Respeak: A Voice-based, Crowd-powered Speech Transcription System

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

Pooja Sethi
University of Washington
psethi17@cs.washington.edu

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

ABSTRACT

Speech transcription is an expensive service with high turnaround time for audio files containing languages spoken in developing countries and regional accents of well-represented languages. We present *Respeak* — a voice-based, crowd-powered system that capitalizes on the strengths of crowdsourcing and automatic speech recognition (instead of typing) to transcribe such audio files. We created *Respeak* and optimized its design through a series of cognitive experiments. We deployed it with 25 university students in India who completed 5464 micro-transcription tasks, transcribing 55 minutes of widely-varied audio content, and collectively earning USD 46 as mobile airtime. The *Respeak* engine aligned the transcript generated by five randomly selected users to transcribe Hindi and Indian English audio files with a word error rate (WER) of 8.6% and 15.2%, respectively. The cost of speech transcription was USD 0.83 per minute with a turnaround time of 39.8 hours, substantially less than industry standards. Using a mixed-methods analysis of cognitive experiments, system performance and qualitative interviews, we evaluate *Respeak*'s design, user experience, strengths, and weaknesses. Our findings suggest that *Respeak* improves the quality of speech transcription while enhancing the earning potential of low-income populations in resource-constrained settings.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

HCI4D, Crowdsourcing, Speech, Transcription, India

INTRODUCTION

Speech transcription — including general, medical and legal transcription — fuels a massive industry; medical transcription alone is expected to reach USD 60 billion globally by 2019 [10]. Transcription of recorded audio is demanded for a wide variety of content, including public speeches, movies, songs, television programs, advertisements, news, interviews, recorded lectures, online videos, and telephone calls. Manual transcription via typing is both slow and expensive. The advent of crowdsourcing has resulted in the rise of crowd-powered transcription organizations, such as CastingWords [3] and SpeechPad [15]; these businesses typically charge USD 1 – 6 per minute of recording and require a one-week turnaround time for the cheapest pricing alternative. Existing automatic speech recognition (ASR) based transcription solutions, like Nuance Dragon, work satisfactorily but only for individual use of well-represented languages, such as English and Spanish, and require respeaking to transcribe audio files. However, no crowdsourcing platform or ASR system currently supports transcription of audio files in languages spoken in developing countries and localized accents of well-represented languages. Though online services like Quick Transcription [11] and Scripts Complete [14] support such transcription using a fleet of transcribers, their transcription cost starts at USD 5 per minute.

The emergence of crowd-powered speech transcription platforms has demonstrated the potential to provide additional earning opportunities to low-income people. However, several inclusion criteria — a minimum typing speed of 40 words per minute (WPM) [15], an active PayPal account connected to a banking institution [3, 12, 15, 16, 17], and access to an Internet-connected computer [3, 12, 15, 16, 17] — makes it difficult for many in developing regions to receive benefits of these platforms. Other crowdsourcing platforms, such as Samasource [13] and MobileWorks [37], designed especially for people in developing countries require workers to have acceptable typing skills: an acquired skill for English language and onerous for local languages like Hindi.

Our primary contribution is the design, deployment and evaluation of *Respeak*: a voice-based, crowd-powered speech transcription system that combines the benefits of crowdsourcing with ASR to transcribe audio files containing local languages like Hindi and localized accents of well-represented languages like English. *Respeak* lets people use their speaking rather than typing skills to transcribe audio files with a low WER, turnaround time, and transcription cost. The *Respeak* engine works by segmenting an audio file into utterances that are each three to five seconds long. Each audio segment is sent to multiple *Respeak* smartphone application users who then listen to the segment and re-speak what they heard into the application in a quiet environment. The smartphone application uses Google's Android speech recognition API to generate an instantaneous transcript for the segment, albeit with some errors. The user then submits this transcript to the *Respeak*

CHI 2017, May 6-11, 2017, Denver, CO, USA.
Copyright © 2017 ACM ISBN 978-1-4503-4655-9/17/05 ...$15.00.
http://dx.doi.org/10.1145/3025453.3025640

engine. For each segment, the *Respeak* engine combines the output transcripts obtained from users into one best estimation transcript by using multiple string alignment and majority voting. Each submitted transcript earns *Respeak* users a reward of mobile airtime depending on the similarity between the transcript they submitted and the best estimation transcript. Finally, the engine concatenates the best estimation transcript for each segment to yield a final transcript of the original audio file.

We conducted a series of cognitive experiments with 24 university students to obtain key design insights into the segmentation process, ordering of tasks, and comparison of speaking and typing skills of potential users. We then seeded *Respeak* with 21 audio files in Hindi and Indian English containing 55 minutes of widely-varied content and deployed the *Respeak* application for one month with 25 university students in India. The *Respeak* engine segmented these audio files to obtain 756 short segments and presented them as micro-transcription tasks to the application users. Collectively, *Respeak* users performed 5464 tasks with an individual average WER of 23.7%, and earned USD 46. The *Respeak* engine aligned the transcripts generated by five randomly selected users to reduce the transcription WER to 10.6%. *Respeak* was particularly effective for Hindi and produced transcriptions with an average WER of 8.6%. The cost of speech transcription was USD 0.83 per minute, and the turnaround time was 39.8 hours, substantially less than the industry standard of USD 5 per minute for Hindi and Indian English transcription. In addition to providing users mobile airtime, with an expected payout of USD 1.16 per hour of usage, *Respeak* also provided them instrumental benefits, such as improved vocabulary, pronunciation and oral skills, a new-found interest in content, and a fun cognitive exercise. We discuss the lessons learned from the cognitive experiments and deployment, strengths and weaknesses of *Respeak*, and the implications for the future of voice-based crowdsourcing marketplaces.

RELATED WORK

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

The advent of crowdsourcing has had a profound effect on the speech transcription industry. Several service providers — such as SpeechPad [15], CastingWords [3], TranscribeMe [17], Rev [12], CrowdSurf [5], and Tigerfish [16] — are capitalizing on the strengths of crowdsourcing and manual transcription. However, most of these providers support only popular accents of English, and none of them support any local language spoken in developing countries. Moreover, their cost varies from USD 1–6 per minute depending on several parameters noted earlier. Workers also transcribe files that can be up to several hours long, making transcription a high cognitive load exercise. Moreover, inclusion criteria noted previously severely limit the adoption of these platforms in developing countries. In India alone, 47% of the 15+ year population does not have a bank account [2] making it hard to compensate them for their work, 97% of households do not have access to a computer connected to the Internet [1], and the typing speed, even of college students, is around 29.5 WPM (more details in the next section). Though several online non-crowdsourcing portals [11, 14, 18] transcribe content in languages spoken in developing regions (like Hindi, Marathi, Urdu and Indian English) using a fleet of transcribers, the cost of transcription starts at USD 5 per minute.

Recent years have seen the rise of several crowdsourcing marketplaces specifically designed to bring additional earning opportunities to low-income people in resource-constrained settings. mClerk [24] and MobileWorks [37] let low-income people transcribe picture SMS sent to their feature mobile phones or perform optical character recognition tasks; TxtEagle lets basic phone users answer survey questions and perform translation tasks using text messages [21]. The rapidly decreasing cost of smartphones and Internet access is helping many of those with low-income and limited technology skills gain access to low-end smartphones and Internet. Jana [9] capitalizes on this trend by transferring mobile airtime to their users who watch videos produced by brands and download smartphone applications. Samasource [13] maintains a dedicated workforce of low-income people who perform varied tasks, including categorization, data mining and transcription. However, the transcription tasks on these platforms require workers to have acceptable typing skills in English and other local languages. *Respeak*, on the other hand, lets low-income people perform speech transcription by listening to the audio and speaking back into an ASR system, which is more natural and usable than typing.

Prior speech transcription research has designed editing tools to improve transcripts generated by an ASR system [20, 26, 41, 44], built speech acquisition systems using Mechanical Turk to expand language corpora [27, 29, 32], and used gamification for speech labeling, identifying accents and prosody annotations [19, 22, 23, 31]. Parent and Eskenazi [38] and Lee and Glass [30] used a two-stage, crowd-powered speech transcription process, where audio files were broken into short segments to reduce cognitive load on workers. Parent and Eskenazi requested that workers first label each utterance and corresponding transcript generated by ASR as understandable/non-understandable and correct/incorrect. The workers were then asked to transcribe understandable but incorrect segments by typing them. A majority vote on the output generated by three workers for each utterance was used to attain the final transcription with a WER of 8.1% without controlling workers' quality. Similarly, Lee and Glass requested workers to type transcriptions for short segments; once each segment was transcribed by one worker, the short transcripts were combined to create the final transcript by using performance estimation and filtering to attain a WER of 10.2%. Lasecki et al. [28] designed a real-time captioning system where non-expert crowd workers transcribed overlapping segments of audio by typing; these segments were merged in real-time by multiple string alignment and majority voting [36] to yield a WER of 45% when attributes were balanced to achieve 66% coverage and 88% precision. *Respeak* draws on existing research by using a two-stage process that segments a large audio file into smaller utterances and then merge transcripts generated for the utterances using multiple string alignment and majority voting. However, unlike other systems, *Respeak* users speak the content into a standard built-in speech recognition engine rather than typing it. Using speaking skills rather than typing skills makes *Respeak* easy and natural to use, especially for people with no or low typing skills.

Prior research exploring re-speaking [25, 39, 42] requires significant data to generate speaker dependent acoustic models and domain dependent language models, making these solutions expensive and untenable at scale. *Respeak*, on the other hand, uses an off-the-shelf generic ASR system and combines transcripts generated by multiple users to reduce ASR errors. Rather than relying on high-skilled re-speakers that have undergone an intensive training of several months and capable of handling multiple hours of captioning without break in a controlled environment [39], *Respeak* rely on multiple unskilled crowd workers to perform micro re-speaking tasks in their everyday environment. Lastly, *Respeak* provides transcription for resource-constrained languages and accents that yield lower ASR accuracy than the well-represented languages and accents, such as English and Japanese, used in prior works.

RESPEAK PROCESS OVERVIEW

Respeak — a voice-based, crowd-powered speech transcription system — combines the benefits of both human intelligence and ASR systems while mitigating their weaknesses by using a five-step process, as follows:

1. **Segmentation:** The *Respeak* engine segments a large audio file for transcription into short utterances that are easier for *Respeak* smartphone application users to remember.

2. **Distribution to Crowd Workers:** Each segment is sent to multiple *Respeak* smartphone application users.

3. **Transcription by Crowd Workers using ASR:** *Respeak* users listen to the segment and repeat the same words into the application in a quiet environment. The application uses the Android ASR API to obtain a transcript for the spoken segment and displays it to the user. The transcript thus produced is expected to have a high WER. The user submits the transcript for the current segment and then receives a new micro-transcription task.

4. **First-stage Merging:** For each segment, the *Respeak* engine combines multiple users' output transcripts into one best estimation transcript using multiple string alignment (MSA) and majority voting. If the errors in the output from the ASR system are randomly distributed, merging the transcripts from different users reduces the overall WER as the correct word is recognized for the majority of users. The transcript sent by the user is compared to the best estimation transcript obtained using MSA and majority voting to determine the reward. Once the cumulative reward amount earned by a user reaches INR 10[1], a mobile airtime credit of INR 10 is sent to the user by the *Respeak* engine.

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

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

COGNITIVE EXPERIMENTS FOR INTERFACE DESIGN

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

1. How audio segment length affects content retention and cognitive load experienced by a *Respeak* user.

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

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

Methodology for Cognitive Experiments

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

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

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

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

Table 1. Self-assessment of participants' language skills.

played the selected segment multiple times and then re-spoke the content once they memorized it.

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

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

Cognitive Experiments Participants' Demographics

We used a campus-wide email list from a university in India to invite participation and randomly selected 24 respondents. Seventeen participants were male, and seven were female. The average age of participants was 24.4 years. Eight participants were summer interns at the university; five were hired as project staff; four were pursuing a bachelor's, four a master's, and three a Ph.D. degree. Twenty participants were from the engineering disciplines and four were from the humanities. All but one participants owned a smartphone with Internet access. The average daily phone and computer usage was reported to be around 5.5 hours and 10 hours, respectively. Nine participants knew about crowdsourcing platforms, but only two had used them previously. As Table 1 shows, the majority of participants assessed their Hindi typing skills as being very bad.

Findings of Cognitive Experiments

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

Experiment 1: Impact of Segment Length on Retention

Figure 1 compares the WER, time taken to retain and re-speak segments, and number of times segments were listened in the first experiment. A repeated measures ANOVA with a Greenhouse-Geisser correction determined a statistically significant difference (at p<.001) in the three parameters for the segments that were 1–7 seconds long. Table 2 highlights the parameters that significantly differ (at p<.05) on a pairwise comparison of the segments of different lengths. The WER and time taken were much higher for segments exceeding five seconds. Our interviews also revealed that several participants could not retain such segments because of complicated sentence constructions and the excessive number of concepts to remember. Thus, using such segments in *Respeak* could result in a poor accuracy speech transcription and put significant cognitive load on users. Eleven participants used synonyms or missed articles while re-speaking content. Three participants found it difficult to retain segments containing an incoherent word. Another three participants found it challenging to retain unfamiliar proper nouns. Four participants found content retention to depend on their familiarity with subject matter rather than on duration. One of them stated:

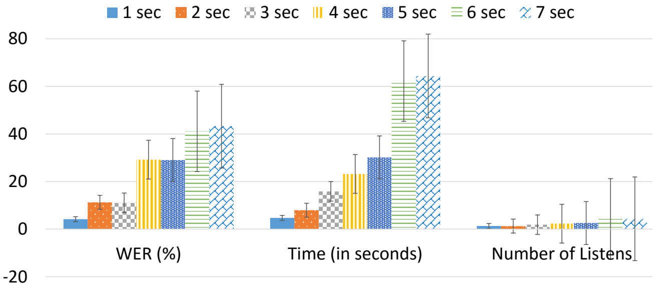


Figure 1. Comparison of varying length segments on several parameters.

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

Table 2. Significant difference in WER (W), completion time (T) and number of listens (L) on pairwise comparison of varied length segments.

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

Twelve participants found it difficult to retain segments containing partial sentences. Abrupt cuts resulting in an incomplete or incoherent word made it substantially more difficult to retain the segment. A participant stated:

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

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

Experiment 2: Impact of Segment Ordering on Retention

We conducted independent samples t-test to analyze the effect of segment ordering on content retention. We found a significant difference in the WER when segments were played

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

Table 3. Mean (M) and standard deviation (SD) for computer typing (CT), phone typing (PT) and speaking (S) tasks.

sequentially (M = 16.59, SD = 3.85) rather than randomly (M = 32.27, SD = 13.18); $t(22)$ = 3.96, p=.001. We did not find a difference in task completion time or the number of times participants listened to segments. These results suggest that content retention is much higher when segments are presented sequentially. In the interviews, five participants specifically mentioned that sequential ordering increased their understanding of context, making it easier to estimate incoherent and clipped words. One of them stated:

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

Experiment 3: Speaking versus Typing

Table 3 shows descriptive statistics for the tasks in the third experiment. The average speed of computer typing, phone typing, and speaking was 29.5, 19.3, and 161 WPM, respectively. A repeated measures ANOVA with a Greenhouse-Geisser correction determined a statistically significant difference in task completion time, $F(1.03, 23.74)$ = 25.41, p<.001. Post hoc tests using the Bonferroni correction also revealed a significant difference (p<.05) in task completion time, even for pairwise comparisons of all three tasks. Though the average WER for speaking was lower than for typing, we did not find any statistical evidence to substantiate this.

We also requested participants to rate the three tasks on a ten-point scale for NASA TLX parameters to assess subjective workload. As seen in Figure 2, participants found that speaking caused the least mental demand, physical demand, effort, and frustration. Moreover, they perceived their performance for the speaking task to be higher than for the typing tasks. A participant explained the ease of speaking vs. typing content:

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

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

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

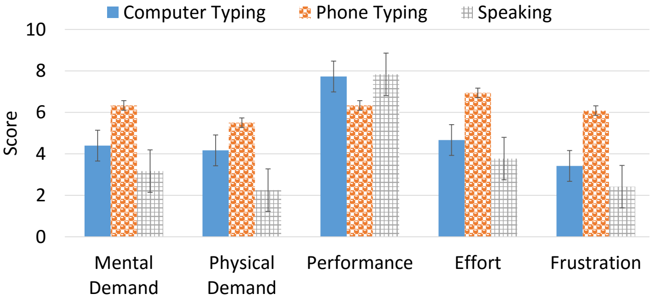


Figure 2. Evaluation of output modes on NASA TLX parameters.

FIRST-STAGE MERGING USING CROWDSOURCING

Respeak's speech transcription efficiency depends on the performance of the multiple string alignment (MSA) algorithm, and the majority voting process that is the core of first-stage merging. For each segment, the *Respeak* engine combines the transcripts submitted by *Respeak* users to produce a best estimation of actual word sequence in the segment. The MSA algorithm in our implementation uses word as the individual atomic unit rather than character or phoneme. We adapted and implemented the multiple sequence alignment algorithm proposed by Naim et al. [36] that uses the A* search algorithm to reduce the search space of multi-dimensional lattice. Any ties during majority voting are broken randomly. Let us assume that first-stage merging takes as input transcripts generated by K users by repeating the same phrase in the *Respeak* mobile application. If the ASR errors are uncorrelated across users, then the WER of the hypothesized word sequence should decrease as K increases. Let P be the average accuracy of speech recognition for individual users. The WER then is $1-P$. Assuming that the errors are randomly distributed across users, the accuracy of the alignment of segments (P_{final}) for N users computed using majority voting is:

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

Figure 3 depicts the estimated improvement in accuracy achievable by aligning the transcripts generated by one, three, five and seven users for several values of P.

To test the feasibility and performance of first-stage merging, we designed, built and deployed a simple smartphone application to collect data from 29 university students in India who volunteered to take part in the experiment. The participants were requested to read the short segments rather than memorize and re-speak them. The speech-to-text output was instantly displayed on the application using the built-in Google ASR engine. There were 25 short segments in Indian English and ten short segments in Hindi. The average number of words were 13 (min=5, max=33, SD=8.08) for English segments and 15 (min=9, max=27, SD=5.6) for Hindi segments. Seventeen participants collected data for English and twelve for Hindi.

Participants had different WERs due to a variety of factors, such as differing background noise, accents, and speaking

Ground truth:	it	is		strong	in	the	market
Transcript 1:	it	is	BLANK	BLANK	from	the	market
Transcript 2:	it	is	BLANK	strong	in	the	market
Transcript 3:	it	is	a	strong	in	the	market
Majority voting:	it	is	BLANK	strong	in	the	market
Reconstructed segment:	it	is		strong	in	the	market

Table 4. Alignment of segment using MSA and majority voting on transcripts obtained from 3 speakers.

Users (K)	Expected WER	Actual WER
1	33%	33%
3	26%	25%
5	21%	17%
7	18%	19%

Table 5. Word error rates (WERs) after MSA and majority voting.

rate. The best speaker had a WER of 11%, the worst had a WER of 58%, and the average WER was 33%. To measure the degree of improvement in transcription by using MSA and majority voting, we varied the number of speakers used to align transcripts. Table 4 illustrates the alignment of the segment *"it is strong in the market"* from three speakers. The ASR transcript contained an error for all three speakers. However, combining the transcripts generated by ASR for all three speakers mitigated the errors made by individual speakers. Table 5 reports the WERs averaged over ten runs of experiments where K speakers were randomly selected from the pool of 29 participants for each run. The WER of the alignment using seven speakers exceeded that for 28 participants and was comparable to the theoretically estimated improvement. The field evaluation validated our hypothesis that aligning transcripts generated by multiple users decreases the WER when ASR errors are randomly distributed.

FIELD DEPLOYMENT IN INDIA

Since many university students have smartphones connected to the Internet and also have financial constraints that might motivate them to earn mobile airtime, we sent an email inviting students in a university in Mumbai, India to participate in our controlled deployment. We randomly selected 25 respondents as users and conducted a face-to-face orientation session with them to install the *Respeak* application on their personal smartphones, show them how to use the application, and collect demographic information.

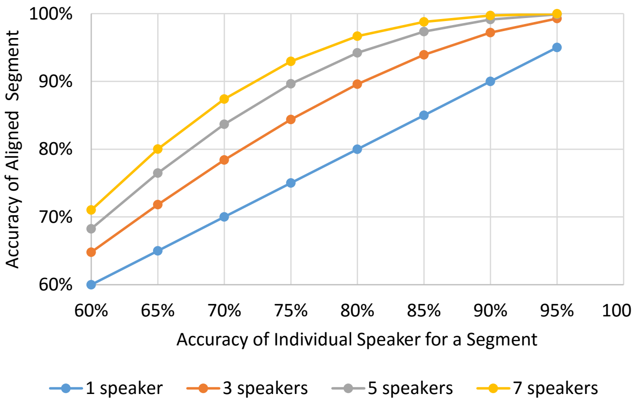


Figure 3. Improvement in accuracy by using MSA and majority voting.

Tasks

We submitted thirteen Hindi and eight English audio files to the *Respeak* engine for transcription. To stress test *Respeak*, we selected audio files that had ambient noise and heavily localized Hindi or English accents. The files contained varied content, including public speeches, telephone conversations, news, television advertisements, songs, interviews, YouTube content, and online lectures. The total duration of the Hindi and English files was 43 minutes and 12 minutes, respectively. The *Respeak* engine partitioned Hindi files into 499 segments and English files into 257 segments to yield 756 unique micro-transcription tasks (see Table 6). The threshold length for the segmentation scheme was based on the speaking rate in the audio file: the length for public speeches and songs was 5–6 seconds, news and YouTube videos was 4 seconds, and interviews and phone calls was 3 seconds. The collective download size of all tasks was 85 MB and the cost of downloading them was roughly 20 INR (USD 0.30) on a 3G connection. Each task could be performed by a maximum of ten users who could see a high-level overview of their transcription accuracy, amount earned, payment processed, and completed tasks.

Payment Scale

The *Respeak* reward structure was designed to keep the cost of transcription below USD 1 per minute. Each transcription task was assigned a reward equal to 0.2 INR multiplied by segment length in seconds. We hypothesized that for each segment, if we aligned the transcripts generated by five users and if all of them performed the task with a high accuracy, the maximum transcription cost would still be USD 0.92 per minute. Each time a user submitted a transcript for a segment, we compared their output with the pre-computed ground truth. If the transcript's accuracy was $\geq 80\%$, we added the entire task reward to the user' earning. If the accuracy was $\geq 50\%$, we added a proportionate percentage of the task reward to the user' earning. A user received no reward if the accuracy was $< 50\%$. This reward structure gave users the incentives to produce speech transcription with more than 80% accuracy, gave proportionate returns to average performers, and penalized poor performers. Once the cumulative earnings of a user reached 10 INR, we processed a mobile airtime credit of the same value to them. The maximum amount a *Respeak* user could earn by doing Hindi tasks was 514 INR (USD 7.80) and by doing English tasks was 152 INR (USD 2.30). The reward structure could also be designed differently to satisfy other optimization goals.

Ideally, the transcripts submitted by *Respeak* users should be evaluated by comparing them to the best estimation transcript generated in first-stage merging. However, at the time of experiment we were unsure about how and when people would use the application, forcing us to use the pre-computed ground truth for comparison. We ran the comparison module every

	Interview			Song			TV ad			News			Public speech			Phone call			YouTube video			Lecture		
	File	Task	Len	File	Task	Len	File	Task	Len	File	Task	Len	File	Task	Len	File	Task	Len	File	Task	Len	File	Task	Len
English	2	177	494	-	-	-	1	10	40	1	10	30	1	15	60	-	-	-	2	35	105	1	9	27
Hindi	-	-	-	3	77	431	-	-	-	1	17	51	5	313	1760	3	37	111	1	54	216	-	-	-
Total	2	177	494	3	77	431	1	10	40	2	27	81	6	328	1820	3	37	111	3	89	321	1	9	27

Table 6. Number of files, tasks and length of files (in seconds) used for each category of transcribed content by language.

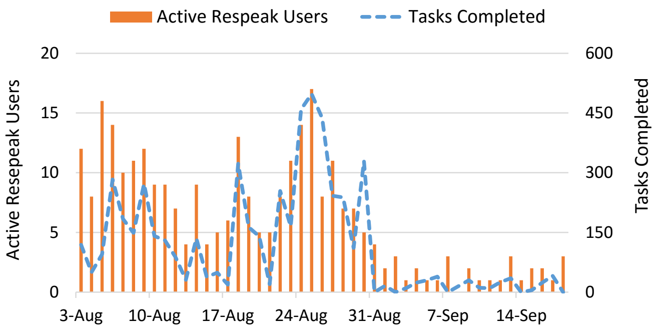


Figure 4. Time series analysis of active users and tasks completed.

15 minutes to balance the desire of users to receive immediate feedback and the need to simulate a delay that would occur awaiting transcripts generated by others for MSA and majority voting process.

Methodology to Evaluate Deployment

We used a mixed-methods approach spanning quantitative analyses of performance, cost, and turnaround time, and qualitative interviews to evaluate *Respeak*. We conducted in-depth, semi-structured, face-to-face interviews with 20 *Respeak* users at the end of deployment. Each interview lasted around 40 minutes and covered several themes, including information on the general technology use, user experience and usability, conception of *Respeak*, and benefits and limitations of the *Respeak* application. The interviews were conducted by the first author (male, 29 years old, Hindi speaking) and were recorded, transcribed, and analyzed using open coding.

FINDINGS

The *Respeak* application was deployed for a month with 25 users. Figure 4 depicts the time series analysis of the number of tasks completed by active *Respeak* users. The low activity between August 10–20 corresponded to an intermittent campus-wide Internet outage at the university where we deployed *Respeak*. Though the deployment ended on August 31, some users continued using the application for another 20 days. 756 audio segments were presented as 5464 micro-tasks to the users, who transcribed the segments successfully with an individual average WER of 23.7%. On average, *Respeak* users listened to segments 2.7 times and re-spoke them 2.1 times before moving on to the next task. The median time for task completion was 36 seconds, and the cost of transcription was USD 0.83 per minute. Collectively, *Respeak* users spent 39.8 hours using the system and earned 3036 INR (USD 46). The expected payout for an hour of their time was 76 INR (USD 1.16), one-fourth of the average daily wage rate in India [7]. The *Respeak* engine combined the transcripts generated by five users for each segment, reducing the average WER to 10.6%. The best alignment yielded a WER of 6.8%.

Respeak Users Demographics

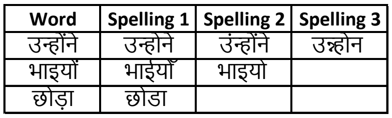
Fifteen *Respeak* users were male and ten were female. Fourteen were students, six were contractual staff, and five were summer interns. Twenty users were from varied engineering departments, and five from the humanities. Eighteen users had or were pursuing a bachelor's degree, six had or were pursuing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

All users owned an Android smartphone, had cellular Internet access, and used their phones for an average of 5 hours a day. Despite heavy and ubiquitous phone usage, 17 participants had a shoestring budget for mobile airtime and data, and relied on the free WiFi provided by the university. Like participants in the cognitive experiments, all users rated their English language skills and Hindi speaking skills highly. However, 22 users reported their Hindi typing skills to be bad. Sixteen of them did not even know how to type in Hindi. Sixteen users were unaware of crowdsourcing systems, and only three had used them previously. After we explained crowdsourcing to them, they expressed an interest in using such systems to earn money (N=14), gain knowledge (N=13), help others (N=6), and spend their copious time productively (N=3).

Efficiency of Speech Transcription

The average WER of the transcription generated by ASR engine for individual *Respeak* users was 23.7%. We performed a series of experiments to measure the improvement in transcription using MSA and majority voting. For each segment, we conducted ten runs of experiments. In each run, the transcripts generated by three randomly selected *Respeak* users were used for MSA and majority voting in first-stage merging. The WERs obtained in each of the ten runs were averaged for evaluation. By aligning the transcripts generated by 3 speakers, the WER dropped to 15.1% — an improvement of 36.3%. We used the same setup to align transcripts generated for each segment by 5 randomly selected *Respeak* users, and the WER dropped further to 13.2% — an improvement of 44.3%.

A closer inspection of users' transcripts and the ground truth revealed interesting cases that were registered as errors by the comparison module but were semantically correct. The application's Google ASR engine transcribed several words in English and Hindi differently for different speakers. In English, the words were often contracted or abbreviated (e.g., it is vs. it's; Doctor vs. Dr.), and the numbers were transcribed either in numeric or textual format (e.g., 3 vs. three) for different speakers. In Hindi, multiple spellings with minor variations were output for the same word depending on the stress, intonation and nasality used by speakers (see Figure 5).



Word	Spelling 1	Spelling 2	Spelling 3
उन्होंने	उन्होने	उंन्होंने	उन्होन
भाइयों	भाईयोँ	भाइयो	
छोड़ा	छोडा		

Figure 5. Different spellings generated by Google ASR engine for words in Hindi.

The manual correction of such corner cases in *Respeak* user transcripts lowered the average WER for individual users from 23.7% to 21.9%. We recomputed the set of experiments where transcripts generated by multiple users were aligned, and the WER dropped to 12.5% and 10.6% when transcripts generated by 3 and 5 randomly selected users were aligned, respectively. Thus, the alignment of transcripts generated by five randomly selected users reduced the average WER by 55.3%. For future deployments, we recommend using existing dictionaries or building a customized dictionary to resolve these corner cases automatically in the comparison module.

To evaluate the effect on WER and cost as more transcripts are used for alignment and majority voting, we randomly selected 50% of 391 tasks that were each completed by ten *Respeak* users. We conducted ten runs of experiments; in each run, we used the transcripts generated by *K* randomly selected *Respeak* users. We varied the value of *K* from 1–9 and averaged the WER obtained for each value of *K* over ten runs of experiments. The cost of transcription was calculated using the rate of 0.2 INR per second of transcription per user — an overestimate that assumed that users would receive the entire reward amount promised for each task. As depicted in Figure 6, the WER decreased as *K* increased, and the cost of speech transcription linearly increased with *K*.

To compare *Respeak* with a state-of-the-art speech recognition engine, we submitted the original audio files to the Google Cloud Speech API [4] after noise reduction. The API yielded transcription with the overall WER of 50% — 4.72 times higher than the WER obtained by *Respeak*. The WER for Hindi audio files was 54% — 6.3 times higher than the WER obtained by *Respeak*. These results suggest that *Respeak* capitalized on the benefits of re-speaking and crowdsourcing to outperform transcription generated by the state-of-the-art speech recognition engine.

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

Table 7. WER obtained by *Respeak* for English and Hindi languages.

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

Table 8. WER obtained by *Respeak* for different content categories.

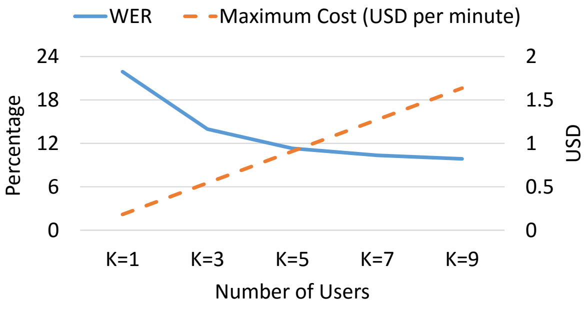


Figure 6. Effect of number of users on WER and cost.

Tables 7 and 8 report the WER of transcription obtained for different languages and content types by aligning transcripts generated by K *Respeak* users during first-stage merging. Our interviews revealed that six users found it easier to do Hindi tasks, and four found it easier to do English tasks. The language preference existed either because of better language skills or faster recognition from the ASR engine in their preferred language. Seven users found it easiest to re-speak song segments while others found interviews (N=3), speeches (N=2), news (N=1), lectures (N=1) and poems (N=1) the easiest. Six users found it very difficult to understand the segments containing an interview of a former president of India, three found it hardest to retain the advertisement segments because of the audio's unclear accent, and the other three found it difficult to re-speak Bollywood song segments because of the *"cheesy"* lyrics. The remaining users found no difference in the difficulty level of tasks based on content type. Three users sang the segments containing songs rather than merely re-speaking them. A user explained how he had to remain aware of his surroundings while re-speaking song segments:

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

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

Figure 7 plots the average WER for segments of varying word lengths. Surprisingly, the WER for individual users did not vary significantly as the number of words in segments increased. However, the improvement in WER by aligning transcripts from 5 users rather than 3 users decreased as the number of words in a segment increased. Though the randomness in errors increased with the increase in number of words in a segment, the errors were sparsely distributed reducing the performance improvement gained by MSA and majority vote.

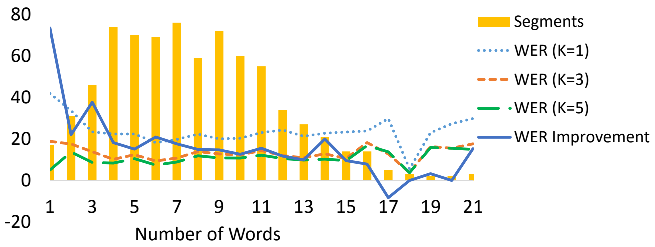


Figure 7. WER for segments of varying word length.

Amount Earned (INR) ≤	*Respeak* Users
100	15
200	5
300	1
400	0
500	3

Table 9. Amount earned by *Respeak* users.

Payments

We processed mobile airtime of 3040 INR (USD 46) to 24 users. The top 3 users earned 44% of the total payments, while the top 20% and 50% users earned 60% and 87%, respectively (see Table 9). Ten users earned more than their monthly phone expense. Several users reported that using the application for ten minutes daily was sufficient to subsidize their phone expenses. *Respeak* became a portal to transfer mobile airtime to their phones. A user reported:

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

All but one users were happy to receive the amount earned as mobile airtime. Some users suggested payments in the form of food coupons (N=6), Amazon gift coupons (N=5), and top-ups of higher value that results in the equivalent mobile airtime [2] (N=3). Two users emphasized the need to process a 10 INR mobile airtime for immediate gratification. A user stated:

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

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

Instrumental Benefits

Seven *Respeak* users reported receiving instrumental benefits from the application use. Three found that *Respeak* improved their language and oral skills. While re-speaking audio segments, they focused on pronouncing the words correctly for faster recognition by the ASR engine. Often, they searched online for the meaning and pronunciation of unfamiliar words, thereby expanding their vocabulary. *Respeak* provided them with the opportunity to speak English aloud *"without being judged by others."* One user reported a new-found interest in the content he transcribed, viz., an old Bollywood song on YouTube. Another found *Respeak* to be a challenging yet fun exercise that improved his cognitive abilities. Two users reported acquiring new knowledge while doing the tasks and found some of the speeches inspiring. One of them stated:

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

Feedback on Respeak

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

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

The users considered the ideal *Respeak* demographic to include: students (N=15), unemployed people (N=4), home makers (N=2), people spending long hours commuting (N=2), and those interested in learning oral skills (N=1). After the deployment, eleven participants expressed an interest in using the application daily, primarily to earn mobile airtime and

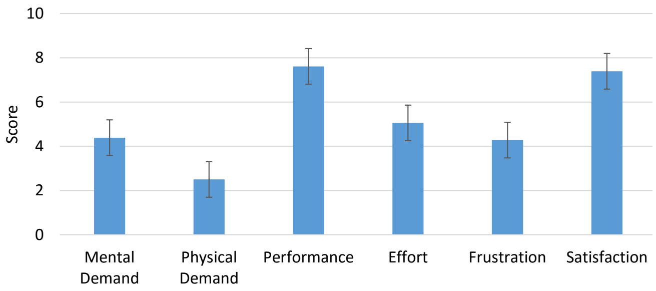


Figure 8. Average ratings by *Respeak* users for several parameters.

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

improve their language skills; six stated that they would use it sparingly when they need mobile airtime; three stated that the lack of time would inhibit their application use in the future.

DISCUSSION AND CONCLUSIONS

Respeak was successful in producing efficient and cost-effective speech transcription with a low turnaround time for widely-varied Hindi and Indian English audio files. Our cognitive experiments were instrumental in designing *Respeak*; they revealed that audio files should be partitioned by detecting natural pauses to yield segments of less than six-second length and presented sequentially to increase retention and decrease cognitive load. In the deployment with 25 university students in India, *Respeak* produced speech transcription with an accuracy of 89.4% and turnaround time of 39.8 hours. The average WER decreased from 21.9% to 12.5% (three users) and further to 10.2% (five users) when the transcripts generated by multiple users were aligned using MSA and majority voting to reduce randomly distributed ASR errors. *Respeak* transcribed audio content in Hindi and Indian English with a WER of 8.6% and 15.2%, respectively. The WER for Hindi tasks was much lower than that for English tasks because of users' better listening and speaking skills in Hindi. *Respeak* produced high-accuracy transcription even for audio files with high ambient noises like songs (WER=7.8%) and telephone calls (WER=12.8%) as humans are better than ASR systems at ignoring ambient noise and interpreting unclear speech. Since the cost of transcription using *Respeak* (USD 0.83 per minute) is only one-sixth of the industry standard of USD 5 per minute for Hindi and Indian English content [11, 14, 18], it is feasible to increase the worker payout rate of USD 1.16 per hour in future iterations to make *Respeak* more lucrative for workers.

One key strength of *Respeak* is its voice-based implementation: *Respeak* let users generate transcripts by speaking rather than typing. Though our deployment had technology-savvy users who were engineering students, the majority did not know how to type in Hindi, and those who knew had poor Hindi typing skills. *Respeak* capitalized on users' speaking skills rather than typing skills that are scarce for languages that do not use Latin script. Though several technologies, like Google Input Tools [6] and India Typing [8], take Latin script input and produce Devanagari script output using transliteration, speaking was much easier for *Respeak* users since voice is a natural and accessible medium of interaction. The reliance on speaking skills makes *Respeak* more inclusive of people with no or poor typing skills.

Respeak has some limitations, however. Its users were unsure what words to re-speak when multiple people simultaneously spoke in a segment. Moreover, the *Respeak* engine did not distinguish speakers in transcription for audio files with multiple speakers. Future versions could consider an improved segmentation scheme that is cognizant of speakers in an audio file containing multiple speakers. Further, the transcription generated by *Respeak* lacked punctuation marks. Though punctuation marks could be added automatically based on the identification and length of natural pauses, a better algorithm is needed when it would be difficult to detect natural pauses due to ambient noise. One possible solution could be to send an audio segment and corresponding transcript generated by *Respeak* to users, who are then asked to identify speakers and place punctuation marks.

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

One of the most direct ways to empower low-income communities in resource-constrained settings is to provide them with additional earning opportunities. Recent years have seen rapid increase in penetration and decrease in cost of smartphones and Internet in developing countries [34, 33]. *Respeak* provides a definite step forward in realizing a smartphone- and voice-based crowdsourcing marketplace. *Respeak* was the first crowdsourcing platform for 88% of its users, who were technology-savvy university students. Fifteen users were financially dependent on their family members, and ten others earned USD 9.76 per day through their job. The amount earned by using *Respeak* was significant for many technology-savvy low-income literate people in our deployment, and we believe *Respeak* has potential to be transformative for other marginalized communities including low-literate and visually impaired people. Our immediate next step is to conduct more deployments with these populations to investigate the external validity of our findings.

Respeak could also be used to subsidize the cost of participation of low-income, low-literate people on voice-based social computing platforms, such as CGNet Swara [35], Sangeet Swara [43], and Polly [40]. More work is needed to adapt *Respeak* to basic mobile phones, where users could call-in to an interactive voice response system, listen to a short segment, re-speak it, and earn mobile airtime. In doing this work, we confront several interesting challenges, such as ensuring reasonable accuracy of ASR systems on telephone lines, providing immediate feedback on transcripts generated by ASR systems, and managing the overhead cost of voice calls. The *Respeak* call-in service could have a high potential to provide additional earning opportunities through crowdsourcing marketplaces to billions of low-literate people with access to a basic phone.

ACKNOWLEDGMENTS

We are grateful to Gaetano Borriello, Aaron Jaech, Ganesh Ramakrishnan, Bill Thies and the anonymous reviewers for their helpful suggestions. This work is funded by a grant (#AID-OAA-G-14-00014) from USAID.

REFERENCES

1. 2012. *Press Note on Release of Data on Houses, Household Amenities and Assets, Census 2011*. Technical Report. Ministry of Home Affairs, Government of India. http://censusindia.gov.in/2011census/hlo/Data_sheet/India/HLO_Press_Release.pdf
2. 2014. *Global Findex 2014 - Financial Inclusion*. Technical Report. World Bank. http://datatopics.worldbank.org/financialinclusion/country/india
3. 2016. CastingWords. (2016). https://castingwords.com/.
4. 2016. CLOUD SPEECH API: Speech to text conversion powered by machine learning. (2016). https://cloud.google.com/speech/.
5. 2016. CrowdSurf. (2016). http://crowdsurfwork.com/.
6. 2016. Google Input Tools. (2016). https://www.google.com/inputtools/.
7. 2016. India Average Daily Wage Rate Forecast 2016-2020. (2016). http://www.tradingeconomics.com/india/wages/forecast.
8. 2016. India Typing. (2016). http://indiatyping.com/.
9. 2016. Jana. (2016). https://www.jana.com/.
10. 2016. *Medical Transcription Services Market - Global Industry Analysis, Size, Share, Growth, Trends and Forecast, 2013 - 2019*. Technical Report. Transparency Market Research.
11. 2016. Quick Transcription Service. (2016). http://www.quicktranscriptionservice.com/Hindi-Transcription.html.
12. 2016. Rev. (2016). https://www.rev.com/.
13. 2016. Samasource. (2016). http://www.samasource.org/.
14. 2016. Scripts Complete. (2016). http://scriptscomplete.com/Hindi-Transcription-Services.php.
15. 2016. SpeechPad. (2016). https://www.speechpad.com/.
16. 2016. Tigerfish. (2016). http://tigerfish.com/.
17. 2016a. TranscribeMe. (2016). http://transcribeme.com/.
18. 2016b. Transcription Services Us. (2016). http://www.transcription-services-us.com/Language-Transcription-Rates.php.
19. Rio Akasaka. 2009. *Foreign accented speech transcription and accent recognition using a game-based approach*. Ph.D. Dissertation. Swarthmore Department of Linguistics.
20. Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. 2012. Tools for Placing Cuts and Transitions in Interview Video. *ACM Transaction of Graphics* 31, 4 (2012).
21. Nathan Eagle. 2009. Txteagle: Mobile Crowdsourcing. In *Proceedings of HCI International*. Springer-Verlag.
22. Keelan Evanini and Klaus Zechner. 2011. Using crowdsourcing to provide prosodic annotations for non-native speech. In *Proceedings of Interspeech*.
23. Alexander Gruenstein, Ian McGraw, and Andrew Sutherland. 2009. A Self-Transcribing Speech Corpus: Collecting Continuous Speech with an Online Educational Game. In *Proceedings of SLaTE*.
24. Aakar Gupta, William Thies, Edward Cutrell, and Ravin Balakrishnan. 2012. mClerk: Enabling Mobile Crowdsourcing in Developing Regions. In *Proceedings of CHI*.
25. Toru Imai, Atsushi Matsui, Shinichi Homma, Takeshi Kobayakawa, Kazuo Onoe, Shoei Sato, and Akio Ando. 2002. Speech recognition with a re-speak method for subtitling live broadcasts. In *Proceedings of ICSLP*.
26. Jennifer Lai and John Vergo. 1997. MedSpeak: Report Creation with Continuous Speech Recognition. In *Proceedings of CHI*.
27. Ian Lane, Alex Waibel, Matthias Eck, and Kay Rottmann. 2010. Tools for Collecting Speech Corpora via Mechanical Turk. In *Proceedings of the NAACL HLT*.
28. Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time Captioning by Groups of Non-experts. In *Proceedings of UIST*.
29. Jonathan Ledlie, Billy Odero, Einat Minkov, Imre Kiss, and Joseph Polifroni. 2010. Crowd Translator: On Building Localized Speech Recognizers Through Micropayments. *SIGOPS Oper. Syst. Rev.* 43, 4 (Jan. 2010).
30. Chia-ying Lee and James Glass. 2011. A Transcription Task for Crowdsourcing with Automatic Quality Control. In *Proceedings of Interspeech*.
31. Ian Mcgraw, Er Gruenstein, and Andrew Sutherl. 2009. A Self-Labeling Speech Corpus: Collecting Spoken Words with an Online Educational Game. In *Proceedings of Interspeech*.
32. Ian Mcgraw, Chia-ying Lee, Lee Hetherington, Stephanie Seneff, and Jim Glass. 2010. Collecting Voices from the Cloud. In *Proceedings of LREC*.
33. Mary Meeker. 2015. *2015 Internet Trends*. Technical Report. KPCB. http://www.kpcb.com/blog/2015-internet-trends
34. Mary Meeker and Liang Wu. 2014. *2014 Internet Trends*. Technical Report. KPCB. https://www.kpcb.com/insights/2014-internet-trends
35. Preeti Mudliar, Jonathan Donner, and William Thies. 2012. Emergent Practices Around CGNet Swara, A Voice Forum for Citizen Journalism in Rural India. In *Proceedings of ICTD*.
36. Iftekhar Naim, Daniel Gildea, Walter S. Lasecki, and Jeffrey P. Bigham. 2013. Text Alignment for Real-Time Crowd Captioning. In *Proceesings of HLT-NAACL*.

37. Prayag Narula, David Rolnitzky, and Bjoern Hartmann. 2011. MobileWorks: A Mobile Crowdsourcing Platform for Workers at the Bottom of the Pyramid. In *In Proceedings of HCOMP*.

38. G. Parent and M. Eskenazi. 2010. Toward better crowdsourced transcription: Transcription of a year of the Let's Go Bus Information System data. In *Proceedings of SLT*.

39. Ales PrazÃąk, Zdenek Loose, Jan Trmal, Josef V. Psutka, and Josef Psutka. 2012. Novel Approach to Live Captioning Through Re-speaking: Tailoring Speech Recognition to Re-speaker's Needs.. In *Proceedings of Interspeech*.

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 CHI*.

41. Venkatesh Sivaraman, Dongwook Yoon, and Piotr Mitros. 2016. Simplified Audio Production in Asynchronous Voice-Based Discussions. In *Proceedings of CHI*.

42. Matthias Sperber, Graham Neubig, Christian Fugen, Satoshi Nakamura, and Alex Waibel. 2013. Efficient Speech Transcription Through Respeaking. In *Proceesings of Interspeech*.

43. Aditya Vashistha, Edward Cutrell, Gaetano Borriello, and William Thies. 2015. Sangeet Swara: A Community-Moderated Voice Forum in Rural India. In *Proceedings of CHI*.

44. Dongwook Yoon, Nicholas Chen, FranÃğois GuimbretiÃÍre, and Abigail Sellen. 2014. RichReview: Blending Ink, Speech, and Gesture to Support Collaborative Document Review. In *Proceedings of UIST*.