Costs and Benefits of Conducting Voice-based Surveys Versus Keypress-based Surveys on Interactive Voice Response Systems

Aman Khullar
aman.khullar@oniondev.com
Gram Vaani
India

Priyadarshi Hitesh
priyadarshi.hitesh.mcs19@cse.iitd.ac.in
IIT Delhi
India

Shoaib Rahman
shoaib.rahman@oniondev.com
Gram Vaani
India

Deepak Kumar
deepak.kumar@oniondev.com
Gram Vaani
India

Rachit Pandey
rachit.pandey@oniondev.com
Gram Vaani
India

Praveen Kumar
praveen.kumar@oniondev.com
Gram Vaani
India

Rajeshwari Tripathi
rajeshwari.tripathi@oniondev.com
Gram Vaani
India

Prince
prince@oniondev.com
Gram Vaani
India

Ankit Akash Jha
ankit.jha@alumni.iitd.ac.in
IIT Delhi
India

Himanshu
himanshurewar14@gmail.com
IIT Delhi
India

Aaditeshwar Seth
aseth@cse.iitd.ac.in
Gram Vaani, IIT Delhi
India

ABSTRACT

Recent improvements through machine learning in speech technologies and natural language processing has prompted active interest in the development of conversational agents for various tasks. We look at the area of data collection in low-resource settings among rural women in North India, and explore the feasibility of using voice-based surveys conducted through IVR (Interactive Voice Response) systems where users may speak their responses in a conversational manner through natural speech. Through an iterative design process and detailed user feedback, we describe several nuances in running voice-based surveys, and compare their accuracy of data collection through equivalent keypress-based surveys. We find strong user preferences for voice-based surveys, and comparable performance with keypress-based surveys for most types of questions. Our results suggest that voice-based conversational interfaces may hold significant potential to build interactive applications for low-income and less-literate populations. Our findings are likely to be useful for other researchers and practitioners using ICTs (Information and Communication Technologies) in developing regions.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

COMPASS '21, June 28-July 2, 2021, Virtual Event, Australia
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8453-7/21/06…$15.00
https://doi.org/10.1145/3460112.3471963

KEYWORDS

Interactive Voice Response systems, data collection, speech recognition, natural language processing, entity extraction, surveys

ACM Reference Format:
Aman Khullar, Priyadarshi Hitesh, Shoaib Rahman, Deepak Kumar, Rachit Pandey, Praveen Kumar, Rajeshwari Tripathi, Prince, Ankit Akash Jha, Himanshu, and Aaditeshwar Seth. 2021. Costs and Benefits of Conducting Voice-based Surveys Versus Keypress-based Surveys on Interactive Voice Response Systems. In *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '21), June 28-July 2, 2021, Virtual Event, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3460112.3471963

1 INTRODUCTION

Data collection through IVR (Interactive Voice Response) systems is widely used in low-resource settings where people may not have access to smartphones or a stable internet access, or may be less literate to participate in text-based surveys. IVR surveys are commonly done through keypresses (also called DTMF - Dial Tone Multi Touch) [10], in which the respondents are asked to press keys on their phone key-pads to answer questions. An alternative approach to process voice-based inputs also exists and is the topic of our research, to understand the feasibility and relative merits/demerits of the approach as compared to the keypress-based method.

Earlier studies [8, 17] had by and large arrived at a consensus of using keypress-based instead of voice-based surveys because speech recognition was not robust enough, and it was unrealistic to expect users to provide one-word responses instead of speaking out the answer in a sentence [17]. Recent advances in speech recognition and the wide availability of commercial libraries in several languages [6], coupled with advances in natural language processing to parse narrative responses and extract relevant entities, merits a revisit to this understanding.

We carry out this research in two parts. First, we describe an iterative design process that we undertook to get better at conducting voice-based surveys. Through several rounds of experiments in controlled and uncontrolled settings, we discuss several nuances to improve doing voice-based surveys. Second, we run a direct comparison study to compare a voice-based survey with an equivalent keypress-based survey, for several common types of questions. We investigate the relative merits and demerits for MCQs (Multiple Choice Questions), single-digit and multi-digit numeric questions, multi-level complex questions that have several parts, and location-based input questions. The evaluation is conducted through an analysis of usage logs, coupled with observations and qualitative feedback from users who participated in these surveys. We find that users have a high preference for voice-based surveys, and for most question-types voice-based surveys are able to do better or just as well as keypress-based surveys in terms of task completion and response accuracy.

The experiments were conducted through a voice-based participatory media platform called JEEViKA Mobile Vaani, which runs on IVR systems in several blocks of the Nalanda district in the state of Bihar in India. The platform is predominantly used by women who are members of the JEEViKA SHGs (Self Help Groups), and is operated by the social enterprise Gram Vaani, in partnership with the Government of Bihar. The primary goal of the platform is to create awareness among SHG members on health and nutrition practices for pregnant mothers and small children, along with providing information to them on livelihood opportunities and agricultural practices, and most significantly during 2020 on COVID-19 news and updates [3, 24]. Users can access the platform by placing a call to a unique phone number publicized in the community. These calls are regular phone calls that can be made through simple non-smartphones. Upon receiving a call, the IVR servers cut the call and call the person back, thereby making the service free of cost to the users. Users can then listen to audio messages published on the platform, or record their own message, wherein they may want to ask a question or share an opinion or experience related to audio programmes that they hear. These recorded messages are moderated by a team of content moderators, and published back on the platform if they pass certain editorial checks. A large volume of research on voice-forums has investigated many aspects of such systems [14, 15, 22].

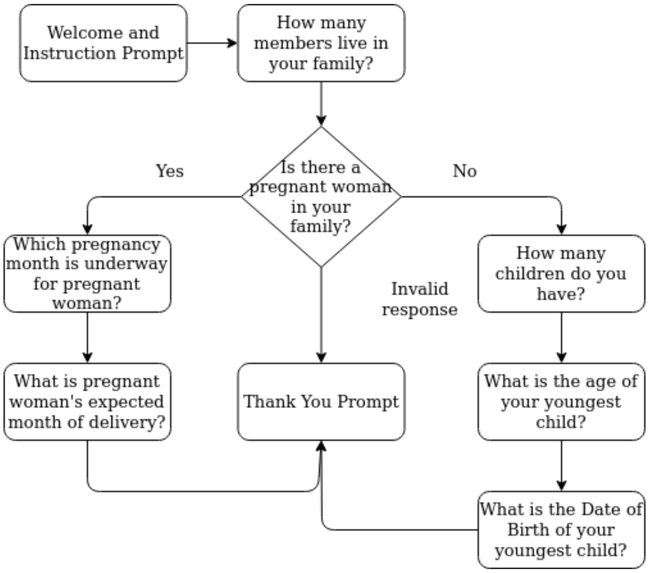

Welcome and Instruction Prompt
How many members live in your family?
Is there a pregnant woman in your family?
Yes
No
Which pregnancy month is underway for pregnant woman?
How many children do you have?
Invalid response
What is pregnant woman's expected month of delivery?
Thank You Prompt
What is the age of your youngest child?
What is the Date of Birth of your youngest child?

Figure 1: Voice Survey questions and flow

The voice-based and equivalent keypress-based survey we designed was kept in line with the goal of JEEViKA Mobile Vaani, and sought details from the platform users about pregnant women and small children in their family, as shown in Figure 1. Given such data, pregnant women and young mothers can be pushed customized voice messages to inform them or send reminders for vital health checks, vaccination schedules, nutrition related advisories, and hygiene and sanitation practices. The benefit of such systems has been widely documented and governments have adopted them at a national scale [9, 12, 16]. We next describe related work, followed by a detailed description of the several rounds of design iterations and user feedback, and finally discuss the scope of conducting voice-based surveys among rural and less-literate populations. User consent was sought for all activities, through audio prompts on IVR systems, or phone conversations for semi-structured interviews, or verbally during field meetings.

2 RELATED WORK

The choice of suitable modalities for taking user inputs on digital devices has been a long standing question. Early work has shown strong user preference to text-free [13] and voice-based interfaces [5] by less-literate users, although the accuracy of speech recognition for voice-based input was found to be a limitation. With improvements in speech recognition however, the widespread adoption of tools like Google Assistant on Android smartphones has more recently re-validated the usefulness of voice-based interfaces and demonstrated their feasibility for several tasks [21, 23, 27–29].

In this paper, we are concerned about data collection tasks through simple mobile phones, of whether keypress-based or voice-based inputs are more suitable. The usability and accuracy of using keypress-based inputs for data collection through IVR surveys has been compared against phone-based surveys by live human operators [4], and was found to be satisfactory. An early study [11] compared the two modalities of keypress-based and voice-based surveys, and found that keypresses performed better than voice inputs, due to a loss of accuracy with speech recognition. This study was conducted with working professionals, but other studies with less-literate users made similar observations [8, 17, 18]. Project HealthLine [25, 26] on the other hand concluded that a well-designed voice-based interface can significantly outperform keypress-based interface for both less-literate and high-literate users. More recent work benefiting from the wide availability of better speech recognition and natural language processing also found that voice-based interfaces should be considered as the main modality for less-literate users, while keypress-based interfaces can act as a fallback option if speech recognition fails or the user is not able to complete the task [19].

Our work is related, of furthering the comparison between the two modalities, but with a difference that most previous work has considered single-word voice responses whereas we focus on user inputs coming in the form of natural speech. The limitation of single-word responses has been acknowledged in previous work [17], as being new and difficult for users to learn. Our approach is therefore meant to be closer to methods involved with building chatbots or voicebots, where a natural interaction between the user and machine is expected: users can respond in natural speech, which the machine can process and respond back. This therefore involves speech recognition to convert the voice-based inputs to text, followed by entity extraction from the text using natural language processing methods. To the best of our knowledge, such a comparison between keypress-based inputs and natural voice-based inputs has not been done before, especially through IVR systems in the context of low-income and less-literate populations.

3 VOICE SURVEY DESIGN AND DEPLOYMENT

We followed an iterative process over four rounds of experiments to design voice-based surveys and compare them with keypress-based surveys. We started with an uncontrolled "before" study by creating a voice-based survey that users on the JEEViKA Mobile Vaani platform could voluntarily take, to obtain a broad understanding of the dominant issues likely to arise with voice-based surveys. This informed us to make several improvements, and Round 2 and Round 3 were then conducted as controlled experiments. These were coordinated on the ground by the Gram Vaani field team, who trained users to take the survey, and noted their observations in checklists provided by us. In the final round, we also contacted several users for semi-structured phone interviews to take their feedback. Due to COVID-19 restrictions and quarantine rules for movement across state boundaries, this method worked well where the field team took precautions while interacting in-person with the JEEViKA SHG members, and we were able to interact remotely to take feedback from the users and the field team. Improvements made throughout Rounds 1, 2, and 3 were implemented for a final "after" uncontrolled study in Round 4. During this round, an equivalent keypress-based survey was also administered to compare its pros and cons with those of voice-based surveys. All the rounds used the same survey questions shown in Figure 1. These questions were designed to span different common types of questions, as shown in table 1.

3.1 Implementation Details

To implement voice-based surveys, we integrated the Gram Vaani IVR stack with Google's Dialog Flow engine [7]. Voice recordings made on the IVR and available at the server-side, were streamed to Dialog Flow to obtain a transcript through the ASR (Automatic Speech Recognition) APIs of Dialog Flow. We refer to these as STT (Speech to Text) transcripts. Dialog Flow then provides a rich set of in-built functions to analyze the transcripts through natural language processing methods for entity extraction of data-types such as date, month, time, number, country, state, and district, among others. It can also understand indirect entity references like "tomorrow" and "2 days ago" for dates, and variations like "first", "second", "third"..."tenth" for numbers. Further, it allows customization of in-built entity extraction methods by specifying synonyms, new additions, and regular expressions that can be matched. Likely phrases in which the entities may be mentioned can also be specified. For example, phrases like "*my youngest child is N years old*" can be annotated to indicate the placement of a number entity *N*, or "*I live in city X*" can be annotated to indicate a location entity. The Dialog Flow documentation loosely refers to this process as "training the engine". We initially trained the Dialog Flow engine with 106 phrases and local entity names based on our familiarity with the deployment site, and as we explain later (in section 4.5), we continuously improved this by adding examples based on user responses that arrived through the various study rounds.

A 3-seconds silence detection was used on the IVR to detect the end of a user's response, while also giving sufficient time to the user to think about and speak their response after having heard a question. This value was determined through significant prior experience and internal tests with the Gram Vaani field team, to not have too aggressive a silence detection threshold for the user to miss out on saying what they want to say, and neither a large waiting time in advancing to the next question that could make the user-experience slow.

3.2 Round 1: Prototyping

The first round saw participation from 346 users of the JEEViKA Mobile Vaani platform. It was carried out in an entirely uncontrolled environment: the survey was announced to the users along with a disclaimer that it was meant to test a new technology feature, and users could volunteer to participate in the study. No specific training instructions were provided other than a welcome prompt which stated that users were supposed to speak their responses instead of pressing buttons on their phone. Each subsequent question prompt was carefully designed to be clear on what information was sought from the user, as shown in Table 2.

Figure 2a shows the success statistics in getting responses. The main issue that clearly arose was of blank responses: users seemed to be confused with this new concept of speaking their responses, and the IVR advanced to the next question after the silence detection duration. Upon listening to some of the responses and reading the corresponding STT output from Dialog Flow, we also noticed that users may speak local words that Dialog Flow may not recognize, or fail to recognize correctly. These pointers helped outline design questions that we proceeded to investigate in the subsequent study rounds.

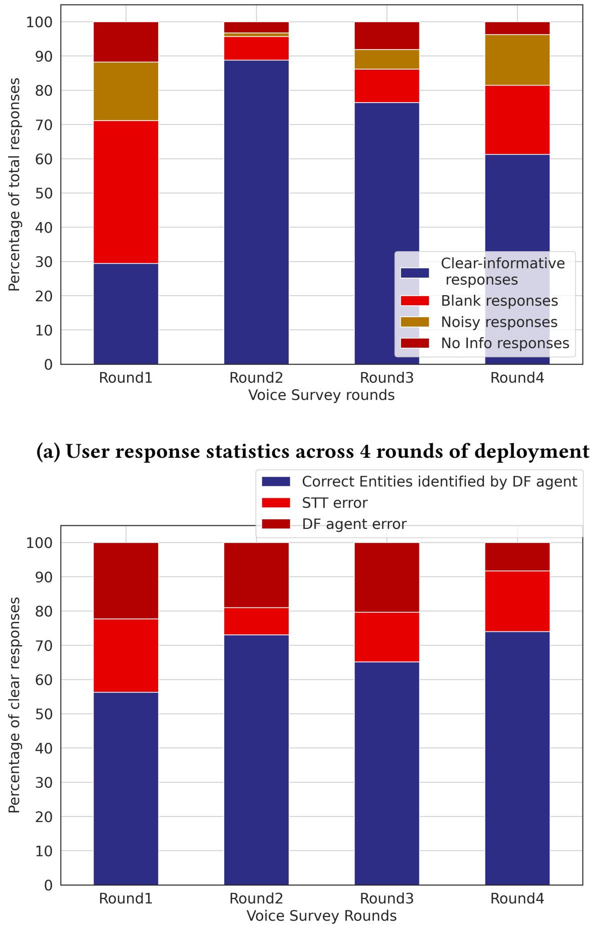
3.3 Round 2: Controlled Environment Study

Based on observations from the previous round, we improved the voice-based survey implementation in two ways. First, to reduce blank responses we introduced a loop on all the questions, to loop for up to two times if no entity was detected by Dialog Flow in the user responses. This detection was done in real time. Second, the content moderators used the responses from Round 1 to build a list of words that the Dialog Flow speech recognition was not detecting, or recognizing incorrectly. These were added as custom phrases to

Table 1: Taxonomy for voice-based survey question entities and keypress-based survey question types

S. No.	Question	Entity Type	Keypress typology
1.	How many members live in your family?	Number	Multi-option MCQ
2.	Is there a pregnant woman in your family?	Yes/No	Two-option MCQ
3.	Which pregnancy month is underway for the pregnant woman?	Number	Numeric
4.	What is pregnant woman's expected month of delivery?	Month(Date)	Multi-option MCQ
5.	How many children do you have?	Number	Numeric
6.	What is the age of your youngest child?	Age	Numeric
7.	What is the Date of Birth of your youngest child?	Date	Multi-level

improve the STT output of Dialog Flow, as also suggested in prior work to reduce speech recognition errors [2].



(a) User response statistics across 4 rounds of deployment

(b) Accuracy performance of Dialog Flow across 4 rounds of deployment

Figure 2: A round-by-round analysis of how users responded to the voice survey and how much information was Dialog Flow able to accurately identify.

This study round was carried out with the help of five field team members who were informed about the voice-based survey, and who in turn took consent and trained 195 JEEViKA SHG members to participate in the study. A training video was developed which helped the field team to understand this new survey methodology, as well as the video was used in their training sessions with the SHG members. We also provided the field team members with an observation checklist to record their observations when users were taking the survey.

As shown in Figure 2a, a 60% increase was seen in getting clear informative responses, as compared to Round 1. We surmise that this was due to a combination of user training as well as the looping introduced by us to repeat the questions if no entities were detected in the response. Figure 2b also shows a 16% decrease in STT errors, likely due to providing custom phrases for speech recognition.

Feedback from observations conducted by the field team members provided two additional insights. One, some users expressed dissatisfaction upon being asked the same question repeatedly, even though they felt that they had given the response:

"*Didi (elder sister) answered the question in her own local dialect which the system was not able to understand and then asked the question multiple times. Didi felt uncomfortable in giving the same answer twice*" — Sanjay from field team, Nalanda, Bihar.

Two, while the welcome prompt and the question prompts asked the users to speak their answer after the "beep", many users did not wait for the beep and started speaking prematurely:

"*A lot of Didis did not wait for the 'beep' tone and started answering before it*" — Santosh from field team, Muzaffarpur, Bihar.

We used this feedback in the next round to make some changes.

3.4 Round 3: Further Design Improvements

This study was carried out in a similar controlled setup as the previous round, to evaluate two changes. 333 users participated in this round.

3.4.1 How to Loop. Looping until "no entity found" as in the previous round was found annoying by the users if Dialog Flow's speech recognition or entity extraction gave an erroneous output. We changed this condition to loop until "empty STT", i.e. to repeat a question if its STT output came out empty. Our hope was that this would avoid entirely blank responses while still keeping it acceptable for users even if they spoke with a strong local accent or dialect that was not recognized by Dialog Flow.

As seen in Figure 2a, we found through usage log analysis that in Round 2 when looping on the more stringent condition of "no entity", 90% of the responses had been clear. This reduced to 77% in the "no STT" case in Round 3. The percentage of blank responses also increased, and as shown in Figure 2b, a smaller percentage of correct entities was identified as compared to Round 2. This was expected given the weaker condition for looping used in the new round. On the other hand, we found that user dropouts decreased in Round 3. Of all cases of dropouts when users disconnected before completing the entire survey, 72% of the dropouts happened when a question was repeated, indicating that users indeed found the question repetitions to be annoying. This however reduced to 40% in Round 3 with the new looping policy. Feedback from the field team also this time did not indicate any cases of users dissatisfied with question repetitions.

This trade-off suggests that voice-based surveys should start with looping on "No STT" because of the gentler user experience it provides, and move towards looping on "No Entity" when the speech recognition and entity extraction becomes stronger, possibly with providing custom phrases to improve the speech recognition results.

3.4.2 How to Beep. The field observations from Round 2, that some users would begin answering the question before the beep sound, motivated us to modify the prompt structure in Round 3. The same problem was noticed in prior work as well [19]. The authors experimented with removing the beep sound altogether, and reported that it did not lead to any significant improvement [19]. We therefore tried a different strategy, to remove the prompt ("*Please answer your question after beep*") altogether and instead just play a beep as an indication that the user could now give their response.

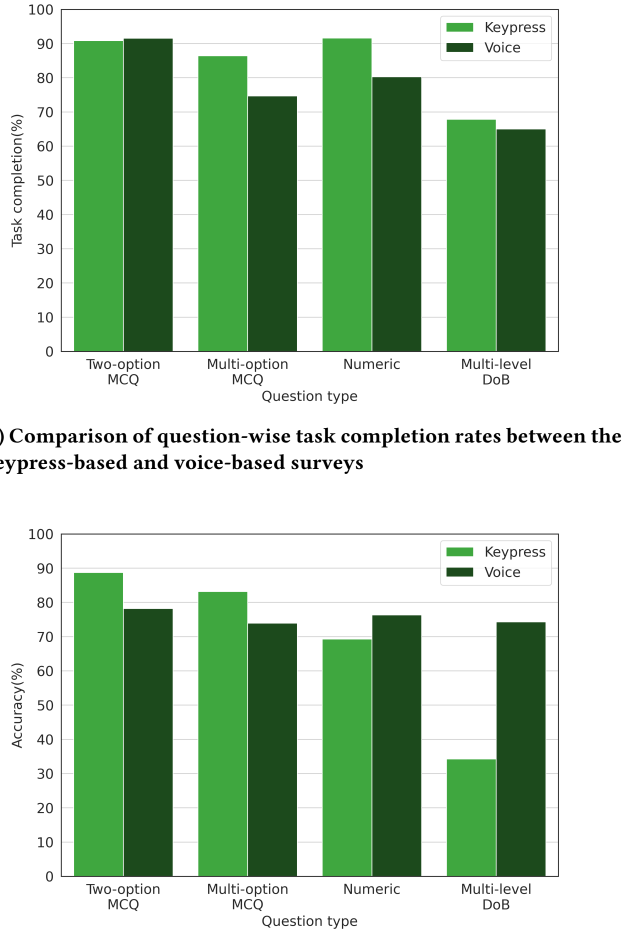
As observed through the usage analysis, blank responses only increased marginally from 7% in Round 2 to 10% in Round 3, with this change. The field feedback also did not suggest any significant changes: Several users would still prematurely start speaking. In the final round, we therefore decided to retain the structure of Round 3 to avoid repetition in the prompts, and noted that more improvement may come by creating tutorial audio clips and with repeated exposure of users to voice-based surveys.

We also sought feedback via the field team to confirm our choice of 3 seconds as the silence detection threshold. Most users did not report any issues and therefore we continued with it unchanged.

3.5 Round 4: Voice-based vs. Keypress-based Surveys

Based on the iterations made through insights gained from the controlled studies, we finally carried out a fourth round for the voice-survey, this time in an uncontrolled environment. Users calling into the main JEEViKA Mobile Vaani platform were given an option to participate in the survey.

Additionally, for comparison with a keypress-based survey, we prepared an equivalent survey that took only keypress inputs. Table 2 shows the corresponding prompts for the keypress-based survey. This included simple MCQs with two options, complex MCQs with five or more options, single-digit and multi-digit numeric questions, and a multi-level question for date of birth which took inputs separately for the year, month, and date of birth of the child.



(a) Comparison of question-wise task completion rates between the keypress-based and voice-based surveys

(b) Question-type wise accuracy comparison between the keypress-based and voice-based surveys

Figure 3: Task completion and accuracy results for keypress-based survey versus voice-based survey from Round 4

213 users opted to participate in the voice-based survey, among them only 7 users had received any form of training (from Round 2 or Round 3) for using voice-based surveys. Approximately ten days after having completed the survey, these same users were pushed the keypress-based survey as well and 154 users completed the additional survey. These responses were used to compare the efficacy of the two methods with each other.

We compare the modalities on two metrics: task completion and accuracy. Task completion for keypress-based questions was simply considered as whether or not a button was pressed, while for voice-based questions it was based on whether the user provided an informative answer. Content moderators heard the audio responses to determine if they contained the required information or not.

Table 2: Prompts and instructions for both voice-based and keypress-based surveys

Voice-based Survey	Keypress-based Survey
Welcome to this survey. Please speak your answer to the question after the beep tone.	Welcome to this survey. Please answer the questions by pressing phone keys.
How many members live in your family?	How many members live in your family? if 1 person press 1, if 2 people press 2, if 3 people press 3, if 4 people press 4, if 5 or more than 5 people press 5
Is there a pregnant woman in your family? Please respond by saying "yes" or "no"	Is there a pregnant woman in your family? If Yes press 1, if No press 2, if don’t know press 3
Which pregnancy month is underway for the pregnant woman? Please answer by saying a number between 1 to 9	Which pregnancy month is underway for the pregnant woman? Please press any phone key from 1-9 to answer this question. For example, if 4th month is going on then press 4 in the phone and if 8th month is going on then press 8.
According to the health care worker or Doctor, what is the pregnant woman’s expected month of delivery? Please answer by speaking the month name.	According to the health care worker or Doctor, what is the pregnant woman’s expected month of delivery? if this month in March press 1, is next month in April press 2, if in May press 3, if in June press 4, if July or later press 5
How many children do you have?	How many children do you have? Please answer by pressing the number on you phone. For example, if 2 children, then press 2, if 1 child then press 1 and similarly press the number for other responses
What is the age of your youngest child?	What is the age of your youngest child? Please answer by entering the age by pressing the phone keys. For example, if child is 1 year old, then press 1, if child is 10 years old then press ‘1’ and ‘0’. Similarly, press the keys to tell the age.
What is the Date of Birth of your youngest child? Please answer by speaking the day, month and year of birth like speaking 27th December 2020.	In which year was your youngest child born? Please answer by pressing phone-keys. For example, if the year is 2021 then press ‘2’, ‘0’ ‘2’, ‘1’. Similarly answer by pressing any other keys.
	In which month was your youngest child born? Please answer by pressing the number of any one of the 12 months in a year by pressing the phone keys. For example, if the child was born in March, then press 3 and if the child was born in December then press ’1’ and ’2’. Similarly, answer by pressing any other keys.
	On which date was your youngest child born? Please press a number among the 31 days in a month. For example, if the child was born on 4th then press ’4’ and if the child was born on 30th then press ‘3’ and ‘0’. Similarly, press phone key to give your answer.

For calculating the accuracy of keypress-based questions, since we had equivalent voice-based responses from the same users, we considered a keypress response as accurate if the response tallied with the audio response provided by the user. This ground-truth construction based on the audio responses was done manually by the content moderators. Accuracy for the voice-based responses was however determined based on whether the Dialog Flow entity extraction output tallied with the ground truth. This therefore allowed us to compare the different sources of inaccuracy in the keypress-based and voice-based surveys: Keypresses are easy to do but users could make mistakes while punching buttons, voice responses may also be easy to provide but can lead to mistakes due to speech recognition or entity extraction errors.

3.5.1 Results. Figure 3a shows the task completion rates for different question-types, through the keypress-based and voice-based

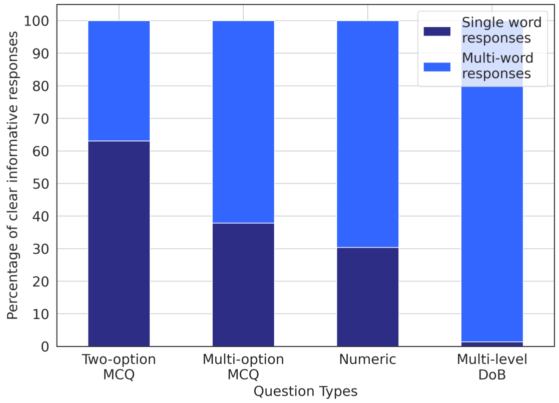


Figure 4: Analysis of how users respond to the various question-types

survey methods. Both modalities perform similarly, with keypress-based task completion being typically slightly better than voice-based task completion. The accuracy comparison shown in Figure 3b presents a more interesting picture. Users seem to be able to respond to MCQs more accurately through keypresses, while voice-inputs can suffer from speech and natural language processing errors. However users make mistakes when using keypresses for numeric inputs, both for single and multi-digit numeric questions, and for numeric questions that were a part of the multi-level question on date of birth. The accuracy of voice-based responses to these questions is much better.

We further analyzed all the voice responses to understand if users would just speak the specific entity word, or did they tend to speak entire sentences in their response. Figure 4 shows that the two-option MCQ question which simply asked for a yes/no response of whether the user had a pregnant woman in their family or not, did have a high percentage of single word utterances. However in all other question types, users did tend to speak in a natural manner, which was cited in earlier studies [17] as the main reason why voice-based input did not work as well as keypress-based input. In the current context however, with improvements in speech recognition and natural language processing, this problem seems to be solvable.

While these observations makes intuitive sense, we followed up with a user feedback exercise to seek qualitative information about their relative preferences for these modalities.

3.5.2 User Feedback. A qualitative study was carried out with 27 users in two groups, comprised of 14 women and 13 men respectively. The users were first informed about the study objectives and provided demonstrations of the surveys by the field team members. Each group was then divided into two sub-groups, one of whom took the voice-based survey first followed by the keypress-based survey, while the other was asked to take the surveys in the reverse order. The field team members also noted down their own observations and experiences. We then conducted a phone interview of the users and the field team members on the same day itself or on the following day, to take their feedback.

We observed a strong preference of both the users and the field team members towards voice-based surveys. All 5 of our field team members expressed that it was easier to conduct a training of the users for voice surveys. They reported that usually a single orientation was sufficient to help users understand about the modality, whereas the keypress-based survey required multiple rounds of demonstrations. Despite these multiple rounds, users faced difficulty with pressing the correct phone keys for many reasons. Their comfort with numeric literacy, being able to recollect information and at the same time think about which corresponding buttons to press, and many phones being in a poor condition with stuck keys or erased numbers, were some reasons observed by the field team. All the 27 users also gave similar feedback, that they found the voice-based surveys to be easier to understand. Their main concern with keypress-based surveys was the lengthy instructions that they had to listen to, and then follow:

"*Keypress-based survey contained very dense set of instructions and sometimes when a Didi (elder sister) is not literate enough, they get confused (on how to respond)*" — Female user, Nalanda, Bihar.

Many users added that in the keypress-based survey they unintentionally pressed some incorrect keys because their phones are old and the keypad is worn out. They in fact take help from others while making phone-calls, although they are able to receive calls on their own. These issues were more common among female users as compared to men.

We next took question-wise feedback, to understand if the relative preferences depended upon the different question-types. While the simple MCQ question with just two options was found to be straightforward on both modalities, several users reported that MCQs with more options can sometimes get confusing on keypress-based surveys:

"*My family has 8 members, but in the IVR the last option is '5 or more', I got stuck and could not understand what should I press*" — Santosh Kumar, field team.

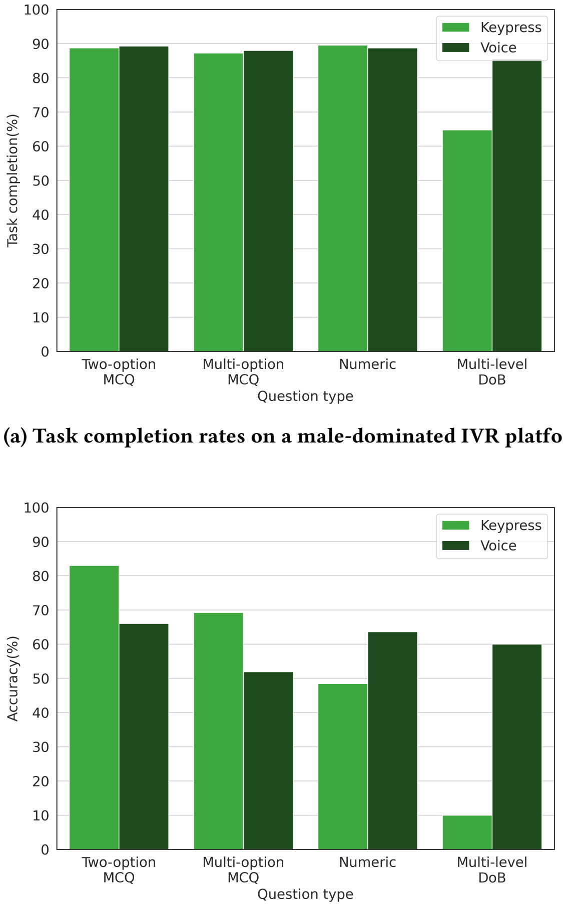
A programme manager experienced with creating keypress-based surveys on IVR systems similarly reported that it is often challenging to create such surveys and ensure that all possible responses can be covered within five or six choices only because anything more than that would be confusing for users to remember.

Users also found single-digit numeric questions to be easier than multi-digit numeric questions on keypress-based surveys because of literacy and phone condition related issues. All the users were more comfortable with voice-based responses for numeric questions.

80% of the users reported that they faced the most difficulty with the multi-level date of birth question. Many of them were not able to recollect the exact date, some knew the dates according to the Hindi calendar (which has months like Chaith, Baisakh, Jeth, . . . , Phagun) but not the English calendar, and the multi-digit responses required for the date, month, and year, made it even harder:

"*Those parents who have had their child 5-6 years back, are not able to remember the exact day, month and year. Moreover, it is absolutely not possible to remember these month names in English.*" — Male user, Munger, Bihar.

We can thus see that voice-based surveys are preferred by the users for most question-types. With further improvements in speech



(a) Task completion rates on a male-dominated IVR platform

(b) Accuracy of answers on a male-dominated IVR platform

Figure 5: Task completion and accuracy results for keypress-based survey versus voice-based survey from a different demography dominated by Male users

recognition and natural language processing capability, such voice-based surveys can potentially substitute keypress-based surveys for data collection.

4 DISCUSSION

Given some clear indications of the different contexts in which one of voice-based or keypress-based surveys should be preferred over the other, we next discuss some additional aspects. The current evaluation was conducted with women in rural areas, but do the results generalize to other demographics? Given the greater ease of use of voice-based surveys, how can the accuracy of voice-based surveys be improved? What operational processes would deployment managers need to consider to use voice-based surveys?

4.1 Different Demography

As part of our Round 4 iteration, we conducted an identical study with users of another Mobile Vaani platform that is predominantly used by men from the same region of Bihar. These platforms have been in use since several years, for local news, improving grievance redressal and social accountability in government schemes, agricultural information, etc [14, 24]. The heavy tilt towards a male userbase, and exposure to IVR platforms for a much longer duration, provide a distinct ground for evaluation as compared to the women users of JEEViKA Mobile Vaani.

225 users opted to participate in the voice-based survey, in an uncontrolled environment. These users were then called-back and invited to participate in an equivalent keypress-based survey, to which 188 users responded. As before, we evaluated for these users the task completion rates for different types of questions, and the accuracy of the answers. Figure 5a shows the task-completion rates and figure 5b shows the accuracy evaluation.

The trend is similar to that seen with JEEViKA Mobile Vaani users. Users were is able to answer numeric and multi-level question-types more accurately through voice, while the keypress-based survey worked better for MCQ questions. The high task-completion rates for voice-based surveys is noteworthy given that these users had not been exposed to such a modality earlier, although they are likely to have participated in keypress-based surveys that are run regular on Mobile Vaani for various purposes to seek feedback from the users.

4.2 Hybrid Modality

The distinct preference of keypress-based input for some question-types, and voice-based input for some other question-types, indicates that a hybrid modality may be suitable for data collection. We received similar feedback from Gram Vaani programme managers:

"*After conducting all these rounds of surveys, we strongly feel a mix of both modalities would be good for data collection.*" — Programme Manager, Gram Vaani, Gurgaon.

A hybrid approach however may become confusing to users. We plan to evaluate this carefully in the future.

4.3 Location Input

An additional question-type often required is for location input. Several methods have been used with keypress-based surveys, such as MCQs when the number of options are small and known in advance, or a multi-level sequence of MCQs to narrow down from taking input for a state followed by a district in that state and then a sub-district in the district, or through a multi-digit input for pin-code[1] [19]. As part of another survey conducted by Gram Vaani to understand the level of COVID-19 awareness among rural communities, we added a question about the location of users and experimented with comparing these modalities against taking a voice-based input for location.

In one setup of the survey conducted within the district of Nalanda, we stated the location question as an MCQ with options for five blocks in the district where the platform was thought to be most popular: "*Where do you stay? For Harnaut press 1, for Sarmera press 2, for Noorsarai press 3, for Motipur press 4, for Mushahri press 5, for others press 6*". On a different state-level platform in the state

[1] India and Pakistan have numeric zip-codes, unlike many countries that have alpha-numeric zip-codes.

of Bihar, we formed another MCQ with options for different districts: *For Gidhaur press 1, for Aliganj press 2, for Chakai press 3, for Jamalpur press 4, for Kharagpur press 5, for Others press 6.* In another setup of the survey on different platforms in use in some specific districts, we asked for the pin-code as a multi-digit numeric input: "*Please type the PIN Code of where you live. For example, if you live in Nalanda's Chandi block then press Chandi's PIN Code which is '8' '0' '3' '1' '0' '8'*". For each of these surveys, we also prepared an equivalent survey where the location input was taken in voice: "*Where do you stay? After the sound of beep, please tell the name of your State and your District*". In an A/B testing format, users were randomly given one or the other type of survey.

The task completion rates for these question-types are shown in figure 6. The results are as expected, that if it is feasible to keep the number of options small then users are most comfortable with MCQs. However, 62% of the respondents selected the *Others* option with the MCQ questions, indicating that our assumption about the geographic popularity of some of these platforms was quite incorrect and the MCQ option should be used with greater care. In this case, since the state of Bihar has 30+ districts, and each district typically has more than 10 blocks, even converting the questions into multi-level MCQs may not be feasible. PIN codes are generic but hard to provide, both because of the multi-digit modality which is challenging for people, and also as pointed out by the Gram Vaani field teams that many people do not know or remember their pin codes. A voice-based input is most convenient in this case.

The main challenge with voice-based questions as we have however seen, is the accuracy of obtaining a good STT, followed by the accuracy for entity extraction. We discuss next about the need for custom entity extraction modules to improve the performance over what Dialog Flow is able to provide even despite a large number of training examples.

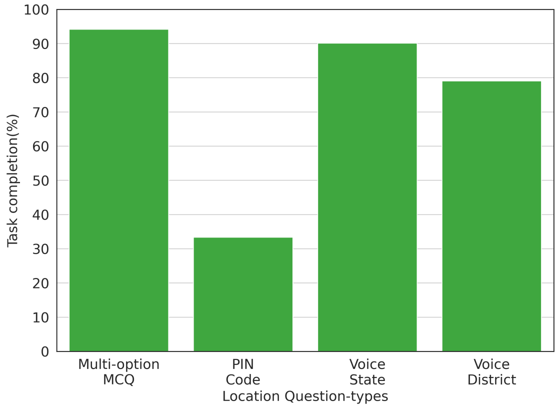


Figure 6: Task completion results for the location question, when asked as different question-types

Table 3: Accuracy comparison between DF and custom modules using good STT

Entity		Bad STT (%)	Accuracy of DF with good STT (%)	Accuracy of custom modules with good STT (%)
Location	State	28.71	51.81	87.78
	District		51.15	75.57
Date of birth	Date	26.93	78.28	95.95
	Month		77.77	87.87
	Year		72.22	92.42
Number		22.65	90.37	95.72

4.4 Accuracy Improvement in Voice-based surveys

As mentioned in the earlier sections, voice-based surveys may suffer from sources of inaccuracy for two reasons. First, the STT output from ASR APIs can some times be poor if users speak with a strong local accent or in a different dialect, or the recording produced through IVR systems is noisy. Second, NLP methods to extract entities from the STT may fail, especially if local phrases are used on which the ASR engines may not be trained. The first column in Table 3 shows the percentage of voice inputs recorded during the multiple rounds of surveys that had a poor STT, determined by the moderators as cases where the audio contained sufficient information but the STT output was incomprehensible even for humans to be able to extract any useful information. We can see that this ranges from anywhere between 20% to almost 30%. The next column shows that for cases where the STT was of a good quality, what was the entity extraction accuracy of Dialog Flow. This is respectable, at a range of 90% and 75% for numeric and date inputs in voice. It is however only 50% for location inputs, even when the Dialog Flow module was provided with a list of Indian states and districts as per the 2011 population census in India.

4.4.1 *Custom Modules for Entity Extraction.*

To improve STT processing for entity extraction, we built three custom modules for location, date of birth, and number entities, respectively. The location module uses a library for multilingual text processing called polyglot which identifies location entities based on the sentence structure [1]. Location entities thus identified are compared through string and phonetic matching with the list of Indian states and districts according to the Indian Census. The Date of Birth module uses several rule-based heuristics for sentence parsing based on how users speak the information, followed by string matching, to identify the date, month and year. We also took care of matching the STT against Hindi month names. The number entity extraction uses a Parts of Speech tagger available in the Stanza library for Indian languages [20], followed by string matching to identify the numbers. The last column of Table 3 shows the accuracy improvements achieved through these custom modules. The performance is substantially better than Dialog Flow for especially location input, and we are currently integrating these modules into the Gram Vaani technology stack.

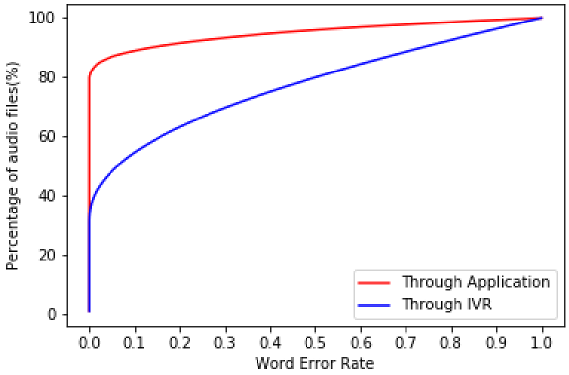


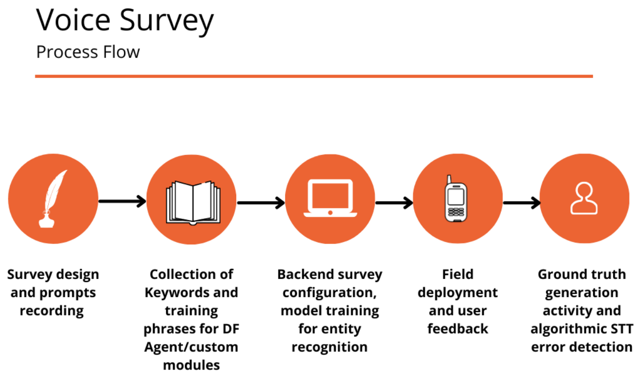
Figure 7: CDF for Word Error Rate of transcripts

4.4.2 STT Transcription Quality. While the custom entity extraction modules can be useful, they can only work post the availability of good STT. We wanted to evaluate if the 20% to 30% errors noticed due to a poor STT output, are due to a poor audio recording quality on IVR systems, or due to deficiencies with ASR performance of speech recognition engines. As part of ongoing work to replicate IVR functionality on a mobile application, we prepared an Android application for voice-based surveys. The application can load a survey structure specified in a custom JSON format, read out the questions through Google's Text to Speech conversion available on Android phones, accept an audio input and then process the STT for the recorded audio.

The field team facilitated a download of this application among 25 users in the JEEViKA ecosystem who had smartphones, from whom we received 175 responses across the various questions. Figure 7 shows a CDF of the WER (Word Error Rate) for these responses, compared with the WER for voice inputs recorded during the various rounds on the IVR. Against 70% of the IVR recordings that had a WER of 30% or less, 90% of the recordings done on a smartphone had a similar or less WER. Our findings corroborate with the results from earlier studies [27] and clearly show strong promise in the use of voice-based surveys on smartphones. The code for our custom modules for entity extraction and the Android application is available online[2].

4.5 Operational Overheads

Despite the challenges of inaccuracies stemming from STT, our study has shown that voice-based surveys hold strong promise in easing the collection even of complex data from less-literate populations. We want to however outline some components of operational overhead that are essential to realize the benefits of voice-based surveys. Figure 8 describes a process flow required to deploy a voice survey. Starting with the survey design, the first step is for experts and programme teams to provide an initial set of expected keywords or phrases in the responses, to train Dialog Flow or to create or modify custom entity extraction modules. This step is essential for complex surveys with conditional branching.



Voice Survey
Process Flow
Survey design and prompts recording
Collection of Keywords and training phrases for DF Agent/custom modules
Backend survey configuration, model training for entity recognition
Field deployment and user feedback
Ground truth generation activity and algorithmic STT error detection

Figure 8: Voice survey flow with a continuous ground truth generation activity

The user responses then need to be constantly monitored to spot new words or phrases that should be modeled to improve entity extraction. Strong internal communication between the programme, moderation, and technology teams is required at this point, so that model improvements can be documented, prototyped, evaluated, and finally incorporated into the voice survey. The survey can be publicized in stages for this purpose, first unrolling it within a small controlled group of users or volunteers, followed by a larger scale rollout. We feel that such an iterative process will need to be a part of any voice-based survey, adding overhead and making it expensive for smaller groups to conduct such surveys.

Further, systematic errors in the STT may also be noticed and providing custom keywords to ASR engines can improve their STT output [2]. We included in our process a step of continuous word to word transcription done manually by the moderators, for a sample of the audio inputs. This output is compared with the ASR generated STT through a dynamic programming based algorithm, which is used to identify words that had been mis-spelt in the STT, missed out, or incorrectly inserted. This list was put up to the moderators and helped them to suggest custom keywords that could be added to the Google ASR engine used by us.

5 CONCLUSIONS

Our study revealed strong user preference for voice-based surveys as compared to keypress-based surveys on IVR systems. With the current capabilities of speech recognition and natural language processing functionality available with commercial platforms such as Google's Dialog Flow, and further improvements possible through additional tools, we found that voice-based surveys are able to give comparable performance on task completion and accuracy as keypress-based surveys. Specifically with numeric questions, complex multi-level questions that are comprised of several sub-parts, and location inputs, voice-based surveys are easier for users to handle and are able to provide better task completion and accuracy. We also provided several insights on the process to operate voice-based

[2] https://github.com/ICTD-IITD/Voice_App_Custom_Entity_Extraction.git

surveys, and nuances observed through usage by users from among rural women in India. This study is likely to be helpful for other researchers and practitioners working in similar settings.

ACKNOWLEDGMENTS

We would like to thank the entire Gram Vaani team without whom this work would not have been possible. We would also like to thank the entire field team who helped conduct the field training and meticulously provided us with daily feedback. We would also like to thank Bill & Melinda Gates Foundation for funding this work and the JEEViKA team in Bihar for their cooperation to conduct the studies. Finally, we would like to thank the anonymous reviewers for their valuable feedback, which helped us improve the paper.

REFERENCES

[1] Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed Word Representations for Multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Sofia, Bulgaria, 183–192. http://www.aclweb.org/anthology/W13-3520

[2] Pranav Bhagat, Sachin Kumar Prajapati, and Aaditeshwar Seth. 2020. Initial Lessons from Building an IVR-based Automated Question-Answering System. In *Proceedings of the 2020 International Conference on Information and Communication Technologies and Development*. 1–5.

[3] Dipanjan Chakraborty, Akshay Gupta, and Aaditeshwar Seth. 2019. Experiences from a mobile-based behaviour change campaign on maternal and child nutrition in rural India. In *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development*. 1–11.

[4] Dipanjan Chakraborty, Indrani Medhi, Edward Cutrell, and William Thies. 2013. Man versus machine: evaluating IVR versus a live operator for phone surveys in India. In *Proceedings of the 3rd ACM Symposium on Computing for Development*. 1–9.

[5] Sebastien Cuendet, Indrani Medhi, Kalika Bali, and Edward Cutrell. 2013. VideoKheti: Making video content accessible to low-literate and novice users. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2833–2842.

[6] Google. 2020. *Ok Google: How is voice making technology more accessible in India?* https://www.thinkwithgoogle.com/intl/en-apac/country/india/ok-google-how-is-voice-making-technology-more-accessible-in-india/

[7] Google. 2021. *Dialog Flow*. https://cloud.google.com/dialogflow

[8] Aditi Sharma Grover, Madelaine Plauché, Etienne Barnard, and Christiaan Kuun. 2009. HIV health information access using spoken dialogue systems: Touchtone vs. speech. In *2009 International Conference on Information and Communication Technologies and Development (ICTD)*. IEEE, 95–107.

[9] Aparna Hegde and Riddhi Doshi. 2016. Assessing the Impact of Mobile-based Intervention on Health Literacy among Pregnant Women in Urban India.. In *AMIA*.

[10] Louis H Janda, Michael Janda, and Eric Tedford. 2001. IVR Test & Survey: a computer program to collect data via computerized telephonic applications. *Behavior Research Methods, Instruments, & Computers* 33, 4 (2001), 513–516.

[11] Kwan Min Lee and Jennifer Lai. 2005. Speech versus touch: A comparative study of the use of speech and DTMF keypad for navigation. *International Journal of Human-Computer Interaction* 19, 3 (2005), 343–360.

[12] Amnesty LeFevre, Smisha Agarwal, Sara Chamberlain, Kerry Scott, Anna Godfrey, Rakesh Chandra, Aditya Singh, Neha Shah, Diva Dhar, Alain Labrique, et al. 2019. Are stage-based health information messages effective and good value for money in improving maternal newborn and child health outcomes in India? Protocol for an individually randomized controlled trial. *Trials* 20, 1 (2019), 1–12.

[13] Indrani Medhi, Aman Sagar, and Kentaro Toyama. 2006. Text-free user interfaces for illiterate and semi-literate users. In *2006 international conference on information and communication technologies and development*. IEEE, 72–82.

[14] Aparna Moitra, Vishnupriya Das, Gram Vaani, Archna Kumar, and Aaditeshwar Seth. 2016. Design lessons from creating a mobile-based community media platform in Rural India. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*. 1–11.

[15] Preeti Mudliar, Jonathan Donner, and William Thies. 2012. Emergent practices around CGNet Swara, voice forum for citizen journalism in rural India. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*. 159–168.

[16] Nirmala Murthy, Subhashini Chandrasekharan, Muthu Perumal Prakash, Aakash Ganju, Joanne Peter, Nadi Kaonga, and Patricia Mechael. 2020. Effects of an mHealth voice message service (mMitra) on maternal health knowledge and practices of low-income women in India: findings from a pseudo-randomized controlled trial. *BMC Public Health* 20 (2020), 1–10.

[17] Neil Patel, Sheetal Agarwal, Nitendra Rajput, Amit Nanavati, Paresh Dave, and Tapan S Parikh. 2009. A comparative study of speech and dialed input voice interfaces in rural India. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 51–54.

[18] Neil Patel, Deepti Chittamuru, Anupam Jain, Paresh Dave, and Tapan S Parikh. 2010. Avaaj otalo: a field study of an interactive voice forum for small farmers in rural india. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 733–742.

[19] Muhammad Qasim, Haris Bin Zia, Awais Athar, Tania Habib, and Agha Ali Raza. 2021. Personalized weather information for low-literate farmers using multimodal dialog systems. *International Journal of Speech Technology* (2021), 1–17.

[20] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082* (2020).

[21] Shan M Randhawa, Tallal Ahmad, Jay Chen, and Agha Ali Raza. 2021. Karamad: A Voice-based Crowdsourcing Platform for Underserved Populations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[22] Agha Ali Raza, Mansoor Pervaiz, Christina Milo, Samia Razaq, Guy Alster, Jahanzeb Sherwani, Umar Saif, and Roni Rosenfeld. 2012. Viral entertainment as a vehicle for disseminating speech-based services to low-literate users. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*. 350–359.

[23] Johan Schalkwyk, Doug Beeferman, Françoise Beaufays, Bill Byrne, Ciprian Chelba, Mike Cohen, Maryam Kamvar, and Brian Strope. 2010. "your word is my command": Google search by voice: A case study. In *Advances in speech recognition*. Springer, 61–90.

[24] A Seth, A Gupta, A Moitra, D Kumar, D Chakraborty, L Enoch, O Ruthven, P Panjal, RA Siddiqi, R Singh, et al. 2020. Reflections from Practical Experiences of Managing Participatory Media Platforms for Development. In *Proceedings of the 2020 International Conference on Information and Communication Technologies and Development*. 1–15.

[25] Jahanzeb Sherwani, Nosheen Ali, Sarwat Mirza, Anjum Fatma, Yousuf Memon, Mehtab Karim, Rahul Tongia, and Roni Rosenfeld. 2007. Healthline: Speech-based access to health information by low-literate users. In *2007 International Conference on Information and Communication Technologies and Development*. IEEE, 1–9.

[26] Jahanzeb Sherwani, Sooraj Palijo, Sarwat Mirza, Tanveer Ahmed, Nosheen Ali, and Roni Rosenfeld. 2009. Speech vs. touch-tone: Telephony interfaces for information access by low literate users. In *2009 International Conference on Information and Communication Technologies and Development (ICTD)*. IEEE, 447–457.

[27] Aditya Vashistha, Abhinav Garg, and Richard Anderson. 2019. ReCall: Crowdsourcing on basic phones to financially sustain voice forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[28] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2017. Respeak: A voice-based, crowd-powered speech transcription system. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 1855–1866.

[29] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2018. BSpeak: An accessible voice-based crowdsourcing marketplace for low-income blind people. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.