VideoKheti: Making Video Content Accessible to Low-Literate and Novice Users

Sebastien Cuendet
EPFL, Lausanne
Switzerland

Indrani Medhi
Microsoft Research
India

Kalika Bali
Microsoft Research
India

Edward Cutrell
Microsoft Research
India

ABSTRACT

Designing ICT systems for rural users in the developing world is difficult for a variety of reasons ranging from problems with infrastructure to wide differences in user contexts and capabilities. Developing regions may include huge variability in spoken languages and users are often low- or non-literate, with very little experience interacting with digital technologies. Researchers have explored the use of text-free graphical interfaces as well as speech-based applications to overcome some of the issues related to language and literacy. While there are benefits and drawbacks to each of these approaches, they can be complementary when used together. In this work, we present VideoKheti, a mobile system using speech, graphics, and touch interaction for low-literate farmers in rural India. VideoKheti helps farmers to find and watch agricultural extension videos in their own language and dialect. In this paper, we detail the design and development of VideoKheti and report on a field study with 20 farmers in rural India who were asked to find videos based on a scenario. The results show that farmers could use VideoKheti, but their success still greatly depended on their education level. While participants were enthusiastic about using the system, the multimodal interface did not overcome many obstacles for low-literate users.

Author Keywords

Speech interface; Novice users; Low-literate users; Mobile design; Multimodal interfaces; ICTD; HCI4D

ACM Classification Keywords

H.5.2. User Interfaces: User-centered design

General Terms

Human Factors; Design

INTRODUCTION

Mobile technology is spreading rapidly and offers new opportunities to reach people in the developing world. It is estimated that in 2011 there were 5 billion mobile phone subscriptions in developing countries, a number growing at 20% a year [29]. Despite this remarkable penetration, there remain many barriers to fully utilizing many of the capabilities offered by this technology. Basic challenges include infrastructural constraints in power and connectivity; poverty and other financial limitations; and problems acquiring and maintaining hardware. But beyond these serious difficulties, there are other problems that are central to HCI. The first is the huge variety of languages spoken in developing regions. While most technology systems in the "Global North" are designed with at most a half-dozen languages in mind (and often only English), designs for the developing world must consider orders of magnitude more; more than 400 languages are spoken in India alone and the number of languages spoken in Africa is estimated at over 2100 [14]. The second major problem is the widespread lack of formal education (and corresponding low literacy) combined with a lack of exposure to information technology. While millions of low-literate people use mobile phones, they use them primarily for voice calls, avoiding the use of SMS, address books or other more complex functions [4].

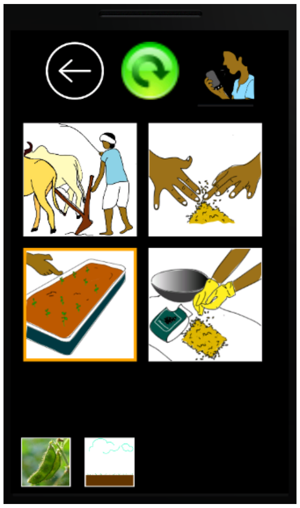


Figure 1. A screenshot of the application.

Researchers have explored a variety of interfaces to make systems accessible to novice and low-literate users. These interfaces often feature multiple modalities such as audio, graphics, and text to help clear the hurdles of language, low-literacy and unfamiliarity with technology [17, 25, 10, 11, 28]. Medhi et al. found that common text-based interfaces were completely unusable for low-literate users, and they are error-prone for literate, but novice, users. In contrast, graphical interfaces with audio output were more successful. Speech interfaces met with mixed results: they were more usable

CHI 2013, April 27–May 2, 2013, Paris, France.

than a corresponding graphical interface for some users but led other users to give up on the task [17]. This echoed a previous observation from the Tamil Market kiosk, for which the original, pure speech interface was augmented with graphics, touch, and typing to accommodate unsophisticated or new users [25].

A conclusion from this research is that there are advantages and drawbacks to touch, graphics, and speech, but their complementarity may lead to a better experience for novice or non-literate users [17, 25]. Speech is a natural means of expression well suited to input, but spoken output can be hard to understand and remember. Graphical symbols and photos are excellent for output, capable of conveying large amounts of non-linguistic information, but there is a danger of ambiguity and confusion about what the images intend to convey.

Despite the potential benefits, speech-based systems for low-literate users remain rare. While there are many challenges, the main obstacle remains the absence of speech recognition engines for many of the languages spoken in the developing world. Training a single automatic speech recognizer (ASR) for a given language/dialect/accent requires many hours of manually annotated speech, and for most languages in these regions, such corpora simply do not exist. In an effort to overcome this problem, Qiao et al. [26] developed SALAAM, a method that requires only a fraction of the training data as traditional ASRs to create a speech recognizer. While ASRs are much cheaper and easier to create using SALAAM, the method does have several limitations: it only works for small vocabulary (about 100 words or fewer) and only allows one vocabulary item to be recognized at a time.

These limitations do impact the natural interaction promised by speech; free speech is not possible, and one must shape the interaction for short utterances. Despite this limitation, we were curious to see whether such a method could be used to build a speech interface for a language/dialect with no available commercial ASR. Could we combine a SALAAM ASR with touch and graphics to build an application usable by low-literate and novice users? While SALAAM requires less training data than a full-fledged ASR, it still requires more work than developing a graphics-only interface. Is it worth the extra engineering effort (to build the ASRs) and network bandwidth for sustained use? How would low-literate and novice users react when faced with this multimodal interface? To answer these questions, we developed a multimodal system to allow low-literate farmers from a remote area of Madhya Pradesh (MP), India, to access videos on farming-related subjects. Our system, VideoKheti, features a multimodal interface with speech, graphics, and touch on a smart phone or tablet. We detail the development of the system and report the results of a field study involving 20 farmers. To our knowledge, this is the first study exploring the use of a multimodal speech interface in a novice rural population using an ASR for local language/dialects. The results show that (1) successful usage of VideoKheti was correlated with the education level of the users, and the speech interface was not able to help low-literate participants overcome problems in using the system; and (2) speech interaction worked best for cases where there was a long list of choices and selections comprised short and familiar words or expressions.

BACKGROUND

This work was done in partnership with Digital Green[1], an NGO active in agricultural extension and training for smallholder farmers in India [6]. Digital Green's core idea is to screen videos of farmers demonstrating farming practices relevant for the crops and context of an area in the local dialect and idiom of that region. Videos typically feature a local farmer in his field explaining and demonstrating an agriculture-related technique. These videos are screened by a local mediator who chooses what videos to play and answers questions about the practices that are shown. The videos are commonly stored on an SD card and projected against a wall by means of a pico-projector (some villages still use televisions and DVD players). Digital Green's model has been extremely successful, allowing them to reach 120,000 viewers across rural India in just a few years. VideoKheti was designed to address two main issues: First, farmers cannot view these videos without the mediator, so there is little opportunity for them to help themselves by reviewing details that they may have forgotten. Second, the current system is difficult for mediators, since they have to access the video library directly through the projector interface (or by thumbing through a large stack of DVDs). Basic features such as searching or browsing by crop, season or activity do not exist. The goal of VideoKheti is to address both of these issues by providing an easy-to-use interface on a single smartphone or tablet. The interface could be used by the mediator, but also directly by the villagers.

RELATED WORK

Interfaces for low-literate users

Research into the design of interfaces for users with little or no formal schooling is still somewhat new. Fifteen years ago, researchers developed a healthcare application on a Newton PDA and tested it with 10 health workers in rural India [7]. They found that text interfaces led to issues of localization and translation for technical terms. Others confirmed problems with text interfaces for low-literate users [8, 19, 23]. However, it was soon discovered that while the problem of reading could be overcome by audio output [8, 18], other interaction issues such as object selection, using menus and buttons, and determining the state of the system were harder to overcome [8]. To help first-time users, video clips that included dramatizations of the scenario were effective [20]. Static, hand-drawn representations proved to be better understood than photographs or icons [18]. Using audio output through voice annotation of graphics generally helped with speed and comprehension [18, 23, 25]. Big Board attempted to side-step these problems by using the camera on phones to allow users to query for information from public places using bar codes [15].

While these systems used graphical interfaces complemented by audio output, other work explored the use of speech input. Speech is appealing because it is natural and avoids some of

[1] http://www.digitalgreen.org, last visited: January 11 2013

the issues related to literacy. In addition, speech interfaces avoid many language-related complications, such as the absence of keyboard standards and unique fonts/scripts (or, in the case of some languages, a lack of *any* written script) [8, 3]. Comparing a Wizard-of-Oz spoken dialog interface with a graphical interface revealed that the graphical interface led to a higher completion rate but that users who understood the dialog system were able to complete the task faster [17]. However, using audio as an input mechanism is technically challenging and costly because of the large amount of data needed to train an ASR with an acceptable word error rate [21]. These data typically are not available for languages in developing countries, and collecting them would be a huge effort with small impact because of the variations in dialect and accent [25]. However, research has shown that by limiting the vocabulary to less than 100 words, one can develop a reasonable speech-based system for some applications [25, 28].

Exploiting the ubiquity of inexpensive mobile phones, Avaaj Otalo and Spoken Web are two speech applications designed for farmers in India. Avaaj Otalo [24] is an interactive voice application that allows users to ask questions and browse others' questions and answers on agricultural topics through a simple mobile-phone call. The system is controlled by simple speech prompts or by numeric inputs and is used to collect information about farmers' harvests in the state of Gujarat, India. The ASR was adapted from American English to be usable by Gujarati speakers. Spoken Web attempts to create a secondary audio version of the web, accessible by any phone. The framework presented in [10] allows users to create ”voice-sites” by means of a voice interaction. These voice-sites can then be accessed by a phone call. There have been several demonstration applications on Spoken Web, one of which was to provide farmers with crop and market information. These systems demonstrate the utility of voice-based systems for providing information to low-literate farmers in rural India. VideoKheti seeks to combine the advantages of voice-based interaction with a rich graphical interface to address farmers' information needs. In addition, VideoKheti is designed to be used with an ASR that can be created for a variety of languages and dialects at a reasonable effort and cost.

Multimodal interfaces and speech-based systems

The benefits of multimodal interfaces have been widely described in the literature. They have been shown to increase efficiency by 10% on average, and to increase satisfaction by allowing users to choose their preferred style of interaction [22]. However, the main advantage of multimodal interfaces is probably their ability to significantly improve error handling and reliability [22]. This is an important asset in the context of this work as our target users are generally unfamiliar with technology, and the cost of system errors can become dramatically amplified. There are a number of reasons for including speech in multimodal interfaces [5], including being able to interact with the system without using one's hands and not having to give full visual attention to the screen. On the other hand, graphics allow one to display options efficiently on a screen instead of streaming them in a potentially lengthy audio list [12]. In systems using speech and graphics, the satisfaction of users was correlated with the performance of the ASR, and users preferred directed prompts to open-ended ones [27]. The Multimodal Access To City Help system (MATCH) was one of the first speech-enabled mobile multimodal interfaces [9]. MATCH users could provide input through speech or by drawing on a display with a stylus. Currently, most new mobile phones are inherently multimodal, including a graphical touch interface combined with voice assistants such as Siri (iOS), Google Search (Android), and Bing Search (Windows Phone). VideoKheti builds on previous work on multimodal interfaces and is the first system targeted at low-literate users that combines touch, graphics, and speech input on a mobile interface.

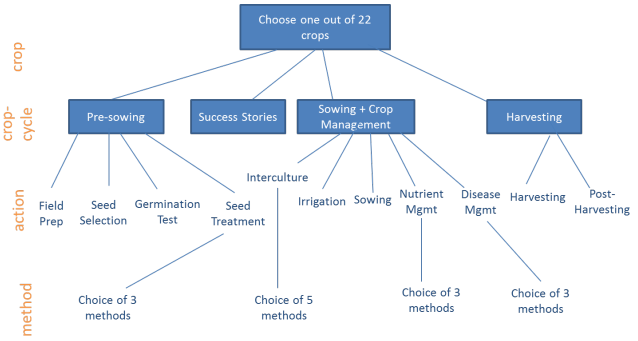


Figure 2. The navigation tree.

VIDEOKHETI: DESIGN AND DEVELOPMENT

VideoKheti allows its users to find and watch agriculture-related videos on a mobile phone. The interaction between the user and the system is multimodal; the user can use speech or touch to navigate the system, and output is a combination of graphics and audio.

Navigation tree

There are a total of 147 videos available in VideoKheti, classified by farming experts according to the features of their content. These features were organized in four levels: *crop*, *crop cycle*, *type of action*, and *type of method*. We built a top-down navigation tree based on these four levels (Figure 2). The user therefore has to make a maximum of four choices before reaching a list of videos to play (choosing a video out of the list is arguably a fifth choice). In the first level, the user chooses one of 22 available crops. The crops are shown on two screens that can be navigated using a simple slide gesture. The second level is the choice of the *crop cycle* (e.g., pre-sowing, harvesting). The third level is the *type of action* (e.g., field preparation, disease management) and the fourth level is the *type of method* (e.g., organic, conventional). Some paths lead to the video choice screen in fewer than four levels; for example, the choice of *type of method* does not make sense for concepts such as irrigation or sowing.

GUI description

The interface of VideoKheti is completely text-free. Figure 1 shows a screenshot of the application. The screen of the phone is split into three main parts: a header panel, which is present on every screen; icons showing the choices available at that stage of the navigation; and breadcrumbs showing

Figure 3. A researcher collecting the audio samples.

what has been selected so far and the position in the hierarchy. The header panel includes three buttons. The back button, on the left, allows the user to go back to the previous screen. The middle button allows the user to play the system prompts for the current screen. Finally, the button on the right is the speech input button. When the system begins listening for speech input, this image is changed to an ear and a distinctive sound is played. Speech input stops either automatically after 7 seconds, or when the user presses the speech input button. When the recording finishes, the skin of the button is switched back to the original image. The main body of the page is occupied by the graphics displaying the navigation choices available. The choices are organized as a grid where each navigation choice is a grid item. Except for the first navigation level where the various crops are represented by pictures, all graphics are hand-drawn, as recommended in the literature on low-literate users [18]. Finally, the breadcrumbs are a graphical reminder of the choices made so far at each of the four levels of navigation.

Audio output and touch

Previous work has underlined the benefits of combining voice annotation with graphics [17]. In VideoKheti, the combination of graphics and voice annotation is present for all the graphics representing the navigation choices. On every page, the system asks users what information they want to know about and explicitly names all choices available (except at the first level, for which there are too many choices to enumerate). A choice uttered by the system is simultaneously highlighted on the graphical interface. Once all prompts are done playing, a first touch on an item will highlight it and play the corresponding word. A second touch validates the choice and the system navigates to the corresponding page.

Speech input

One aspect of natural speech interaction is the ability for the user to barge in. However, in an open-microphone system like VideoKheti, which itself plays human-voice prompts, reliably detecting barge-in from users is not possible without significant signal processing work beyond the scope of this research. We tried two alternative options. The first was to introduce a ”push-to-talk” button that the user pressed to start and stop the audio recording. The second was to automatically start the recording at the end of the system prompts and to stop it after a certain amount of time. After informally testing both, we opted for the second option and set the length of the recording window to 7 seconds, enough time to comfortably record the longest of the valid inputs. We opted against the ”push-to-talk” button because it would have required as many operations as navigating with the graphics (2 taps on the screen), but with more uncertainty about the success of the operation. Moreover, it would have removed the advantage of not having to directly handle the device (beyond holding it), one of the advantages observed for speech interaction [17]. This came at the price of a more system-driven interaction, since the speech recording started automatically after the end of the prompts.

Speech recognition

As noted above, developing a full-fledged ASR for the language communities we wish to target was not an option. VideoKheti therefore uses the SALAAM method [26] to recognize the Hindi dialect of the rural villages we were working with. SALAAM allows small-vocabulary recognition by using the acoustic model of any existing ASR and performing cross-language phoneme mapping between the language of the ASR and the target language (Hindi, in our case). It is fully automatic and requires a very small amount of training data. The limitation of this approach is that it can only be used efficiently with vocabulary of 100 word types or less, where a word type refers to a single word or phrase.

Data collection

The original vocabulary of VideoKheti consisted of 79 word types. In order to match the accent and dialect of the target population, the data required for SALAAM was collected in the villages of the farmers we worked with. A total of 23 participants (10 women) recorded two samples of each of the 79 words. An ad-hoc application was deployed in a mobile phone and used to prompt the participants and record their input. The order of the prompts was randomized and a four-second silence separated the prompt from the start of the recording to prevent mimicking. More than 3500 samples were recorded over two days in two different villages in the region of Rajgarh, MP. The recordings were made in an open environment, which included background noises from the village as well as a variety of distractions for participants (see Figure 3). The specific vocabulary provided by our NGO partner turned out to be problematic, as some participants had trouble memorizing and repeating some of the longer or more technical expressions. While we expected the NGO to provide the appropriate local vocabulary for the agricultural information in these areas, there was often a mismatch between the language the NGO used and that of the farmers. As we describe later, this had a significant impact on the usability of the speech system.

Training and accuracy

Audio samples were used to generate a list of phoneme sequences using the SALAAM method with very little training data for a 79 word vocabulary. To prevent navigation errors in the application, the system was tuned to reduce the number of false positives (and consequently augmenting the number of true negatives). The final system we deployed on the field was trained on the speech of 20 users and had an accuracy of 96% on test data. In the field, evaluation was complicated due to background noise, long silences and partial input. However, on excluding the samples containing silence and discussions between the researchers and the farmer, the ASR had an accuracy of 90% (including partial phrases), and 94% (excluding partial phrases) on field data. A deeper and more rigorous analysis of the speech performance is provided in [1]

FIELD STUDY

To test the usability of the system, we ran a study with 20 farmers living in rural villages in the state of Madhya Pradesh in central India. The study took place over three days in three different villages. The goal of the study was to assess whether farmers would be able to use the VideoKheti system and to observe the differences in the usage between a system using only graphics and touch, and a system with speech, graphics and touch.

Conditions

There were two conditions: speech and touch with graphics and audio output (STGA), and touch with graphics and audio output (TGA). The participants were split randomly between the two conditions (10 in each). In the TGA condition, the only input means was touch, while output was both graphical and audio (voice annotation). In the STGA condition, the participants could use speech or touch for navigation, while the output was identical to the TGA condition. The system prompts in the STGA condition were adapted to make users aware of the two input modalities and we made a special effort to call out both speech and touch interactions in all instructions.

Participants

The 20 participants (8 females) came from three villages in the region of Rajgargh, MP. All participants belonged to farmer families and had been or were currently involved in farming activities. Except for one participant, all participants were regular users of mobile phones. The age and the education level of participants were balanced across gender, and those three variables were further balanced across the two conditions as much as possible given the availability of villagers (see Table 1). The education level varied greatly among the participants, ranging from no formal education to a bachelor's degree corresponding to 15 years of schooling. We were concerned that differences in participants' level of education might lead to different performance and usage patterns of VideoKheti. Therefore, we split participants into two groups: low-literate (up to completed fifth standard education) and higher-literate (sixth standard or above). In the STGA condition, there were 6 low-literate farmers (2 women and 4 men) with an average of 2.6 years of schooling, and 4 higher-literate farmers (2 women and 2 men) with an average of 13 years of schooling. In the TGA condition, there were 4 low-literate users (2 women and 2 men) with an average of 4 years of schooling and 6 higher-literate users (2 women and 4 men) with an average of 9.8 years of schooling.

Table 1: Participant information.

	Low-lit	**High-lit**
Number of participants	10	10
Mean/median age	45 / 50	28 / 28
Mean/median education level	3.2 / 4.0	11.1 / 11.5
Gender (male/female)	6/4	6/4
	TGA	**STGA**
Number of participants	10	10
Mean/median age	36 / 36	37 / 35
Mean/median education level	7.5 / 7.5	6.8 / 5.0
	Female	**Male**
Number of participants	8	12
Mean/median age	32 /29	40 / 38
Mean/median education level	7.0 / 6.5	7.25 / 6.0

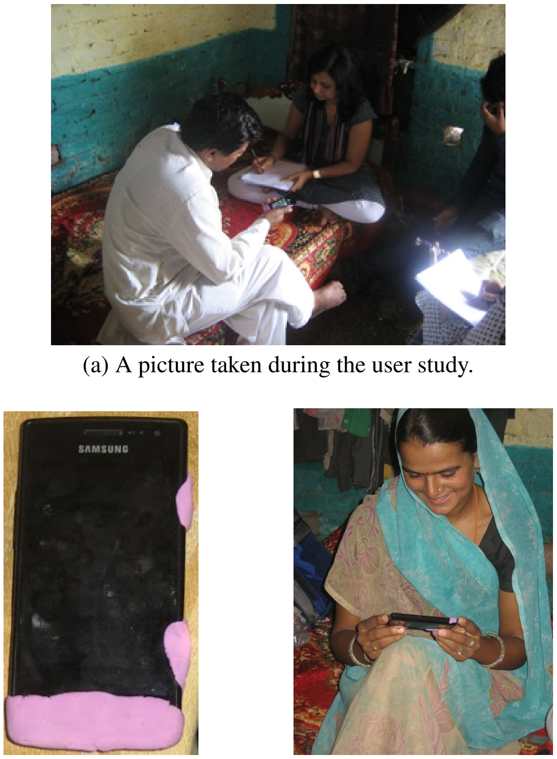
Technical set-up

In each of the villages, the study was conducted in a closed or semi-closed room (Figure 4a). The phone used was a Samsung GT-I8350 running Windows Phone 7.5. To prevent unintentional triggering of the buttons that could not be programmatically disabled, we covered those buttons with Play-Doh (Figure 4b). Because there was no reliable wireless data connection in the villages we were working in, we needed to simulate a broadband data connection to connect the phone to the speech server. The phone was connected to a wi-fi network that was set up locally on a Lenovo T400 laptop that served as a speech-recognition server. The amount of data transferred for each recording was around 700 kilobytes. This could be reduced to improve bandwidth consumption by some simple processing (e.g., silence detection, audio compression).

Experimental process

The same researcher acted as experimenter for all participants and followed a script. Participants came in one by one. The researcher first gathered information about the participant such as age, education level, and phone usage (this took about 2 minutes). She then briefly explained the scope of the project and the tasks that the participant would have to complete (1 minute). The next step was to demonstrate how the application worked and to complete the training task (3-4 minutes). She described the scenario of the training task and completed the task with the participant, demonstrating both the speech and graphics modalities for the STGA condition and only the graphics for the TGA condition. When both speech and graphics were available (STGA), participants were told that they were free to use either input as and when they preferred. In each case, we logged all interactions with the interface for later analysis.

Throughout the experiment, the researcher helped participants when needed. The help provided was classified into three categories by a second experimental observer during the experiment: simple encouragement, spoken reminders, and hand-holding (actually helping them to complete the task). A *general assistance score* was computed as a weighted sum of the number of prompts in each of three prompt categories.



(a) A picture taken during the user study.

(b) The phone used for the study, with Play-Doh.

(c) A participant watching a video on the phone.

Figure 4. Elements of the user study.

The weights reflected the amount of help provided by a given type of prompt: 1 for simple encouragement, 2 for a spoken reminder, and 3 for a hand-holding intervention. This score was our primary usability metric, with higher scores indicating more difficulty completing a task. At the end of the last task, two female researchers spoke with participants about their experience completing the tasks and listened to their feedback. Participants were then thanked and given a gift worth approximately USD 5 for their participation.

Tasks

A total of four tasks were given to each participant. The first task was the training task that the experimenter completed with the participant. All four tasks had the same structure: Find one or more videos that matched a given scenario. Each scenario was read from a script:

1. This year you chose to cultivate maize. You need to treat the seed before you can sow. Find a video that will explain how to treat the seed using an organic method.
2. You notice that some of your soybean crops have a disease, due to some insect. Find a video that will demonstrate how to do insect control on your crops.
3. You now want to cultivate oranges. You have the seeds and the field is ready, but you have never planted oranges before. Find a video that will explain to you how to do it.
4. You are now growing coriander. You wish to know what nutrients you can give to your crops to increase yields. Find a video that will explain to you how to do that using a chemical method.

A task was considered completed when the participant reached the page displaying the videos and signified that she was done. Each task required going down a different path of the navigation tree. All interactions between the researchers and the participants were in Hindi.

RESULTS

As our main interest is how the system was used, we focus primarily on the *general assistance score* and descriptions of specific interactions. All participants in both conditions were able to complete all three tasks, with the exception of one person (in the STGA condition) who could not complete the second and third tasks despite repeated assistance. While most participants could complete the tasks, there were very large differences between users in the assistance required from the experimenter, the usage of speech (when available), and a willingness to use/explore the interface.

Impact of literacy

As shown in Figure 5, there were dramatic differences in the general assistance score for lower- and higher-literate participants in both conditions. A mixed-model 3-way ANOVA (Condition x Education Level x Task) revealed a significant difference for education level [F(1,16)=11.83, $p<0.01$], but no significant effects for Condition or Task and no significant interactions.

It is important to note that literacy is also highly correlated with age for these farmers. The median age of lower-literate farmers was 50, while the median age of the higher literate farmers was 28. Thus any observed differences in literacy are also effectively generational differences. Higher-literate users were much more confident at exploring and using the interface. Occasionally, some of the higher-literate participants would select an option before the audio prompts were done playing; this barging-in never happened in the case of low-literate users. In the TGA condition, higher-literate users selected more items by touch (74 vs. 54 selections), indicating a willingness to experiment with the interface by exploring hierarchy. In the STGA conditions, lower-literate users were far more likely to use touch instead of speech (56 selections via touch vs. only 19 for higher-literate users).

Looking only at the STGA condition, where participants were free to use either the graphics or the speech interface to navigate, we see more effects of education level. When attempting to command the system with speech, there were two pos-

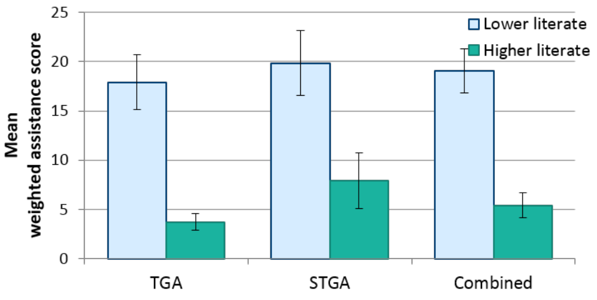


Figure 5. Mean weighted assistance scores (±SEM) per task for each condition by education group.

Table 2: Speech accuracy.

Literacy	Not valid	Valid
All	0.50	0.50
Low	0.62	0.38
Higher	0.33	0.67

sible outcomes: the system recognized what the user said and performed the navigation action (”valid”), or it did not recognize it and nothing happened (”not valid”). A third outcome could have been that the system inappropriately recognized a word, though this never happened in practice. Table 2 shows the detail of the outcomes when using the speech interface. The ratio of valid words for low-literate users was close to half that of higher-literate users. In other words, about 33% of attempts to use speech by higher-literate users were rejected, while rejections were close to 60% for low-literate users.

Figure 6 shows details of actions at each level of the application as a percentage of total actions at each level in the UI. It includes the two types of speech actions described above (valid and not valid) as well as touch interactions. For both lower- and higher-literate users, speech was most heavily used in the first level, crop selection. At this level, the speech interface worked perfectly for higher-literate users, but lower-literate users encountered more rejections than valid detections. For all users, successful speech usage decreased in lower levels and the use of touch increased. This was much more dramatic for lower-literate users where more than half of all interactions at lower levels were via touch. All actions at the fifth level (selection of a video) were via touch because there was no speech selection available for videos.

There are several explanations for the overall lower performance of the lower-literate participants. First, several participants with almost no schooling did not seem to fully understand the concept of searching for a video, despite repeated explanations. It was common for these participants to just start describing the whole scenario to the phone when requested to speak. Sometimes this would be phrased as a detailed request for information, and sometimes as a description of how *they* would manage the scenario in their own fields. Second, as indicated by the large number of spoken reminders, those who understood the concept of searching for a video had trouble memorizing the mini-scenario of the task. A third issue was one of vocabulary and terminology. Although the vocabulary used in the application was provided by the local partner NGO, many participants were not familiar with some of the terms. Some of the words used by the NGO were technical or Sanskritised Hindi not used in their dialect. For example, many participants had issues with the expression ”sowing and crop management” (one of the choices at the second navigation level). Indeed, one participant balked after hearing a scenario described (in Hindi!), saying, *”You said it in English, I speak Hindi, how can I understand?”* Vocabulary issues led to a larger number of navigation mistakes as well as more errors in speech detection. Participants would often say only part of the expression, or use a different expression with a similar meaning, which the recognizer did not know. The problem of terminology was generally more frequent with women than with men, probably because men are more likely to attend meetings and video screenings by the NGO than women, and therefore have more experience with this vocabulary.

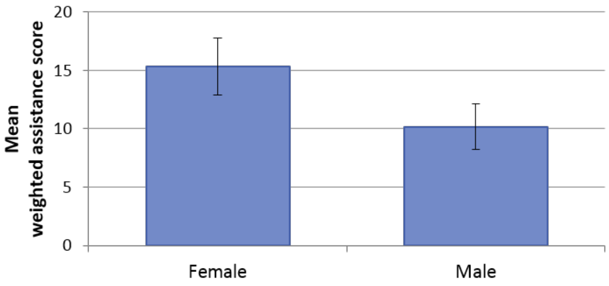


Figure 7. Mean weighted assistance scores (±SEM) by gender.

Gender differences

Indeed, gender was a major source of variability in how participants used the system. Figure 7 shows that women tended to need more assistance than men to complete the tasks, though this difference is not quite statistically significant and is much smaller than the difference due to education. Figure 8

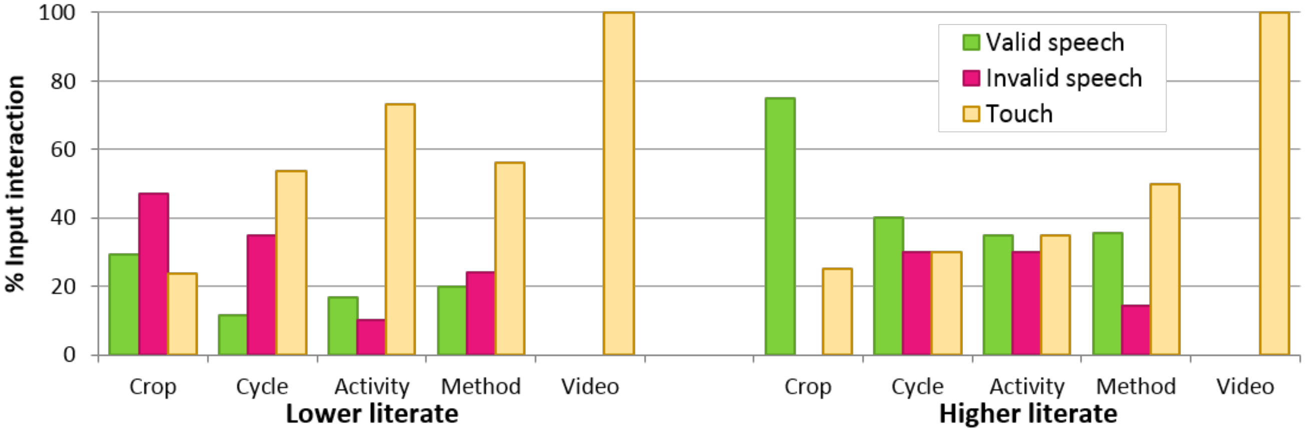


Figure 6. Percentage of interactions at each level for each education group in the STGA condition.

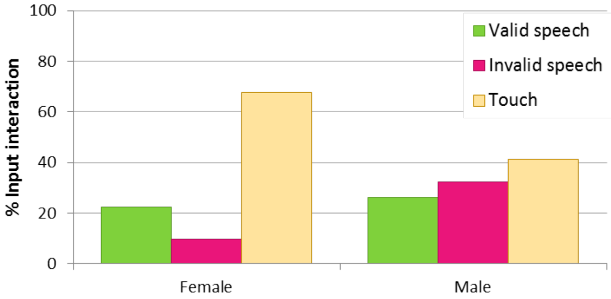


Figure 8. Percentage of interactions by gender for the STGA condition.

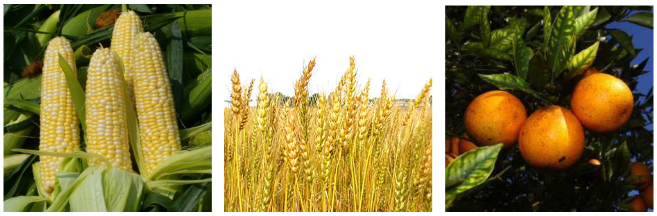
renders the action types performed by females and males as a percentage of all actions taken. Women did not use speech as much as men, but when they did, the ratio of valid words was higher than those not detected, while it was the opposite for men.

Based on our observations and interviews, one possible explanation for women using speech less frequently than men is that women are not used to being the center of attention nor to speaking in front of men. Several of them were shy, did not dare to speak, and when they did, they did not speak loudly. One woman who had used mostly touch said she *”felt shy of speaking because it is my in-laws' house, and there are these people listening outside”*.

Speech usage and performance

All participants in the STGA condition used the speech interface successfully at least once to perform a navigation action. The participants tried to navigate with the graphics in 46% of the cases and used speech for the remaining 54%. All intended touch actions resulted in the expected navigation action, while about half of all attempts at speech did not lead to any response, as shown in Table 2. Most non-detections came from the vocabulary problem mentioned above, from participants speaking before the ”beep” sound, and from some participants not speaking loudly enough. A detailed analysis of the speech interface accuracy is out of the scope of this work and is provided in [1].

Despite the problems with speech, it was the preferred interaction for several higher-literate users, among whom all but one used speech in every task. The participant who did not was from another region and spoke a different dialect; she said she did not feel comfortable using the speech interface because it was not her native language. Of the three others, one used speech exclusively and the two others used speech and *”switched to graphics when the speech did not understand me.”* The one who used only speech said *”I was concentrating on what was being said and if I had started looking for pictures my mind would have been diverted.”* This is consistent with previous research that showed that for novice users, the most efficient method is not necessarily the most preferred one [13].



(a) Maize (b) Wheat (c) Orange

Figure 9. Crop images.

All participants tried to use speech repeatedly, even after encountering rejections. The fact that so many participants were willing to use speech despite many rejections is positive. We interpret it as the novelty effect as well as the ”magic” of being able to speak to a phone in one's own dialect (though one cannot ignore potential demand characteristics of the test, where speech was introduced as a salient feature). Overall, the speech interface was used more and with greater success by higher-literate users. This may be explained by the fact that low-literate participants were often not able to remember the words and therefore used touch and graphics more, particularly at lower levels of the selection hierarchy. We also interpret this difference as a lack of self-confidence in low-literate participants that led them to doubt their speaking, leading to an even poorer recognition by the system. The large number of valid speech interactions at the first level can be explained by three factors. First, the vocabulary problem mentioned above did not affect simple words such as crop names. As one participant reported, *”I said 'orange' because it was a single word. I remembered. It was easy.”* Second, because it was the first choice, crop names came shortly after the end of the scenario explanations, making memorization easier. Finally, crop selection gave users significantly more choices than the other levels (22, as opposed to 4 at the maximum for the following ones). The choices were spread over two screens, making it more difficult for the user to find the right icon to select. For speech navigation, it has been shown that complex menus increase difficulty of task and this complexity can be in terms of long explanatory prompts, difficult or unfamiliar vocabulary, as well as depth and breadth of choices [30]. In this case, although no participant mentioned it explicitly, the vocabulary being short and familiar (the crop name) may have made it easier for the participants to speak than to look for the wanted icon.

Finding the wanted icon

The crops were represented by photographs of either the plant or the fruit of the plant, as shown in Figure 9. Participants, independent of their level of education, had a hard time identifying crops shown with a picture. The crop that caused the most problems was the orange (Figure 9c), with seven out of ten users having trouble finding it. Most of them reported either having been looking for the plant or a green orange (most oranges in this area have green skin even when ripe): *”I could not find the orange icon as the plant does not look like this. You should put unripe, green orange.”*. The hand-drawn icons at the other levels did not lead to as much confusion. As suggested by previous research, this could be because hand-drawn icons are better understood than photographs [18]. However, this could also be because lower levels had a maximum of 4 choices, and because the choices were explicitly mentioned in the audio prompt on navigating to the page.

Understanding and using double touch

One problem we saw in both conditions was with double touch. A first touch on an icon highlighted that icon and played the sound attached to it. Once the icon was highlighted, a second touch triggered the navigation action linked to the icon. Many users had trouble remembering that they had to touch the icon twice and blankly stared at the screen after having touched it once. This is reflected by the fact that for 37% of all touch selections, the user waited for more than five seconds between the two touches to navigate. For 39% of the selections, this time was between two and five seconds, indicating that the users listened to the voice annotation and then touched the icon a second time. About 14% of all navigations happened in less than one second, suggesting that most participants listened to the voice annotation information. This is in line with previous research that has underscored the importance of voice annotation [18, 23, 25].

Desire for speech

As noted above, participants were very enthusiastic about using speech — both talking to the device and hearing it talk back to them. Four participants had the natural tendency to repeat to themselves the choices that were prompted by the application, or to simply repeat only the choice they had selected. Two participants greeted the system back with a *"Namaste!"* when it greeted them at the start of the application. This confirms that speech can be perceived as a natural way of communication even with machines. This is particularly true for mobile phones, for which voice is the primary means of interaction (even if this is usually with a person on the line).

DISCUSSION

Both with and without speech input, we saw similar outcomes when farmers tried to use our system. One major finding was the impact of education: whether or not participants had speech input available, farmers with very little education had trouble using VideoKheti. The difficulties encountered by low-literate users included problems with understanding and remembering scenarios; vocabulary comprehension and reproduction; and even understanding the hierarchical organization of information. One argument that is often made in the literature (e.g., [25]) is that speech interfaces could help overcome low-literacy-related issues and provide universal access. Our study revealed that while participants indeed were enthusiastic to talk to the system, the speech interface did not manage to overcome many of the barriers linked to low-literacy. Indeed, it seems that potential benefits of adding speech to the system were outnumbered by the cognitive overload of adding yet another thing for our participants to remember and deal with. The choices of vocabulary were still lacking user-centrality and there was a limitation of words in the ASR that did not relate to the context of the farmers, thus making the system not ready for the current skill level at hand. Recent work studying low-literate users suggests that many of the problems we saw are related less to literacy *per se*, but rather due to a variety of cognitive skills that are also learned at school [16]. It is also important to note the strong correlation between age and education for these users; younger (and more literate) users were better able to use the system. There were also cultural factors at play that inhibited use, particularly among women, who were reluctant to speak in front of men or their in-laws.

It is also likely that the lack of naturalness in the speech interaction with VideoKheti is partially responsible. The dialogue was initiated by the system, forcing the participants to adapt to its rhythm rather than their own. The ASR only recognized a small number of words, and worse, did not recognize partial sentences or similar but not identical expressions. Participants whose speech was not recognized lost self-confidence and became flustered by a system that did not understand them. This has been observed before [2]. Ideally in such cases, users should have switched to touch and graphics. This happened in several cases, but mostly for higher-literate users. A future research topic would be to see whether switching to the graphics when speech fails becomes more natural as the users get acquainted with the system.

A speech interface is therefore not a miracle remedy against low-literacy issues. However, the choice of speech over graphics to pick the crops and the higher success of speech recognition in this case suggests that a speech interface can work well when there are many choices available, and those choices are represented by short and familiar words. This result, obtained with a touch interface, is in line with what Patel et al. [24] had hypothesized, without being able to show it, when comparing voice command with touchtone input. More generally, we observed that participants liked to use the speech interface even if it was not as reliable as the touch interface. This is in line with previous research on multimodal interfaces that showed that allowing the users to choose their preferred style of interaction increased user satisfaction [22], even if it is not the most efficient one [13].

The initial motivation for this work was to develop an application that would provide easier access to videos for farmers. While the results of the field study show that farmers with 5 years of schooling or less had trouble using the system, they also show that higher-literate farmers were able to use it effectively, with or without speech input. This is an extremely encouraging result because it opens up opportunities to give direct access to the videos to at least some of the farmers. In the context of Digital Green, the application could be used by the mediator and lead to simplified logistics.

CONCLUSIONS AND FUTURE WORK

The goal of this work was to see whether rural Indian farmers with little exposure to technology could use a multimodal system including speech, touch, graphics and audio output to find agricultural extension videos. While we were impressed by how well the SALAAM ASR performed in this context, it was not clear that adding speech to the interface was that useful. When it was available, participants were willing to use speech and all of them managed to use it at least once successfully. Despite the difficulties with it, participants did appear to like the speech interface, especially when the list of choices was long and the choices were short, familiar words. However, the usability of the system was highly correlated with the education level of the users irrespective of whether speech was available; low-literate users had much more difficulty than their more educated peers. Counter to our expectations, the speech interface was not able to overcome the issues related to a lack of education.

A key question raised by these results is whether using a limited-vocabulary system (such as the SALAAM technique used here) provides enough benefit to offset the cost and difficulty of implementing and using it. We interpret these results as mixed: it is clear that users want to use speech, and they are willing to endure a relatively inflexible system to do so. Particularly in cases where there are many choices available and the vocabulary comprises short, familiar words, a speech system like ours may be worth the effort. However, this decision should be tempered by other realities as well. At the time of our testing, there was no wireless broadband available in the villages we were working in. Until infrastructural constraints such as this are resolved, the challenges for implementing speech systems are only amplified. Further, it seems clear that a limited vocabulary ASR such as we implemented is only useful in a very constrained domain using short, familiar words. Our multimodal speech interface using SALAAM is not a clear winner for low-literate users.

While one can learn a good deal through the kind of prototype testing we did here, we are very interested in understanding how users might adapt to a system like VideoKheti over time. The interactions in our system were extremely novel for all our users; in time, we believe that users will adapt to the novelty and begin using the system as it is intended to be used (or not!). In particular, we are interested in understanding how practice and exposure might affect our low-literate users. Will they adapt and become experts, or will they simply depend on their more educated (junior) peers to access the information for them? Hopefully, by working with Digital Green and other partners, we will be able to parlay our findings into robust systems for making valuable information available to people such as the farmers we were working with.

ACKNOWLEDGMENTS

We would like to thank Digital Green for their support and collaboration in this project. Thanks to Roni Rosenfeld, Hao Yee Chan, and Sunayana Sitaram for the help with the SALAAM algorithm, and to Bill Thies for the many useful suggestions throughout the project.

REFERENCES

1. Bali, K., Sitaram, S., Cuendet, S., and Medhi, I. A hindi speech recognizer for an agricultural video search application. In *ACM Dev'13: Annual Symposium on Computing for Development Proceedings* (2013).
2. Boyce, S., and Gorin, A. User interface issues for natural spoken dialog systems. *Proc. ISSD 96* (1996), 65–68.
3. Boyera, S. The mobile web to bridge the digital divide. *ISTAfrica Conference* (2007).
4. Chipchase, J. Understanding non-literacy as a barrier to mobile phone communication. Tech. Rep. June 17, Nokia Research, 2005.
5. Cohen, P., and Oviatt, S. The role of voice input for human-machine communication. *Proc. the National Academy of Sciences 92*, 22 (1995), 9921–9927.
6. Gandhi, R., Veeraraghavan, R., Toyama, K., and Ramprasad, V. Digital green: Participatory video for agricultural extension. In *Proc. ICTD* (2007).
7. Grisedale, S., Graves, M., and Grnsteidl, A. Designing a graphical user interface for healthcare workers in rural india. In *Proc. CHI*, ACM (1997), 471–478.
8. Huenerfauth, M. P. *Developing Design Recommendations for Computer Interfaces Accessible to Illiterate Users*. PhD thesis, University of Pennsylvania, 2002.
9. Johnston, M., Bangalore, S., Vasireddy, G., Stent, A., Ehlen, P., Walker, M., Whittaker, S., and Maloor, P. MATCH: an architecture for multimodal dialogue systems. In *Proc. ACL*, Association for Computational Linguistics (2002), 376–383.
10. Kumar, A., Agarwal, S. K., and Manwani, P. The spoken web application framework. In *Proc. W4A*, ACM Press (2010).
11. Kumar, A., Reddy, P., Tewari, A., Agrawal, R., and Kam, M. Improving literacy in developing countries using speech recognition-supported games on mobile devices. In *Proc. CHI*, ACM (2012), 1149–1158.
12. Lamel, L., Rosset, S., and Gauvain, J.-l. Considerations in the design and evaluation of spoken language dialog systems. In *In Proc. ICSLP* (2000).
13. Lee, K. M., and Lai, J. 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).
14. Lewis, M. P. *Ethnologue: Languages of the World*, 16th ed. SIL International, 2009.
15. Maunder, A., Marsden, G., and Harper, R. Making the link-providing mobile media for novice communities in the developing world. *Int. J. Hum.-Comput. Stud. 69*, 10 (2011), 647–657.
16. Medhi, I., Menon, S. R., Cutrell, E., and Toyama, K. Correlation between limited education and transfer of learning. *ITID* (June 2012), 51–65.
17. Medhi, I., Patnaik, S., Brunskill, E., Gautama, S. N., Thies, W., and Toyama, K. Designing mobile interfaces for novice and low-literacy users. *Proc. ToCHI*, 1 (2011).
18. Medhi, I., Prasad, A., and Toyama, K. Optimal audio-visual representations for illiterate users. In *Proc. WWW* (2007).
19. Medhi, I., Sagar, A., and Toyama, K. Text-free user interfaces for illiterate and semiliterate users. *ITID 4*, 1 (Oct. 2007), 37–50.
20. Medhi, I., and Toyama, K. Full-context videos for first-time, non-literate PC users. In *Proc. ICTD* (2007), 1–9.
21. Moore, R. K. A comparison of the data requirements of automatic speech recognition systems and human listeners. In *Proc. Eurospeech, Geneva* (2003), 2582–2584.
22. Oviatt, S. Multimodal interactive maps: designing for human performance. *HCI*, 1 (Mar. 1997), 93–129.
23. Parikh, T., Ghosh, K., and Chavan, A. Design studies for a financial management system for micro-credit groups in rural india. In *Proc. CUU*, ACM (2003), 15–22.
24. Patel, N., Chittamuru, D., Jain, A., Dave, P., and Parikh, T. S. Avaaj otalo: a field study of an interactive voice forum for small farmers in rural india. In *Proc CHI*, ACM (2010), 733742.
25. Plauche, M., Nallasamy, U., Pal, J., Wooters, C., and Ramachandran, D. Speech recognition for illiterate access to information and technology. In *Proc. ICTD* (2006), 83–92.
26. Qiao, F., Sherwani, J., and Rosenfeld, R. Small-vocabulary speech recognition for resource-scarce languages. In *Proc. DEV*, ACM Press (2010).
27. Rahim, M., Fabbrizio, G. D., Kamm, C., Walker, M., Pokrovsky, A., Ruscitti, P., Levin, E., Lee, S., Syrdal, A. K., and Schlosser, K. VOICE-IF: a mixed-initiative spoken dialogue system for AT&T conference services. In *Proc. Eurospeech* (2001).
28. Sherwani, J. *Speech Interfaces for Information Access by Low Literate Users*. PhD thesis, CMU, May 2009.
29. Union, I. T. Measuring the information society 2011. Tech. rep., International Telecommunication Union, 2011.
30. Whittaker, S., Hirschberg, J., and Nakatani, C. H. Play it again: a study of the factors underlying speech browsing behavior. In *Proc. CHI*, ACM (1998), 247–248.