in Maharashtra, and 3) university students in Mumbai, who are the typical target population of crowdsourced data collection. Our goal is to study and compare both the quality of data collected as well as the experiences of the workers across the 3 groups.

Because of the unique demands of our target demographic (rural and urban poor Marathi speakers), we could not rely on an existing crowdsourcing solution. Most crowdsourcing platforms assume access to a computer and the Internet. However, our chosen village in rural Maharashtra, Amale, does not have any mobile signal or Internet. None of our participants in Amale and the slums of Kolhapur had access to a computer. Further, most crowdsourcing platforms assume that their users speak/understand English, and these platforms are typically designed for digitally savvy users. None of our participants in Amale and Kolhapur spoke or understood any English, and for most of them, this was their first time using a smartphone. To solve all of these problems, we created our own Android application (See Section 4.2) to perform the data collection activity.

Finally, we evaluate the collected speech data and qualitative feedback from our participants. We conduct quality-check experiments and error analysis of the data collected, and discuss how such a platform might be realized as a source of supplementary income generation for workers in rural India. Our results support the viability of crowdsourcing speech data collection to poor areas in rural and urban India. All collected data is released to the public, free of cost, to motivate future research.

2. Crowdsourcing for Data Collection

There have been a number of prior efforts that have looked at crowdsourcing as a way to collect speech data for low-resource languages. Ambati and Vogel (2010) probed Amazon Mechanical Turk workers to see if they are capable of translating a number of low-resource languages, including Hindi, Telugu, and Urdu, demonstrating that such workers could be found. Novotney and Callison-Burch (2010) showed that transcriptions for training speech recognition systems could be obtained from Mechanical Turk with near baseline recognition performance and at a significantly lower cost (Callison-Burch and Dredze, 2010).

Even though Mechanical Turk has a high representation of workers from India (and thus, could be used for speech data collection for local Indian languages), most Mechanical Turk workers in India are urban, highly educated, speak fluent English and have access to a computer and high-speed Internet (Khanna et al., 2010). Datasets, collected using Mechanical Turk, would not be very diverse. We strongly believe that traditionally underrepresented rural Indians also have much to gain, and much to contribute, as participants in crowdsourcing platforms.

Takamichi and Saruwatari (2018) used web-based recording and crowdsourcing platforms to construct a parallel speech corpus of Japanese dialects. Hughes et al. (2010) suggest creating an Android application to help build transcribed speech corpora quickly and cheaply for many languages. They mainly engage with university students, and mention that university students require less technical support. BSpeak is an accessible crowdsourcing marketplace that enables visually challenged people in an urban setting to earn money by transcribing audio files in English through speech (Vashistha et al., 2018). To the best of our knowledge, this is the first work to crowdsource speech data collection in Indian languages to rural Indians. Prior research has shown the poor state of speech data collection in Marathi (Shrishrimal et al., 2012). Our work aims to collect widely spoken but under-researched Marathi dialects (ranging from tribal Maharashtra to Kolhapur and Mumbai) and is the first step to capture India's diverse 19,569 languages in a speech database.

3. Bringing Crowdsourcing To Underserved Communities

In 2015, the Government of India launched the celebrated Digital India Mission. The mission advocates for speech data collection in all major Indian languages, and allocates funding for corpus construction in these languages. The government hopes to release the collected data to the public for free, to encourage the creation of local startups.[1]

Simultaneously, researchers have created platforms such as Project Karya (Chopra et al., 2019) that provide digital work to underserved communities. Project Karya deployed an Android application that its users used to digitize local language text. Users were shown images of words in their local language(s) and prompted to type them using a keyboard. Project Karya has shown that when digitizing local language text, communities in rural India outperform traditional digitization firms, achieving higher accuracy at a similar cost. Participants respond to the work with high levels of interest and excitement. Users reportedly like the flexibility digital work provides (they could do the work anywhere, anytime) and the prospect of supplementary income. Users also reported enjoying working in groups (neighbors, co-workers etc) and reported increased use of Hindi in messaging applications like Whatsapp.

Prior work (Chopra et al., 2019) has also observed the various benefits of bringing crowdsourcing to rural communities. Given that 75% of rural Indians live on less than INR 33 (USD 0.5) a day, helping build datasets in their local languages could substantially increase daily incomes of our participants. Crowdsourced work in local languages also has the potential to bring marginalized cultural artifacts (such as tribal languages) into the digital realm. Digital work can also boost digital literacy, potentially giving our participants access to other earning opportunities.

Bringing crowdsourcing to rural Indians could create a win-win situation by : 1) Providing digital work and the promise of additional income to local language speakers in underserved communities and 2) Getting valuable data from a diverse set of local language speakers traditionally underrepresented in such datasets.

Because of the unique demands of our target demographic (poor communities in rural and urban India), we couldn't rely on an existing crowdsourcing solution. As mentioned before, most crowdsourcing platforms assume access to a computer and the Internet. They also assume that their users speak English. Both those assumptions are not true in our case. Given the success of Project Karya in tapping the rural population for text-digitization and given the urgency for speech data collection in India, we decided to focus on expanding Project Karya to collect speech database by creating a speech data collection application (See Section 4.2). Like Project Karya, all files are stored offline and retrieved by the research team at the end of the user study.

[1] Government of India Press Release: `https://meity.gov.in/writereaddata/files/linguistic_resources_sharing_om.pdf`

4. Study Design

4.1. Language

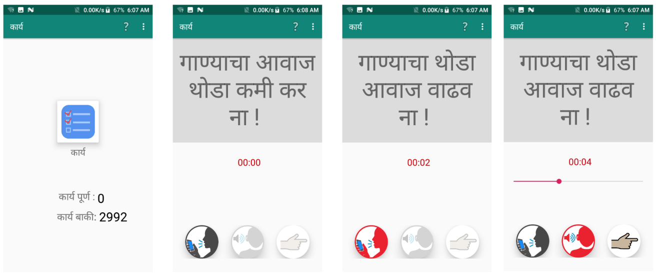
In this study, we focus on Marathi, an Indo-Aryan language spoken primarily in the western Indian state of Maharashtra, along with a strong presence in other neighbouring states of Goa, Madhya Pradesh, Karnataka and Gujarat. Marathi is spoken by over 83 million speakers and has the 3rd largest number of native speakers in India, after Hindi and Bengali. It is also the world's 10th most widely spoken language (Wikipedia, 2020). Despite such a large speech community, little work has been done for the Marathi Language especially with respect to language technology. Prior research has also shown the need to obtain continuous speech datasets recorded in noisy environments, in Marathi. (Shrishrimal et al., 2012)

4.2. Application

As we were working with participants with limited digital literacy, we had to design our application such that novice and low-literate users would feel comfortable using the application, with minimal training and external assistance. Earlier work has shown that static, hand drawn representations are better understood than photographs or icons by low-literate users (Medhi Thies, 2015). Although prior work focuses on low-literate users, we believe their results should extend to users who are new to digital devices. Therefore the main functionalities of our application are presented as hand drawn icons. These buttons clearly represent the action they are intended for. The recording button shows a user speaking to the phone. The listening button shows a user listening to the recorded audio, and the next button shows a hand pointing to the right.

Figure 1 shows four screenshots of our Android application. The first screenshot (from the left) shows the application home screen. This screen shows the number of tasks completed and the number of tasks remaining. When the user clicks on the blue checklist button, it takes them to the recording activity (second screenshot from the left). The recording activity displays a Marathi sentence. At the bottom of the screen, there are 3 buttons that record, listen and go to the next sentence, respectively. Figure 2 shows the 3 buttons. The user can start recording the audio by clicking on the record button, at which point, it turns red (third screenshot from the left). They can stop recording by clicking the record button again. The listen button is automatically activated, and plays back the recording (fourth screenshot from the left). If the user is happy with their recording, they can click on the next button. If they think they have made a mistake, they can hit the record button again and re-record the sentence.

For every sentence, we grey out the listen and next buttons at the beginning. This was done to make it easier for users to know which button to click. After our initial studies, users indicated a preference for a back button to go to the previous sentence, and re-record it, if required. As users became more adept at using the application, it was noted that sometimes, they would press the next button in haste, without listening to their recording properly. A back button was added for our future studies.



कार्य पूर्ण : 0
कार्य बाकी: 2992

गाण्याचा आवाज थोडा कमी कर ना !

गाण्याचा थोडा आवाज वाढव ना !

गाण्याचा थोडा आवाज वाढव ना !

Figure 1: Screenshots of our Android application

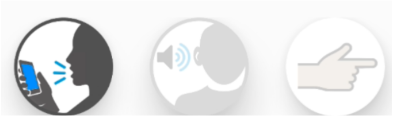


Figure 2: The 3 key buttons

Finally, at any point in the application, if participants need further information, they can click on the "?" button to bring up a video that describes the application. We describe the video in detail in the Section 4.3.1. For the three user studies, the application worked without Internet and saved all recordings on the phone's local storage. This was done to enable the data collection activity even in those areas where cellular or Internet connectivity was intermittent or even non-existent.

4.3. User Study in Amale

We conducted a two-week user study in Amale, a small tribal village in the Wada district of Western India. The local language spoken in Wada and surrounding districts is Marathi. We worked with a local nonprofit, Rural Caravan, based in Wada. In the last 2 years, the nonprofit has operated in several 'adivasi' (tribal) communities in the area.

12 participants (3 men, 9 women) volunteered to be a part of the user study. Every participant had completed at least a 5th standard education in their local language, and were fluent readers in Marathi. The group had an average of a 10th standard education, and the most educated participants were college graduates (Bachelors). The group had people from the ages of 18-63 and every participant belonged to a Below the Poverty Line family (households making less than INR 27000, around USD 385, per annum). Of the 12 participants, no one had access to a smartphone. For the duration of the study, we provided inexpensive Android smartphones that cost less than USD 50 to each of our participants. We also provided earphones (with a microphone) to our participants for the duration of the study, to ensure better audio quality. The tribal village of Amale had no cellular data connectivity. One of our participants informed us that once a day, every week, people who have feature phones climb a nearby mountain to get access to mobile signal.

We created a corpus of 3,000 sentences for the study. The sentences were collected from Pratham's Marathi textbooks[2] and included short stories with lessons on morality and virtues like compassion, secularism etc targeted at chil-

[2] https://storyweaver.org.in/stories?language=Marathi

dren from K1-4. As we discuss later, these stories were well received by participants.

On our first visit to the village for this study, we distributed the Android phones and earphones to our participants. The primary researcher introduced the work, and explained that a payment of INR 1 per sentence would be made on the successful completion of the work. The participants were asked to watch an introductory video, and then perform the tasks.

4.3.1. Training Participants in Amale

A short introductory video is played, once the application is installed. The video (featuring a local Marathi speaker) encourages the user to go to a quiet place before they start recording the sentences. The video demonstrates how to record Marathi sentences, using two examples. The users are asked to listen to their recordings properly and ensure that the given sentence is audibly spoken and that there is limited background noise.

The researchers spent an hour in the village, making sure that the users understand the application. Because of the village's communal lifestyle, participants started helping each other. The younger participants (who understood how to operate the smartphone quicker) started helping the older participants. It is important to note that, for most of our older participants, this was their first time seeing a smartphone. While our younger participants also didn't have access to a smartphone, they had seen them before (in college, in towns, in markets etc). After an hour of observation, the researchers left the village. We returned after two weeks, to collect the recorded speech data.

4.4. User Study in Kolhapur

Next, we worked with 12 participants in the Shenda Park colony, an unrecognized slum dwelling in the small city of Kolhapur in Maharashtra. Most families in Shenda Park have one or more members affected by leprosy. As people suffering from leprosy are ostracized in Indian society, most of our participants were unemployed or begging in the local temples and mosques. We worked with a local nonprofit, Shelter Associates, based in Shenda Park. The nonprofit has worked in the colony for more than 5 years now, and has built excellent sanitation solutions (toilets, access to clean drinking water) for every home.

12 participants (4 men, 8 women) volunteered to be a part of the user study. Every participant had completed at least a 2nd standard education in their local language. The group had an average of a 8th standard education. The group had people from the ages of 23-83 (average age was 41 years) and every participant belonged to a Below the Poverty Line family. Of the 12 participants, no one had access to a smartphone, and only 2 people had used a smartphone before. Like our previous study, we provided inexpensive Android smartphones and earphones to all our participants for the duration of the study.

4.4.1. Training Participants in Kolhapur

Once again, participants were asked to watch the introductory video and perform the given tasks. The older participants (who had leprosy), initially struggled with the smartphone interface and using the application. But, in all cases, they were helped by their younger children and grandchildren. At the end of the two-week study, we returned to the slum and collected the recorded speech data.

4.5. User Study in Mumbai

Finally, we conducted a two-week study with college students in Mumbai, the second most populous metropolis in India. Traditionally, most speech data collection work in India (for research purposes) is done by college students and we wanted to compare both the accuracy of the recorded speech data, and the qualitative experiences of college students with rural poor and urban poor participants.

We selected 12 students from a local government university in Mumbai for our study. The 12 participants (12 women) were engineering students. Most of them were 2nd and 3rd year B.Tech (Bachelor of Technology) students. All of the 12 participants had their own smartphones, and were proficient in using them. Since the students were familiar with the concept of digital work, we didn't need to train them. Participants were asked to watch the introductory video and perform the tasks.

4.6. Payment for Work

All 36 participants had recorded the given 3000 sentences, and they were paid INR 3000 (around USD 45) each for the work. Detailed user interviews were conducted with the participants, and qualitative experiences (while doing the work) were noted.(See Section 5)

We sought to pay participants at a rate that could reasonably be supported by potential future employers and yet higher than the existing minimum wage, which varies from state to state in India. In Maharashtra, the government of India pays INR 201 (USD 2.8) per day or INR 2814 (USD 40) for 2 weeks. It should be mentioned that this wage is for a full day (8 hours) of work.

4.7. User Study Details

We worked with a total of 36 participants, and collected 109.301 hours of spoken data. Of the 36 participants, 29 participants were women. The average age was 31 (ranging from 18-83) and participants had a wide variety of existing occupations and education levels. No personally identifiable information was collected during the study.

5. Qualitative Analysis

We observed participants as they used the application, and at the end of all three user studies, we conducted interviews with all of our participants. The interviews were conducted by members of the research team in a local setting. We attempted to put our participants at ease by conducting the interviews as conversations. Before the study started, we took consent from all our participants. They were informed about the purposes of the research, and that the collected data would be published online.

Overall, participants expressed enthusiasm for completing the work, and were excited about the opportunity to both participate in the digitization of their native language(s) and to work with their friends and neighbors in this process.

Our interviews coalesced around three main themes: elements of linguistic pride, the centrality of the communal aspect of the work, and the importance of storytelling and commensurate lack of previously available reading material. For all themes, we noticed major differences between rural and urban poor, and college students.

5.1. Elements of Linguistic Pride

"*Earlier, we used to give you money so you can teach us English. Now, you are giving us money so you can learn Marathi. Times have really changed.*" - P5, unemployed man, Kohlapur

In working with both the rural and urban poor in Maharashtra, elements of linguistic pride came through strongly in the interviews. Participants repeatedly mentioned the pride they felt, not only in their native tongue and ability to share their knowledge, but also in the fact that English-speaking outsiders also attributed importance to their native language.

With the exception of one student, none of the college students expressed this linguistic pride, with many describing the work as 'boring', something that we did not hear even once in the rural or urban poor communities.

In fact, the only college student who mentioned linguistic pride, and who did not say that the work was boring, was originally from a rural community in Karjat, Maharashtra.

5.2. Communal Working Styles

"*At first I was a bit scared...but then they all helped me through it. And I had fun! It went so fast.*" - P6, older woman, Kohlapur

In rural and urban poor settings, we observed that completing the tasks became a communal activity, and this community aspect was emphasized by participants in our interviews as a crucial benefit of the platform. Friends and family members would complete tasks together, and we observed younger participants spend time helping older participants through occasional difficulties with the interface or with certain words or phrases. This was a reflection of daily life in these communities, where both work and play are often deeply communal activities. Older participants, particularly, were able to build confidence through these communal aspects and become more adept at and more excited about completing tasks, as in the quote above. The communal aspects of the work allowed participants to more seamlessly integrate the platform into their daily lives.

This was certainly not expected. Given that the work has to be done in a quiet setting, we expected participants would do the work in isolation. While participants did work in isolation sometimes (at their homes, in their fields, by the river), they often asked each other for help. Participants also reported sitting next to each other, and taking turns speaking the same sentence.

University students, however, did not report working together or asking each other for help (even though they all went to the same university, and lived together). This could be due to how comfortable they already were with their smartphones.

5.3. Reading Material and Storytelling

"*I loved reading Marathi as a child. But, as there are no books or newspapers in the village, I am no longer able to read as often. Sometimes, traders from the city bring "samosas" wrapped up in local newspapers. They throw the newspapers, and I would pick them and read the stories in the newspaper.*" - P11, older woman, Amale

The lack of reading material in the villages was keenly felt by participants. As Amale is a relatively remote village, participants did not have access to newspapers, television, books, or other types of content, to the point where they found themselves reading stories from snack wrapping paper, as in the quote above. Rural participants especially described the speech data tasks as bringing welcome content to the village. They particularly emphasized the element of storytelling in the tasks, which were constructed as individual narratives, as a reprieve from having very little interesting content to consume otherwise in their daily lives. Participants in the urban slums of Kolhapur reported slightly less excitement about the stories. College students reported significantly less excitement about the stories.

6. Quantitative Analysis

We collected a total of 109 hours of Marathi speech data from across the 36 participants. Table 1 shows the number of sentences and hours of data collected from each of the three user groups.

Category	Number of Datapoints	Total Hours
College	29561	30.625
Rural Poor	26593	34.764
Urban Poor	36317	43.912
Total	92471	109.301

Table 1: Data Collection Statistics

6.1. Evaluation of Collected Data

One of the main challenges in speech data collection, especially when crowdsourcing, is evaluating data quality. With no direct method of measuring it, bad data quality can significantly hamper the performance of tools and data-driven models which are trained on it. Commercial data collection, typically employed by industry, has a rigorous set of quality standards imposed upon them in order to ensure that the most can be leveraged from the dataset. In order to make the accumulated data useful to the consumers of data, it is imperative to have indicators that our data matches up to certain quality levels. Manual inspection, although the most reliable, is often not feasible as the size of the data collection scales up. An automatic or semi-automatic method needs to be employed. In our specific scenario, we had crowdsourced data from low-income workers who were not used to doing digitized work, and did not have exposure to advanced technologies to be comfortable in using online annotation platforms. This made it all the more important to meticulously set and record data quality statistics. Typically, this would not be an issue if either an Automatic Speech Recognition (ASR) system in that language, or manual gold standard annotations for the speech data quality were available. However, our scenario is common with many other low resource scenarios where we do not have such existing resources and tools. Therefore, we have devised an alternative method to determine data quality.

6.1.1. Experiments

For our data assessment, we carried out a set of experiments which serve as an indicator of data quality as well as a measure of potentially how much the data can improve industrial standard speech recognition engines. For this, we employed an industry-level subtitle alignment pipeline, which aligns speech clips to their transcripts. They are often used in aligning captions in videos (Shin et al., 2016) (Park et al., 2010) . We used the Kaldi system to build our alignment system (Kaldi, 2020). The pipeline does the following:

1. Attempts to convert the speech clip to text using an ASR. In our case, the ASR consisted of a pre-trained Hindi acoustic model, and a restricted language model which was trained on the given transcripts. This made up for a Marathi ASR, since there wasn't sufficient data to train a high-performing Marathi acoustic model from scratch. Marathi and Hindi are related languages with close phonemic inventories, thus the use of the Hindi ASR was considered.

2. An aligning module attempts to align the decoded text with the original transcript, allowing for a certain error margin (which is bounded by a given threshold). If the module can align the text within the error margin, it classifies the datapoint as of good quality, else it gets classified as misaligned/bad quality. We posit that this serves as a metric for data quality, because if the pipeline, which has been pre-trained heavily on high-quality data, can align the speech-transcript pairs, the data can be used effectively for training other speech models.

Using the pipeline above, we carry out the following steps for quality assessment:

1. Using the subtitle alignment pipeline and a hold-out set of the collected data from our user studies, check the percentage of datapoints that have been aligned in the hold-out set.

2. Train (fine-tune) the acoustic model of the alignment pipeline with the remaining collected speech-transcript data, and repeat step 1 after. Look at the improvement of the pipeline.

We fine-tune the ASR in the pipeline because although Marathi and Hindi have similar sound systems, they have very distinctive prosodic features that differentiate them. Fine-tuning can result in dramatic performance improvements if the fine-tuning data is of good quality. Thus, a strong improvement in the pipeline results will reinforce the good quality of the data.

6.1.2. Experimental Results

As seen in Table 2, there is approximately an *11% increase* in hold-out set performance after fine-tuning. Looking at all categories of data, we can see that there are significant jumps on all fronts. This shows that the data collected positively affects the training of an industrial-level ASR instead of hampering it. An interesting point to note is that the largest increase comes from the data collected from rural workers (17.8%). We can gather that this type of data is under-represented in the previous training of the ASR, which again stresses the diversity of data that our platform brings about. In order to accommodate different communities, with varying speaking styles, and allow them to access these speech technologies, there needs to be training data diversity to maintain high performance across different users.

6.1.3. Error Analysis

After the initial experiment was carried out (without the fine-tuning), we analyzed the mismatched audio-transcripts, for around half an hour of audio. We classified the errors into categories, as shown in 3. We realized that approximately 31% (Categories 3 and 8) of the mismatched audios are a result of the inability of tuned Hindi ASRs to recognize Marathi speech. Fine-tuning and the creation of a Marathi ASR can substantially cut down the mismatches, again emphasizing the importance of collecting diverse data at a larger scale. After fine-tuning on the collected data, we noticed that most of these errors reduced substantially. Also, errors due to categories 1,2,4,5 and 7 are all easily addressable by better controlling the collection environment and introducing these aspects into the training of participants.

Looking at the individual types of data collected, it can be noted that college audio transcripts were slightly better aligned than rural and urban poor audio transcripts initially (2-3% higher). However, after fine tuning, rural and urban poor audio transcripts align better than audio transcripts of college students (5-10% higher). This shows that the data collected from rural and urban poor participants was of comparable quality to the data used to pre-train the industrial level ASR.

Further, we did another manual analysis of around 30 errors from each type of data (college,rural,urban) which occurred after the fine-tuned experiment. As seen in Table 4, there are more errors which occurred due to stuttering, pausing and murmuring in the rural data collection as compared to college data or urban data. We believe that these errors are easily fixable, with introduction of more advanced training exercises. We also expect these errors to decrease as familiarity with the task and the application increases.

As a result of this analysis, as future work, we recommend that during the training of crowdsourcers, participants are made aware of the errors that render the data faulty. This can prevent a large percentage of human errors and environmental errors from taking place. Further, we could also think of ways to automatically flag such errors during data collection to prevent them from being added to the standardized collection, thus saving a large amount of post-processing effort.

Stage	**% Datapoints aligned**			
	College	Rural Poor	Urban Poor	Total
Before Fine-Tuning	67.6%	63.9%	65.1%	**65.5%**
After Fine-Tuning	71.2%	81.7%	76.9%	**76.4%**

Table 2: Results of Automated Alignment

7. Discussion And Future Work

Our work proves the viability of a crowdsourcing platform that employs poor rural and urban Indians to collect speech data in Indian languages. Our participants came from disadvantaged backgrounds, and most of them had never used a smartphone before. Yet, the collected speech dataset was similar in quality to the one collected from college students (who are traditionally employed to do this work). Participants in both rural and urban poor communities reported high levels of enthusiasm in finishing this work. Elements of linguistic pride, communal working styles and excitement to read stories in their language were uniquely noticed in the rural and urban poor communities. Even in the absence of a proper Marathi ASR, we describe a way to evaluate the quality of the collected data. We also analysed the errors in the collected data, and note ways future work could preempt them.

Our results prove our initial hypothesis, that employing rural and urban poor participants to collect local-language speech data creates a win-win situation : 1) it provides them with supplementary income and 2) provides us with a diverse data set we can use to build new local language tools and systems.

We have already started extending our work. We are currently working on a speech data collection for Hindi with 110 villagers from Soda, a village in rural Rajasthan. Soda ranks among India's most resource-constrained villages, and our initial results look quite promising. In the future, we would like to build more and more diverse data sets in Indian languages with our rural and urban poor communities, and release these data sets to the public.

8. Acknowledgements

We would like to thank the following people for their help: Anuja Dumada and Pranita Harpale, our research associates, who helped us with the study in Amale. Rohit Pillai and Nikesh Ingle from Rural Caravan, our project partner in Amale. We would also like to thank Lata Shrikhande and Pratima Joshi, from Shelter Associates, for helping us with our study in Kolhapur. Finally, we would like to thank Sanjay Pawar from SNDT College, who helped us with our study in Mumbai.

Category	**% of observed**	**Category Details**
1) Correct clips	24%	The speech had no issues and matched the gold transcription.
2) Incorrect speaking, misunderstood word	19%	The reader mispronounced the word/phrase.
3) Stuttering, pausing, murmuring	15%	The reader stuttered/murmured through the word, or took pauses in between saying a word.
4) Low volume or inaudible speech	11%	-
5) Paraphrased the reading text while speaking	11%	The reader paraphrased the given transcript without reading it word-for-word. Certain words were unknowingly omitted,inserted, or substituted for similar sounding/meaning words.
6) Background noise	8%	-
7) Speaking in a local/slang manner	7%	The reader uttered a word/phrase in a slang manner, which sounds different than how it would be uttered in a ”Hindi”-way of speaking. This may have confused the Hindi ASR.
8) Repeating a word or sentence	4%	The reader repeated a phrase half-way through saying it, perhaps due to the conception that they mispronounced or wrongly uttered the transcript.

Table 3: Error analysis on random sample of half an hour of mismatched audio samples, carried out after initial experiments on the collected speech data.

Category	% Category Total		
	College	Rural	Urban
1) Stuttering, pausing, murmuring	2.02%	4.32%	2.3%
2) Repeating a word or sentence	0%	0%	0%
3) Speaking in a local/slang manner	2.02%	0%	2.3%
4) Incorrect speaking, misunderstood word	2.9%	5.08%	1.39%
5) Low volume or inaudible speech	3.74%	4.32%	7.62%
6) Paraphrased the reading text while speaking	0%	0.56%	0%
7) Background noise	7.78%	1.13%	2.31%
8) Correct clips	10.66%	3.2%	6.93%

Table 4: Split of errors as a percentage of total category datapoints which occurred after fine-tuning of the pipeline.

9. Bibliographical References

Ambati, V. and Vogel, S. (2010). Can crowds build parallel corpora for machine translation systems? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 62–65, Stroudsburg, PA, USA. Association for Computational Linguistics.

BusinessStandard. (2019). Number of smartphone users in india likely to double to 859 million by 2022 — business standard news. `https://www.business-standard.com/article/news-cm/number-of-smartphone-users-in-india-likely-to-double-to-859-million-by-2022-119051000458_1.html`. (Accessed on 12/01/2019).

Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 1–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chopra, M., Medhi Thies, I., Pal, J., Scott, C., Thies, W., and Seshadri, V. (2019). 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. ACM.

Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P., Lebeau, M., Research, G., and Usa. (2010). Building transcribed speech corpora quickly and cheaply for many languages. 01.

IndiaToday. (2019). 'more than 19,500 mother tongues spoken in india'. `shorturl.at/bfo04`. (Accessed on 12/01/2019).

Kaldi. (2020). Kaldi. `https://kaldi-asr.org/doc/about.html`. (Accessed on 13/03/2020).

Khanna, S., Ratan, A., Davis, J., and Thies, W. (2010). 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*, page 12. ACM.

Medhi Thies, I. (2015). User interface design for low-literate and novice users: Past, present and future. *Found. Trends Hum.-Comput. Interact.*, 8(1):1–72, March.

Novotney, S. and Callison-Burch, C. (2010). Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. pages 207–215, 07.

Park, S., Kim, H., Kim, H., and Jo, G. (2010). Exploiting script-subtitles alignment to scene boundary dectection in movie. In *2010 IEEE International Symposium on Multimedia*, pages 49–56, Dec.

Shin, A., Ohnishi, K., and Harada, T. (2016). Beyond caption to narrative: Video captioning with multiple sentences. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3364–3368, Sep.

Shrishrimal, P., Deshmukh, R., and Waghmare, D. V. (2012). Indian language speech database: A review. *International Journal of Computer Applications*, 47:17–21, 06.

Takamichi, S. and Saruwatari, H. (2018). CPJD corpus: Crowdsourced parallel speech corpus of Japanese dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Vashistha, A., Sethi, P., and Anderson, R. (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, pages 57:1–57:13, New York, NY, USA. ACM.

Wikipedia. (2020). Marathi wikipedia page. `http://tiny.cc/gga9kz`. (Accessed on 13/03/2020).