Gurgaon Idol: A singing competition over Community Radio and IVRS

Zahir Koradia[§]
IIT Bombay
Mumbai, India

Piyush Aggarwal
IIT Delhi
New Delhi, India

Gaurav Luthra
IIT Delhi[♣]
New Delhi, India

Aaditeshwar Seth[♣]
Gram Vaani[§]
New Delhi, India

ABSTRACT

In this paper, we describe several IVR usage and learnability insights that emerged from a singing competition held by a community radio station located in an urban community of low-income migrant workers. Our community radio station partner, Gurgaon Ki Awaaz, relies heavily on folk songs to build its content repository and develop a close rapport with its community; the station organized a competition called Gurgaon Idol, in which community members could call into an IVR system to record their songs, and vote to select the best songs. Our research yielded several insightful results on how to best solicit audio recordings on IVR, methods for crowdsourced voting on IVR, cultural preferences towards certain voting methods, how to help first-time IVR users learn the system, and practical tips to keep in mind when running such a competition. To the best of our knowledge, we are the first to explore usability of voice user interfaces for recording audio and for crowdsourced voting over IVR systems.

1. INTRODUCTION

Community Radio (CR) stations, small range FM stations typically run by not-for-profit organizations, are heavily leveraging the growing penetration of mobile phones to engage with their listeners [23, 8]. Stations run several activities such as soliciting questions from their listeners, requests for songs and specific programs, running quizes, surveys, etc. In one such experiment, an urban CR station located in northern India, Gurgaon Ki Aawaz, organized a singing competition using IVR systems to engage with young listeners. We worked with the station to design and execute this competition. Our work yielding several interesting insights into the use of IVR systems for recording audio, crowdsourced voting, cultural implications on voting methods, and helping novice IVR users learn to use these systems. These insights form the subject of this paper.

DEV'13, January 11-12, 2013. Bangalore India

We begin with describing the context in which Gurgaon Ki Aawaz operates and their motivations for holding this competition. We then present an overview of the Gurgaon Idol competition, and specific IVR related research questions we are interested in answering. Section 4 describes in detail several usability tests we ran with over 80 subjects, and Section 5 describes the actual competition as it unfolded on the ground. Finally, a discussion section is presented on our reflections from running the competition.

2. THE CONTEXT

Gurgaon Ki Aawaz (GKA) is a community radio station based in Gurgaon, a city in northern India. The primary listener base of the station are migrant workers who have moved to Gurgaon from several states, and even from the neighboring country of Nepal. The station makes programs on women's issues, child health, micro-entrepreneurship, and civic issues, among others. Being a radio station, GKA strives to deliver its content in an entertaining manner to keep its listeners engaged. Folk music and songs especially crafted as a creative commentary on current affairs forms a significant portion of GKA's broadcast. Much of this content is recorded by members from the station's community itself. The station also uses several technologies to engage with its listeners over phone calls and SMS to solicit program requests and opinion, which are also put on air. However, the station has a constantly moving target to increase its repository of folk song content, and improve and widen the base of its engagement with the community.

The culture of music at the station: Gurgaon has a rich culture of encouraging local folk music and musicians. *Raginis*, descriptions of mythological events in song form, and *Sangs*, dramatization of mythological stories in the form of a musical, are two of the most popular forms of local music. While artists are invited to perform *Sangs* in family events, *Ragini* competitions are more common and are frequently held to celebrate events like child birth, religious festivals, and other social events with a generous remuneration being earned by the performers in both cases.

GKA has brought this rich culture of folk music to radio. Of the 22 hours that it broadcasts in a day, roughly 14 hours of the content is local folk music, and 1 hour is non-local folk music. A two hour long song request program called *Apni Pasand*, roughly translated in Hindi as *My Request*, is aired every afternoon and repeated later at night. *Bhajans* (devotional songs) are aired for one hour every morning. In

addition, topical programs on health and employment are also accompanied with songs on a related theme.

Gurgaon Ki Aawaz needs to have a significant repository of songs to be able to air such a large amount of folk music on a daily basis. It has evolved several innovative strategies to do this. The station staff regularly visit Ragini competitions, Sang performances, and various local music events, to record the songs and later air them on the station. Because the station has established itself as a promoter of folk music, it now also regularly receives invitations to such events. Other methods of sourcing music include inviting musicians to the station for higher quality studio recordings, soliciting pre-recorded music from local professional singers, and visiting musicians' homes for recordings. GKA now has a bank of more than 800 folk songs collected over a period of three years. In spite of this collection, obtaining new songs remains a continuous challenge for the station as its listeners continue to request for songs that the station does not have.

Community representation at the station: One of the core principles of community radio is ensuring community participation in content creation. At GKA this principle manifests in the form of receiving calls from listeners and putting their voices on air. The station uses a variety of methods to achieve this: It uses (a) GRINS [7, 9], a radio automation system, to put calls live on air, (b) PhonePeti [10], an answering machine system, to receive feedback 24 hours a day, and (c) a dicta-phone to record received calls on their office phone. Across all these methods, the station receives more than 50 calls a day.

Although the number of calls received are admirable, a closer look at the demography of callers reveals a bias: according to the station, most of the calls are received from men aged greater than 30. The station has attempted to even out this bias through explicit requests for women and youth participation and relevant radio programs but with limited success.

To address the above challenges to enhance the bank of songs and singers, and to encourage the youth and women to build a relationship with the station, we worked with GKA to design and execute a singing competition called Gurgaon Idol. We next describe the design of the competition.

3. COMPETITION DESIGN

To enhance the song bank and participant plurality, and to create excitement about the competition, we wanted it to be easily accessible for participation by listeners and to involve them in judging the best songs. We therefore designed the competition as a two phase event conducted over radio, IVR, and SMS:

3.1 Phase 1 - Participation

Radio promotions were aired to encourage participation, and listeners could take part in the competition by simply calling into an IVR system and recording their name, age, and the song they wanted to enlist. To encourage participation from the youth, the participants were divided into two groups: (a) those aged below 30 and (b) those aged above 30[1]. Participants could also call repeatedly to re-record their songs in case they were not satisfied with their previous recordings. We used the phone numbers as an identity for the singers to allow them to revise their entries.

[1]We had also divided each of the two groups into male and female, but merged them later due to lack of enough female participants

3.2 Phase 2 - Voting

In the second phase, all the songs were assigned an entry number and considered for voting. The songs with their entry numbers were aired on the radio to popularize them, and listeners were encouraged to vote through SMS and IVR systems set up by us. The songs were to be judged exclusively through these votes, thus ensuring that winners were chosen by the community. Two winners from each group were chosen at the end of the voting. The winners were invited to the station and several songs of each winner were recorded in the studio, which the winners were free to sell.

3.3 Research questions

Through the competition, we wanted to answer several questions about the usage of IVR systems in low-income and poorly literate communities. The questions on which we focused are as follows:

- **Audio recording method:** In earlier work [10] the authors noticed that when callers were asked to record their message after "the beep", many of them did not understand the instruction. In this paper, we explore two options to prompt a caller to record their name, age, and song: (a) the commonly used "record after the beep", which we call *Beep Voice User Interface (VUI)* and (b) "press a button and then start recording", which we call *Button VUI*. Our intuition was that the second method would be easier and give time to the caller to prepare herself to start recording. Insights from this study would apply to other IVR systems as well that ask callers to leave messages.

- **Song voting method:** What is a *good* way to allow callers to vote on songs? We consider three different voting methods: (a) *Thumbs-up/Thumbs-down* - one song at a time is played to the caller and she is asked to vote the song up or down, (b) *Best of two* - two songs are played to the caller and she is asked to choose the better of the two, and (c) *Best of four* - caller chooses the best of four songs played to her. Insights from this study would be useful in other crowdsourced ranking and rating IVR systems as well.

- **Learning to use IVR systems:** Given that most listeners are unlikely to have had past exposure to IVR systems, how can the radio station train listeners in using IVR systems? We consider learning through (a) on air instructions, (b) repeated calls, (c) instructions over a phone call, and (d) in-person handholding. Insights from this study will apply broadly to any IVR system on how to efficiently train new users to interact with the automated system.

We try to answer these questions by first conducting detailed usability tests in a controlled environment, and then obtain macro-level statistics recorded in the actual competition.

4. USABILITY EXPERIMENTS

To answer the above questions, we conducted four usability studies with a total of 88 subjects chosen from the

catchment area of Gurgaon Ki Aawaz. We first describe demographic characteristics of the subjects, followed by the experiment environment, and then provide detailed experiment design and results of each of the four tests.

4.1 Subjects

Our subjects were drawn from the listener base of GKA so that the sample mimics, as much as possible, the actual listeners of the station. Due to the lack of availability of listener demographics, we chose our subjects across a variety of age, gender, education and occupational backgrounds. Figure 1 shows histograms corresponding to the ages of 41 male and 47 female subjects, and highlights the even distribution of subjects across ages. Figure 2 shows the educational qualification of the subjects that seems to be concentrated around class 10 and class 12, and is evenly distributed otherwise. Some of the occupations reported by the subjects were student, house wife, electrician, security guard, and driver. All except five subjects used Hindi as their primary language of communication. 35 subjects reported to have had prior exposure to IVR systems at customer service numbers of cellular providers, but only 9 subjects had used a record-a-message like IVR system such as PhonePeti [10]. Thus, our subjects covered a wide demography range of low-income urban citizens who were new to the kind of IVR systems on which we wanted to evaluate them. Appropriate caution should be taken on the generalizability of our results to a demography outside the one we evaluated.

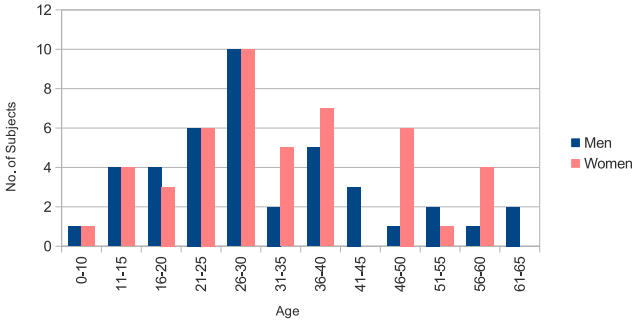


Figure 1: Age distribution of male and female subjects chosen for the experiments.

We conducted four different experiments with these subjects. To avoid confounding the results of the experiments with each other, no subject was used for more than one experiment. We next describe the environment in which the experiments were conducted.

4.2 Experiment environment

All experiments were conducted at the community radio station, to serve as a balance between the practical challenges of conducting the study in the subject's own environment Vs. the risk of impacting experiment results because of the environment being unknown to the user. A separate room was used for the evaluations with subjects invited one at a time to the room, while other subjects waited for their turn in another room.

A low-end color-screen phone was used to make calls into the IVR system. Enough time was provided to the subjects to get familiar with the phone.

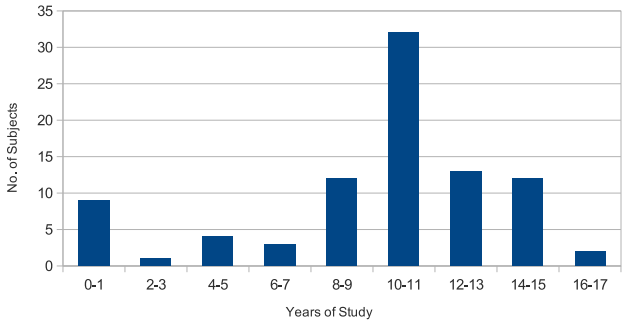


Figure 2: Distribution of number of years of study completed by the subjects chosen for the experiments. Examples: Class 10 pass = 10 years of study, and three year undergraduate program = 15 years of study.

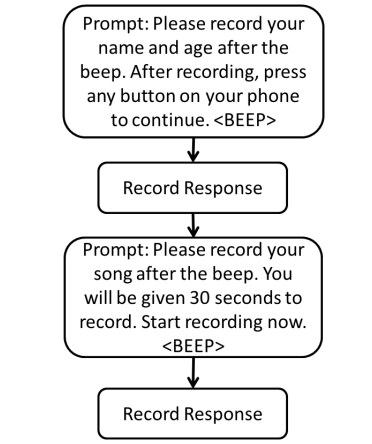
All subjects were given a token money of Rs. 100 and a GKA sticker to acknowledge that they had taken time off work and traveled at their own expense to the station to participate in the experiments. The subjects were informed of the token money before they arrived at the station, and that acted as an incentive for them to participate in the experiments. We next describe each experiment design in detail and present the corresponding results.

4.3 Audio recording methods

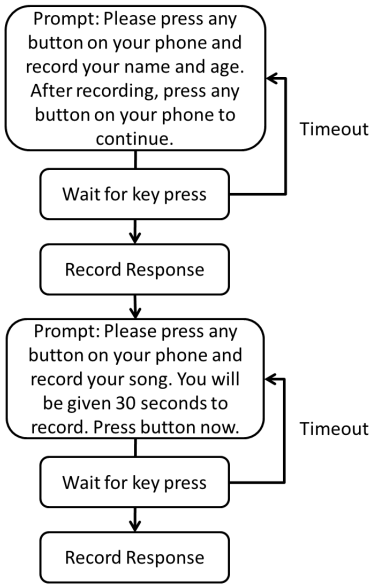
In PhonePeti, the authors observed that first-time IVR callers faced problems in being able to leave recorded messages, and hence we wanted to experiment with another voice user interface (VUI) for the participation phase of the competition to have callers record their songs. We considered two different VUIs: (a) *Beep VUI*, a commonly used interface in answering machine systems asking the caller to record their message after the beep, and (b) *Button VUI*, asking the caller to press a button before starting to record the message. If the caller does not press a button within a timeout, the instructions are repeated. Schematics of both these interfaces are shown in Figure 3. Both VUIs asked the caller to first record their name and age, and then record the song for the competition.

Seven subjects were first made to use the Beep VUI followed by the Button VUI, and four subjects were asked to use the to VUIs in reverse order[2]. Subjects were given a hypothetical situation where they had to participate in a singing competition by calling into a "computer operated service" and record their song. A promotional tutorial prepared by GKA with instructions on mock usage of the IVR was played out to each subject from a laptop to train them on how to use the IVR, and mimic the situation of hearing the promo on radio. Subjects were allowed to listen to the promo as many times as they wanted before actually

[2]Although we started with seven subjects for both the orderings, three subjects in the second grouping got called off during the experiment and we were unable to replace them since the subjects chosen for this experiment were all singers.



(a) Beep VUI



(b) Button VUI

Figure 3: Schematics of Beep and Button VUI

trying out the IVR system. Prompts and encouragements were then provided whenever the subjects were hesitant or confused while using the system. After a subject had used one VUI, she was asked about the difficulties they faced in understanding and using the VUI. After a subject had used both the VUIs she was asked to choose the VUI that was easier for her and her reasons to choose it. Next we present the learning obtained from this experiment:

- Subjects did not understand what was wanted when the IVR system asked them to "record" their name and age. They found "speak into the phone" as a more appropriate instruction.

- When asked to press any button (before or after the recording, depending on the VUI), subjects got confused about which specific button should they press on the phone. They preferred to be told to press a particular button so that they did not make a mistake.

- The promo in the form of a mock usage did not seem to help. Explicit instruction worked better. This was in fact reported as a need by one of the subjects. Five subjects could not differentiate between the two IVRs through the promos, further indicating that mock usage may not work very well as instructions. We verify this more strongly in the third experiment on IVR learnability.

- Subjects were often hesitant to press buttons as they were afraid of doing something wrong or breaking something. This is a well known challenge [13, 2], and it reappeared here in our context as well.

- Many subjects reported that they understood the instructions when they heard them, but during actual usage they forgot what they were supposed to do and were not able to act correctly. This is a common difficulty commonly associated with the "Memory for Arbitrary Things", where people try to remember arbitrary sets of steps to carry out a task instead of building a mental model of the task and logically try to remember the steps [15].

- When using the button VUI, many participants did not wait to listen to the complete instructions. They tried to press a button immediately after hearing "please press any button on your phone and record your name and age", and did not wait to hear "press the button again to stop recording".

The key learning to note are that words like "record" are hard to relate to for a first time IVR user; giving specific instructions instead of a free choice, for example when asking the caller to press any button to start recording, can reduce anxiety; and having people learn IVR systems by listening to mock-ups or pre-instructions is not as simple as it seems.

The above insights helped to improve the actual voice user interfaces presented to the participants in the competition. We next describe the experiments related to voting methods.

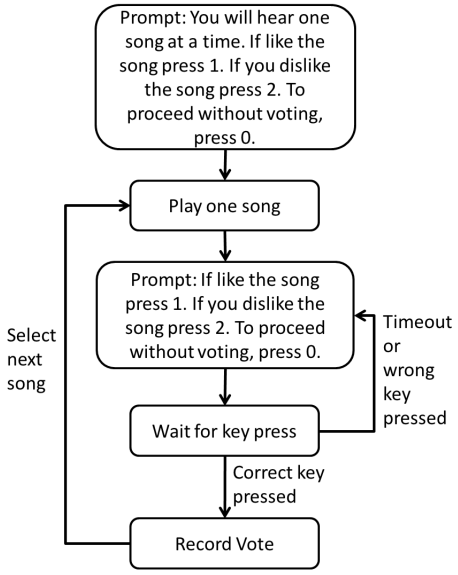
4.4 Voting methods

A common method widely used on the Internet for crowdsourced ranking is for users to do a thumbs-up/thumbs-down or provide a rating between 1-5 on each item. This method aims to arrive at an absolute score for each item, and assumes some degree of prior exposure of the items to the user for her to be able to rate them in a reasonable manner. In our case, we can develop a similar system by playing out songs to callers one at a time, and asking them to do rate up or down each song. Another common method for crowdsourced ranking is to present pairs of items to users and ask them to choose the better one. Several algorithms have then been proposed to aggregate these pair-wise rankings to arrive at a global order [4, 3, 5]. This is another interface with which we want to experiment, where each caller is presented a pair of items and asked to choose the better one. Optimization algorithms can even be applied so that the choice of pairs to be presented in each step is made intelligently to quickly arrive at the top-k songs (k=2 in our case). We experiment with both the thumbs-up/thumbs-down and the pairwise ranking methods.

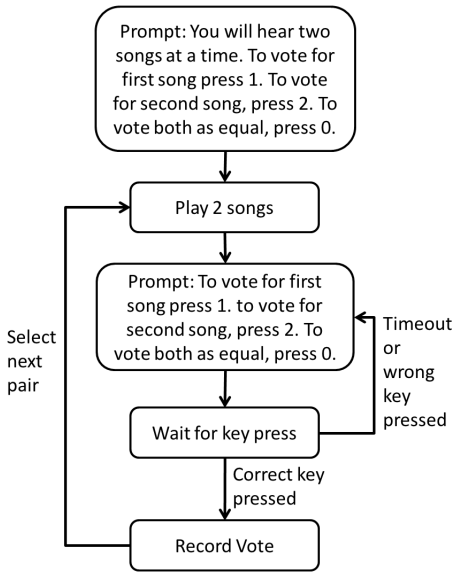
4.4.1 Thumbs-up/thumbs-down vs Best of two

We experimented with two setups: (a) In the *thumbs-up/thumbs-down (TUTD)* method one song was played to

the voter at a time and she was asked to vote the song *up*, *down*, or *neutral*, (b) in the *best of two* method, two songs were played to the voter and she was asked to choose the better one or choose both as equal. The schematics of both the voting methods are shown in Figure 4.



(a) TUTD VUI



(b) Best of Two VUI

Figure 4: Schematics of VUI of Thumbs Up Thumbs Down (TUTD) and Best of Two voting methods.

Nine subjects were asked to use the TUTD VUI followed by the Best of Two VUI and another eight subjects were asked to use the VUIs in reverse order. The experiment design was similar to the audio recording experiment: subjects were give a hypothetical situation where they had to vote in a competition, and an audio snippet giving detailed instructions on how to use the VUI was played through a laptop to mimic hearing the instructions on a radio program. Support was provided if a subject got stuck, and feedback was collected after the subject used each of the VUIs. Below we present the learning obtained from the experiment:

- In terms of usability, all but one subject were able to use both the VUIs. This one subject was not able to use either of the VUIs.
- Some subjects asked before pressing a button whether they should indeed press the button then. This again relates to the hesitation of breaking something or doing something wrong, also observed in the previous Beep and Button VUI experiments.
- There were several interesting reasons reported for preferring one voting method over the other:
  1. One subject preferred best-of-two VUI over TUTD as he felt uncomfortable in rating a song bad and considered it culturally impolite. He felt more comfortable stating that Song A is better than Song B.
  2. Another subject preferred best-of-two VUI over TUTD as he found it easier to compare two songs rather than rate a song up or down without knowing all the songs in the playing field.
  3. One participant preferred TUTD over best-of-two VUI as she did not feel comfortable choosing one song over the other. A closer look at her use of the TUTD VUI system showed that she had actually up-voted all the songs she heard. She had not chosen all songs as equal in her pairwise voting however, which may imply that pairwise rankings can provide more information when voters are not willing to rate songs as bad.
  4. Only two participants reported that pairwise ranking was harder to understand compared to TUTD voting.

Thus, even though both TUTD and best-of-two voting methods were found to be easy to use, there were several cultural and psychological factors that impacted the preference of one voting method over the other.

4.4.2 Best of Two vs Best of Four

One of the limitations of ranking algorithms that use pairwise preferences to arrive at a global order is that they require a large number of pairwise preferences to arrive at an ordering with sufficient confidence levels. For example, the Glicko rating method [4] recommends at least 5 pairwise preferences per song before a song's rating can be obtained accurately. Depending on the number of songs to be ranked, the number of pairwise preferences required can be quite large.

One way to obtain more information from a single vote is by asking the voter to choose the best of n songs. Each such vote then gives us $n-1$ pairwise preferences. However, as n is increased, usability of the IVR may decrease very quickly.
Usability: To explore the usability of such a voting method, we conducted an experiment where 10 subjects were asked to use the best-of-two VUI followed by a best-of-four VUI, and another 10 subjects were asked to use the VUIs in the reverse order. The rest of the experiment design was the same as before. *We found that all the subjects complained about the difficulty of voting in the best-of-four VUI, mainly*

because they could not remember the initial songs by the time they got to the last song.

Voting reliability: If the callers have difficulty in remembering the earlier songs in the best-of-four VUI, then it is likely that the rankings obtained through the best-of-four VUI may be different from those obtained through the best-of-two VUI. To test this hypothesis, we conducted an experiment where 10 subjects were asked to use the best-of-two VUI and vote on 4 songs (ie. 6 pairs), and another 10 subjects were asked to use the best-of-four VUI and vote on the same 4 songs. The subjects chosen for this experiment were graduate students who were comfortable with using IVR systems. We chose these subjects to rule out the possibility of IVR usability and novelty aspects in impacting the results. The songs chosen for voting were 30 second snippets of well known Hindi movie songs.

We used the Glicko ranking algorithm [4] to identify the best songs from the pairwise preferences of both the VUIs. This algorithm accepts preferences between pairs of songs and produces as output, a tuple (R_s, RD_s) for each song s. R_s is the rating, which indicates how good song s is: higher rating indicates better song. And RD_s is the *rating deviation*, which indicates the confidence the algorithm has in the rating it has given to s: lower *rating deviation* indicates higher confidence. Glicko uses a Gaussian distribution to represent a song's rating, with rating deviation representing the standard deviation of the distribution. Thus, for two songs s_1, s_2, if $R_{s1} + RD_{s1} < R_{s2} - RD_{s2}$ then we can safely conclude that s_2 was rated better than s_1 by the voters.

Table 1: Rankings of four songs based on votes cast by 10 subjects using the best-of-two VUI, another 10 subjects using the best-of-four VUI, and a third set of 10 subjects with song order reversed in the best-of-four VUI

	best-of-two		best-of-four		best-of-four (songs order reversed)	
	Rating	**Std-Dev**	**Rating**	**Std-Dev**	**Rating**	**Std-Dev**
Song 3	1835	81	1856	101	1800	105
Song 4	1719	81	1435	129	1557	121
Song 1	1180	79	1074	148	1074	148
Song 2	1290	79	1280	137	1280	137

Looking at the rating scores given by the Glicko algorithm in Table 1, we can see that both VUIs produced the same ranking, indicating that the best-of-four VUI could be used just as well. However, rating deviations show that Glicko is much less confident about the ratings given in the best-of-four VUI. This actually happens because even though a vote in the best-of-four VUI results in three pairwise preferences, all the three pairs have a common song and therefore if some one song is dominantly better than the others then not all songs may get a sufficient number of votes. This was indeed found to be the case because seven out of ten callers who used the best-of-four VUI voted for Song 3. *Thus, although we have some evidence that the choice of VUI may not impact the results, a larger study with more songs and subjects would still be needed to confirm this.*

The original subjects from the station's listener base had indicated that the best-of-four VUI was harder for them because they were not able to remember the initial songs they had heard in the list. We tried to test for this by also reversing the order of the songs with another set of 10 graduate students subjects. The results of this experiment are shown in Table 1, and indicate that *reversing the order did not have any effect on the ranking results in the best-of-four VUI.*

We therefore conclude that both the best-of-two and best-of-four IVRs would yield similar results.

4.5 Learning to use IVR systems

We experimented with a different set of 40 subjects with a slightly modified beep VUI on how to best train them to use IVR systems. Our reason for choosing this VUI instead of one of the voting VUIs was that we noticed this to be the harder of the two types of VUIs. While using the voting VUIs in the previous section, subjects generally required prompting only when they were pressing a button for the first time. But in the audio recording VUIs, subjects required encouragement and instructions at several points in the IVR: (a) to start recording name, age, or song, (b) to press a button after recording, (c) to press a button before recording in the Button VUI, etc. Hence we wanted to choose the harder VUI to evaluate its learnability. We wanted to mimic the following real world scenarios for the actual learning methodologies:

- **Training over radio:** A community radio station can train listeners on using an IVR system by airing an instructional promo. To mimic this, we designed an experiment where subjects could listen to an audio snippet played on a laptop that would contain instructions on how to use the the VUI. They could listen to this promo as many times as they liked but could call into the IVR only once. This corresponds to a situation where a caller is allowed to record a song or a message only once, a good example of which is an answering machine where one cannot change a previous recording. We call this the *many promo, one call* experiment.
- **Repeated calls:** It is possible that being able to use the interface more than once may allow the caller to self-learn how to the IVR system. To study how allowing access to the IVR multiple times impacts learnability, we designed the experiment to allow subjects to call into the IVR multiple times in addition to being able to listen to the promo multiple times. We call this the *many promo, many calls* experiment.
- **Training over phone:** CR stations often build a strong bond with their community members through conversations with them over phone calls. This is particularly true for GKA where listeners call not only to make songs requests but also to ask for information about programs, obtain contact details of civic authorities, and sometimes even just to chat with the station staff without any specific purpose. Instructional conversations over phone calls are therefore likely to arise as another training methodology. We designed a third experiment where subjects were provided instructions on IVR use over the phone. We call this the *phone training* experiment.
- **In person handholding:** Station staff sometimes visit the communities for field recordings, where the

staff can give extra time to train listeners on using the IVR systems. We designed a fourth experiment to study the impact of in-person handholding where close help was given to subjects to walk them through the VUI. We call this the *in-person training* experiment.

Note that the above training methods are listed in order of increasing effort for the radio station. While running a promo multiple times is not costly for a station, allowing repeated calls to change previous entries incurs cost and resources in terms of software development and competition design. Training each person over the phone takes up a significant amount of station staff's time, and in person handhold is even more expensive in terms of time and other resources. However, intuitively, each of the above training methods are also listed in the order of increasing learnability: in-person handholding is likely to help more people to learn the VUI compared to just passively listening to instructional promos. Thus, there is a clear trade-off involved that we try to understand as follows.

We combined the four scenarios described above into a single larger experiment due to shortage of subjects. This combined experiment was executed in a phased manner with each subject. A subject was first informed of the hypothetical situation where she was to participate in a talent competition. The subject was given time to decide what poem, song, joke, or any other contribution she would like to record. If the subject could not decide then the station staff would suggest some poem or joke. Once the subject was ready, a promo was played out to her explaining how the VUI was to be used to record her name, age and contribution. For this study, the beep VUI shown in Figure 3(a) was modified slightly to make the caller press a button to terminate a recording, instead of using a 30 second timeout. This was to explore the ability of the subjects to terminate recordings by pressing a button [3]. After playing the promo once, the subject was told that she could listen to the promo as many times as she liked, but could call into the IVR only once. This acted as the *many promo, one call* experiment.

If after the first call, the subject was not able to use the VUI correctly, then she was told that she could listen to the promo again or call again multiple times. This component represented the *many promo, many calls* experiment.

If the subject was unable to use the VUI correctly even after making several calls, then we talked to the subject in more detail and used other ways to explain the IVR to them. Care was taken not to use any gestures or external indicators during this conversation, to mimic a purely verbal communication as on a phone call. This conversation acted as the *phone training* experiment.

If the subject was still unable to use the VUI correctly after the "phone training", we handholded them through the IVR asking them to press a button or record their name, age, and contribution as and when required. This final component represented the *in-person training* experiment.

We defined two tasks for this experiment: (a) task-rec requires recording name, age, and contribution with or without key presses to terminate recordings, and (b) task-keypress requires executing task-rec as well as pressing a button to terminate recordings. Table 2 shows among 40 subjects that undertook the experiment, how many completed these two tasks through each of the training methods, phase by phase. Each column indicates the number of subjects that could complete the task in the corresponding training phase but were not able to complete the task using the previous phase training methods. There are several interesting insights to note:

[3]In earlier work [10], using silence detection to terminate a recording often caused a premature termination as first time IVR users took time to start speaking into the system. Therefore we wanted to experiment whether having users press a button to terminate a recording would work well

Table 2: Results of experiment on learning to use Beep VUI. Each column corresponding to a training method shows number of subjects that could complete the tasks using the training method but could not do so with previous method.

Task	Many promo, one call	Many promo, many calls	Phone training	In-person training
task-rec	17	4	13	6
task-keypress	6	7	15	12

- Seventeen or 42% of the participants were able to complete task-rec in the first call. This is close to the percentage of *good calls* reported by PhonePeti where training was only available through a promo on radio.
- Running an ANOVA test over the results shows that there is a significant impact of different training methods on the completion rates of task-rec ($F = 18.31, P < 0.0001$) and task-keypress ($F = 47.73, P < 0.0001$), which was expected.
- Applying the Student's t-test to each pair of training methods showed that allowing multiple calls into the IVR did not have any statistically significant impact on the completion rates ($P_{rec} = 0.38, P_{keypress} = 0.77$) as compared to a single call. However, training over phone and in person handholding had a significant impact on task completion rates compared to single and multiple calls ($P < 0.0001$ in all four cases). We also found that in-person handholding had significant impact on task completion rates compared to phone training.
- As a further validation of users not being able to closely follow several steps of instruction, we found that repeated promos and repeated calls did not help them remember to do a keypress to terminate recordings. This learning came to the subjects only after phone training or in-person handholding. Therefore, we recommend that when taking speech input in an IVR, thought should be given to specify a reasonable maximum recording duration after which the recording would terminate automatically.
- In several cases during the phone training phase, we explained to the subjects the reasons why it was required for them to press a key for termination by telling them that the "computer needs to know when you want to stop speaking". This seemed to have helped build a better mental model of the system in their minds and improved

the learning. We want to use these insights in the future to design better promos.

Thus, training over phone and in-person handholding showed significant improvement in task completion rates of the VUI. Also, the use of keypress for recording termination was hard to learn for the subjects in the experiments.

4.6 Summary

Table 3: Summary of usability studies and their learning

Study	Learnings
Audio recording methods	1. Words like "record" are hard to relate to for a first time IVR user. 2. Limiting the choices through specific instructions can reduce anxiety. 3. It is hard for users to remember a long sequence of instructions, and rather building the right mental model of the system is more important for learning.
Voting methods	1. Even though both TUTD and best-of-two are easy to use, several cultural factors impact preference of voting methods. 2. Best-of-four method was hard to use because of the difficulty in remembering earlier songs, but that did not impact the voting results.
Learning to use IVR	1. Training over phone and in person showed significant improvement in task completion rates of the VUI. 2. Use of key press for recording termination was hard to learn through promos and repeated calls.

Table 3 summarizes various usability studies done by us, and their corresponding learning. We used these insights to design the IVR systems for the actual Gurgaon Idol competition. Several of our observations are however limited in conclusiveness and generalizability, and need further research. (a) Cultural preferences noticed for different voting VUIs requires larger studies to evaluate cultural impact on voting results. (b) Studies on best-of-two and best-of-four voting VUIs need to be repeated with more voters and non-bollywood songs not known previously to the voters. (c) IVR learnability and other studies need to be repeated with a larger sample size. We are currently working on another round of the competition with a different community radio station, in which we aim to investigate these issues thoroughly.

We next describe execution of the actual competition.

5. GURGAON IDOL

The Gurgaon Idol competition was executed in two phases: a participation phase lasting 14 days, and a voting phase lasting 19 days. During the participation phase, listeners could call into the IVR system and record their name, age, and song. Information about the competition i.e. who should call, what can they record, etc was provided through radio promos. Once the participation phase was over, we shortlisted the entries that had correctly recorded their name, age, and song. Each entry was given an entry number and grouped according to the age of the participant. All participants aged below 30 formed one group and those above thirty formed another.

During the voting phase, these entries were played on air along with their entry numbers. This was done by clubbing the songs from each group into subgroups of three and airing the subgroup together with instructions for voting. Care was taken to ensure that all the songs were broadcast an equal number of times to avoid bias. Listeners could vote in two different ways: they could call into an IVR, listen to songs that are played, and vote on them, or they could note down the entry number aired on the radio and send an SMS "$GKA <$ *space* $><$ *entry_number* $>$" to the station. Both the methods were equally advertised on radio.

We used the Beep VUI[4] during the participation phase, which was modified based on the usability experiments so that the caller was not required to press a button to terminate a recording. The maximum time for recording name and age, and song were set to 20 seconds and 30 seconds respectively so that recording was terminated automatically after this timeout.

In the voting phase, we used both the best-of-two and best-of-four voting VUIs. Even though our experiments showed a preference for the best-of-two VUI, we still used both the interfaces because the best-of-four VUI would give us 3 pairwise preferences compared to only 1 preference per vote in the best-of-two VUI.

5.1 Participation phase

During the participation phase, we accepted a total of 31 entries out of 85 recordings. The rejected recordings were of poor quality and were replaced by new recordings where we asked the callers to call again. We describe later in Section 6 several practical challenges to keep in mind while conducting such IVR based singing competitions. Among the 31 participants selected for the voting phase, 21 were aged below 30 and 10 were aged above thirty. *Thus, this concept of a competition did attract more youth, which was a success for GKA as the station had so far struggled to engage young listeners. Women participants were however few, and that particular objective was not met.*

5.2 Voting phase

During the voting phase, 221 callers called into the IVR system. Surprisingly, only 10 of the 109 callers that were presented best-of-two VUI and 19 of 112 callers that were presented best-of-four VUI actually voted for a song. All the other callers simply hung up the call without voting. Another surprising aspect of the voting results was that the number of callers that voted in the best-of-four VUI was more than those that voted in the best-of-two VUI, although it is hard to infer anything statistically significant from this. *Thus, we suspected that factors beyond usability of the two voting interfaces impacted the usage of the voting system.*

We called back a sample of 47 callers who did not vote to understand their reasons for not voting. Nine callers said that they did not vote because they did not like any of the

[4]We used Button VUI also with the purpose of comparing task completion rates of the two VUIs, but a bug in the IVR system did not allow the Button VUI to be used correctly. We present data collected only from the Beep VUI in this paper.

songs! The main reason cited by them was that the audio quality was too poor. While the station could have filtered songs to allow voting only on songs they thought were good, it chose not to do so to avoid bias, which resulted in less votes. Another 7 callers we interviewed, had voted for a song before all the songs were played. Although, the IVR did not allow voting before all the two/four songs were played to ensure equal exposure of all competing songs to the voter, this design choice resulted in lesser number of votes being registered. Yet another 18 callers said that the system malfunctioned and they did not hear any songs when they called. Our analysis of the call logs however revealed that these calls were actually disconnected while the instructions were being played out! The callers were therefore making excuses, and would have called only to check out the system, a conclusion validated by discussions with the station staff and our past experience with PhonePeti [10]. All this analysis revealed that much more was going on rather than just the usability of the voting VUIs.

5.2.1 Voting over SMS

As a backup in case the IVR voting mechanism did not work, we had also put an SMS-based voting mechanism in place. We received 264 SMS votes from 68 different phone numbers. These votes were of course direct thumbs-up votes for specific songs, and could not be used in the Glicko ranking algorithm. The higher number of SMS votes could indicate that in this particular demography, SMS may be a more convenient way for the station to engage with its community. It could also indicate that listeners may prefer to directly vote for a specific song rather than differential voting on preference pairs.

5.3 Competition result

We used Glicko ranking algorithm to obtain a global order of the songs based on the pairwise preferences obtained from the IVRS. Recall that Glicko accepts pairwise preferences and produces as output, a tuple (R_s, RD_s) for each song s. Thus, given two songs s_1, s_2, if $R_{s1} + RD_{s1} < R_{s2} - RD_{s2}$ then s_2 is rated better than s_1.

The votes obtained from the IVRS were insufficient to obtain the top two songs in both the groups using the above rule. We were however able to get the top song in both the groups. We used SMS votes to find the second winner from each group.

6. PRACTICAL CHALLENGES IN EXECUTING GURGAON IDOL

While we focused on usability aspects of voice user interfaces and their learnability, we realized that there are several other aspects that need equal or more attention for an event like Gurgaon Idol to succeed.

One of the biggest challenges for us was the poor quality of audio recorded over a phone call. There are several dimensions to this. First, the 8KHz sample rate on telephony networks seems to be insufficient for song recordings. Second, several recordings had audio gaps due to a temporary signal loss while recording. Such breakages may be acceptable in a conversation but they significantly impact a song's quality. Another practical issue was that since participants had never sung into a phone before, they did not know how far they should hold their phone from their mouth to get recordings at an acceptable volume levels. Handheld recorders usually have level indicators to provide feedback to people using the devices. No such feedback could be made available here except for replaying the recorded song. As a result of these problems, several songs had to be discarded or the callers had to be asked to call again to revise their entry.

Deciding the duration of each phase of the competition was also a difficult task as neither Gurgaon Ki Aawaz nor we had conducted such a competition before. Initially we had planned to keep both the phases only 7 days long in an attempt to keep the competition short and exciting. However, we had to change the durations based on the number of participants and votes we got in each of the phases. Running the system off a toll-free number could have impacted the call volumes, something we plan to do in the future to evaluate the relative volumes when callers have to pay Vs. when the line is toll-free for the callers.

We also realized that by conducting only a singing competition as against a generic talent hunt, we ruled out a large percentage of the population that was not interested in singing songs. The decision proved counter productive for the station as only a few hundred members of the community participated in the competition, which is a small number in the context of urban India. We explicitly highlight this failure to enable others to learn from it.

7. RELATED WORK

We categorize related work in three domains and present each of them below:

IVR systems for development: IVR systems have been extensively used in the development context, mostly to play out audio recordings such as announcements and updates [19, 21, 20, 1], or to build peer-to-peer information sharing networks [17, 14]. The Tamil Market project [19], Healthline [21], and the phone broadcasting system for urban sex workers [20] focused on carefully understanding the context in which the IVR system was deployed and experimented with the timing of calls and content to be made available on the system. Information sharing systems such as Avaj Otalo [17] and CGNet Swara [14] focused on the kind of interactions enabled through the IVR system, and how they related to the actual social interactions on the ground. Our work, while done in the context of a competition, focuses on the usability of the IVR interfaces and explores methods to enable learnability of the interfaces.

Graphical user interfaces for low literate users: Recent work has also been done in the domain of building GUIs for low literate users. The need for graphical elements [16] instead of text, voice annotation and audio feedback [6], semi-abstracted graphics [13], and consistent help information across mobile phone applications [13] have been repeatedly highlighted by this body of work. Authors have generally recommended contextual design methods because the domain of audio-visual interfaces for low literate users is relatively new. The same principle also applies to voice user interfaces for low literate users, particularly for those with little past exposure to IVRs. Our work contributes to this relatively nascent field by exploring options for recording audio and voting VUIs with populations that have had little prior exposure to IVR systems.

Voice user interface design for the developing world: Some recent studies on VUIs for the developing world have

compared DTMF Vs. speech as input modality [18, 19, 22], where speech was preferred if a recognizer with a sufficiently high accuracy in the local language was available, and DTMF was preferred otherwise. PhonePeti [10] was used to receive feedback from listeners of a community radio station, and evaluated methods soliciting specific information from callers. Lerer et al. [11] explored the challenges of conducting voice based surveys without giving prior training to the participants on its usage. Medhi et al. [12] studied voice user interfaces for banking activities. To the best of our knowledge, no study has been done so far to understand VUI design for recording audio and for voting interfaces. Additionally, we are the first ones to study mechanisms for helping first time IVR users learn how to use VUIs.

8. CONCLUSION AND FUTURE WORK

We presented a series of usability tests designed to evaluate different methods of recording audio and crowdsourced voting on IVR systems meant for semi-literate people with little prior exposure to IVR systems. We showed that phrases like "record after the beep" do not come intuitively to this demography, several cultural factors affect the type of crowdsourced voting method to use, presenting users with pairwise comparisons to evaluate pairs of song recordings is easier than choosing the best out of four options but the voting results obtained from using best-of-four are comparable to those obtained from using a best-of-two voting method. Through formal usability studies we also showed that allowing users to call multiple times to record and re-record audio does not impact learnability of the IVR system, but training over phone and in-person handholding improves learning significantly. Much future work remains though. Among other things, we plan to evaluate voting accuracy and the usability of best-of-four Vs. best-of-two voting methods in larger studies to confirm some intuitions we have developed in the current work to build crowdsourced ranking systems.

Acknowledgement

The authors would like to thank Ed Cutrell at Microsoft Research India, Amit Nanavati at IBM India Research Labs, and the reviewers for their valuable comments. The authors also acknowledge the significant contribution of Soumya Jha, Gurgaon Ki Aawaz in conducting this research.

9. REFERENCES

[1] Sheetal Agarwal, Arun Kumar, Amit Nanavati, and Nitendra Rajput. User-Generated Content Creation and Dissemination in Rural Areas. In *Journal of Information Technologies and International Development*, 2010.

[2] A. Chand. Designing for the Indian rural population: Interaction design challenges. In *Development by Design Conference*, 2002.

[3] Weiwei Cheng, MichaÃńl Rademaker, Bernard De Baets, and Eyke HÃijllermeier. *Predicting Partial Orders: Ranking with Abstention*. 2010.

[4] Mark E. Glickman. Parameter estimation in large dynamic paired comparison experiments. In *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1999.

[5] Ralf Herbrich, Tom Minka, and Thore Graepel. TrueskillâĎć: A bayesian skill rating system. In *Advances in Neural Information Processing Systems*, 2007.

[6] M. Huenerfauth. Developing design recommendations for computer interfaces accessible to illiterate users. Master's thesis, University College Dublin, 2002.

[7] Z. Koradia, A. Premi, Balachandran C., and A. Seth. Using icts to meet the operational needs of community radio stations in india. In *Proceedings of the First ACM Symposium on Computing for Development*, ACM DEV '10, pages 21:1–21:9, New York, NY, USA, 2010. ACM.

[8] Zahir Koradia. *Tools for Use in Integrating Mobile Phones into Local Educational Programming*, chapter 22. Commonwealth of Learning, 2012.

[9] Zahir Koradia, C. Balachandran, Kapil Dadheech, Mayank Shivam, and Aaditeshwar Seth. Experiences of deploying and commercializing a community radio automation system in india. In *Proceedings of the 2nd ACM Symposium on Computing for Development*, ACM DEV '12, pages 8:1–8:10, New York, NY, USA, 2012. ACM.

[10] Zahir Koradia and Aaditeshwar Seth. Phonepeti: Exploring the role of an answering machine system in a community radio station in india. In *Proceedings of the 5th International Conference on Information and Communication Technologies and Development*, ICTD '12, 2012.

[11] A. Lerer, M. Ward, and S. Amarasinghe. Evaluation of ivr data collection uis for untrained rural users. In *Proceedings of the First ACM Symposium on Computing for Development*, ACM DEV '10, pages 2:1–2:8, New York, NY, USA, 2010. ACM.

[12] I. Medhi, S. N. N. Gautama, and K. Toyama. A comparison of mobile money-transfer UIs for non-literate and semi-literate users. In *CHI*, 2009.

[13] Indrani Medhi, Aman Sagar, and Kentaro Toyama. Text-free User Interfaces for Illiterate and Semi-literate users. In *Information Technologies and International Development*, pages 37–50, 2007.

[14] Preeti Mudliar, Jonathan Donner, and William Thies. Emergent Practices Around CGNet Swara, A Voice Forum for Citizen Journalism in Rural India. In *Information and Communication Technologies and Development*, 2012.

[15] Donald A. Norman. *The Design of Everyday Things*. Basic Books, 2002.

[16] Tapan S. Parikh, Kaushik Ghosh, and Apala L. Chavan. Design Studies for a Financial Management System for Micro-Credit Groups in Rural India. In *ACM Conference on Universal Usability*, 2003.

[17] N. Patel, D. Chittamuru, A. Jain, P. Dave, and T. S. Parikh. Avaaj Otalo - A Field Study of an Interactive Voice Forum for Small Farmers in Rural India. In *CHI 2010*, 2010.

[18] Neil Patel, Sheetal Agarwal, Nitendra Rajput, Amit Nanavati, Paresh Dave, and Tapan S. Parikh. A Comparative Study of Speech and Dialed Input Voice Interfaces in Rural India. In *CHI 2009*, 2009.

[19] Madelaine Plauche, Udhyakumar Nallasamy, Joyojeet Pal, Chuck Wooters, and Divya Ramachandran. Speech Recognition for Illiterate Access to Information and Technology. In *Information and Communication Technologies and Development*, 2006.

[20] Nithya Sambasivan, Julie Weber, and Edward Cutrell. Designing a phone broadcasting system for urban sex workers in India. In *CHI*.

[21] J. Sherwani, N. Ali, S. Mirza, A. Fatma, Y. Memon, M. Karim, R. Tongia, and R. Rosenfeld. HealthLine: Speech-based access to health information by low-literate users. In *ICTD*, 2007.

[22] Jahanzeb Sherwani, Sooraj Palijo, Sarwat Mirza, Tanveer Ahmed, Nosheen Ali, and Roni Rosenfeld. Speech vs. Touch-tone: Telephony Interfaces for Information Access by Low Literate Users. In *ICTD*, 2009.

[23] Bart Sullivan. *Freedom to Learn: Blending Interactive Voice Response and Radio*, chapter 21. Commonwealth of Learning, 2012.