8-22-2008

Speech vs. Touch-tone: Telephony Interfaces for Information Access by Low Literate Users

Roni Rosenfeld
Carnegie Mellon University

Jahanzeb Sherwani
Carnegie Mellon University

Sooraj Palijo
Health and Nutrition Development Society

Sarwat Mirza
Health and Nutrition Development Society

Tanveer Ahmed
Health and Nutrition Development Society

See next page for additional authors

Speech vs. Touch-tone: Telephony Interfaces for Information Access by Low Literate Users

***Abstract*—Information access by low literate users is a hard problem. Critical information, such as in the field of healthcare, can often mean the difference between life and death. In our research, we have developed and tested various spoken language interface prototypes with low literate community health worker. In this paper, we present results from our research that show that well-designed speech interfaces are preferable to touch-tone equivalents. Additionally, we show that it is especially important to localize the system's spoken language output for low literate users.**

***Index Terms*—Speech interfaces, low literate, information access, community health workers.**

I. Introduction

Low literate users face great difficulty in accessing information that is often easily available to literate users. This is especially problematic in the developing world, where there are many more non literate users, and where the importance of information is often greater. One domain where information is especially important is healthcare.

Healthcare is a fundamental, yet often under-serviced need of citizens in developing countries. These regions have the highest maternal mortality and neonatal mortality ratios in the world, and, not surprisingly, also have the largest unmet need for health service providers in the world. Given the high cost of training doctors and nurses, and the low number of medical schools in these parts of the world, many governments have begun community health worker (CHW) programs, where people (usually women) are chosen from their own communities, trained in basic health service provision for a few months, and sent back to provide health services in their communities. In some countries, especially in Latin America, their effectiveness is quite high, reducing infant mortality to below that of the US. These CHWs vary greatly in literacy levels and receive little refresher training. It is not surprising that the need for better information access by CHWs is widely agreed upon: "Providing access to reliable health information for health workers in developing countries is potentially the single most cost effective and achievable strategy for sustainable improvement in health care" [1].

Over the past three years, we have been researching spoken language interfaces for information access by low literate community health workers in Pakistan. We conducted a number of pilot user studies testing speech interface prototypes in Urdu and Sindhi, in different urban and rural sites, with community health workers of varying literacy.

In this paper, we present various findings from these studies:

- We describe why reinforcing existing practice is not always the best strategy, in the context of the choice of content to be provided in a prototype information access interface, for the purposes of usability testing (Section II).
- We outline a novel approach on how to significantly improve speech recognition accuracy for local languages using a standard US English speech recognizer (Section IV).
- We describe our novel solution for a mobile user study lab for speech interface research, and argue that having such a modifiable system available at the field site is essential for rapid iterative development with participatory design (Section V).
- We present lessons learnt from a number of pilot experiments in various urban and rural field sites (Section VI).
- We describe a novel method for teaching speech systems to participants effectively and quickly, and show that an effective tutorial is essential when conducting a user study on a speech interface. (Section VII).
- We present both qualitative and quantitative results from our user studies, which suggest that participants who find a given system design difficult would prefer an interface with a touch-tone input modality to an equivalent interface with speech input. Moreover, we show that it is harder for low literate participants to effectively use systems that aren't perfectly localized to their cultural context (Section VIII).
- Finally, we suggest why low literate users prefer touch-tone systems when the content has not been localized to their context (Section IX).

II. Health Information Content

Based on our prior ethnographic research, we had initially identified specific health topics on which to provide information through any automated interface [anonymized]. However, our prior work was focused on urban community health workers with a minimum of 8 years of education. Since that time, we have shifted to focus on lower literate, rural community health workers. In collaboration with our partner NGO, we initially opted to work with reinforcing the material that the health workers were trained on (maternal and

Manuscript received August 22, 2008.

reproductive health), which is what an eventual deployed system would need to provide. Additionally, this seemed the prudent choice, as it is preferable to reinforce existing systems and practice than to create new ones. The following issues forced us to rethink this approach:

1. **For the participant**: In a user study, even though we clearly stated that "this is not a test of your knowledge", especially when participants are tested on information they are supposed to already know, they believe that it is a test of their knowledge. In our experience, when participants were unable to give answers that they felt they should have known from before, they felt embarrassed and uncomfortable.
2. **For the researcher**: It is impossible to tell whether a response to a question-answer task is being given based on what the participant found through the system, or from prior knowledge. One way to cope with this issue is to conduct a pre-test of their knowledge, but this would further conflict with the previous issue.
3. **For both**: Reproductive health issues are extremely taboo in Pakistani society, and are never discussed in the presence of males. As the primary author (a male) needed to be present during the user studies, this presented a source of discomfort for user study participants (e.g., they sometimes leaned in to give a response privately to the female facilitator).

Based on the above issues in our pilot tests, we have now shifted to working with content that the community health workers have *not* been trained on before, without any taboo elements in it.

III. Telephony Interfaces for Information Access

To provide the information identified above, we have built two primary telephony interfaces that we have tested extensively. The first is a purely non-interactive system, which plays back a specific audio clip from beginning to end. This was primarily created as a baseline, to assess the cognitive load on the participants created by the length of the speech segment.

The second interface is menu-based. It asks the user to select a given topic (e.g., malaria, diarrhea, or hepatitis), after which they are asked to choose from a specific sub-topic (e.g. general information, signs, preventative measures, treatment), after which they are given detailed content broken down into chunks of three bullet points at a time. The interface was created in two 'flavors': one using touch tone input for choosing between the options, and the other using speech input. Here is a sample call for both flavors, translated from Sindhi:

Speech	Touch-tone
Hello, I'm Dr Marvi, and I'm here to give you health information.	
What would you like to hear about? Malaria, Diarrhea, or Hepatitis?	For information on Malaria, press 2, for information on Diarrhea, press 3, and for information on Hepatitis, press 4.
User says Diarrhea	*User presses 3*
Diarrhea. If this isn't the topic you want, say 'other topic'. [Pause]	Diarrhea. If this isn't the topic you want, press 0. [Pause]
Let me tell you about Diarrhea. As a Marvi worker, you need to know that Diarrhea is a dangerous disease that can potentially be life threatening. You should know about its causes, its signs, its treatment, and how to prevent it.	
What would you like to learn about: causes, signs, treatment, or prevention? [Pause] To learn about a different topic, say 'other topic'.	To learn about the causes of diarrhea, press 2. To learn about the signs of diarrhea, press 3. To learn how to treat diarrhea, press 4. And to learn how to prevent diarrhea, press 5. [Pause] To learn about a different topic, press 0.
User says 'causes'	*User presses 2*
The causes of Diarrhea. If this is not the topic you want, say 'other topic'. [Pause]	The causes of Diarrhea. If this is not the topic you want, press 0. [Pause]
Let me tell you about the causes of Diarrhea… [gives 3 bullet points on the topic].	
To hear this again, say 'repeat'. To hear more, say 'more information'.	To hear this again, press 1. To hear more, press 2.
User says 'more information'	*User presses 2*
[The system gives 3 more bullets on the topic, and this cycle continues until there are no more bullets, at which point the following instructions are given.]	
To hear this again, say 'repeat'. For a different topic, say 'other topic'.	To hear this again, press 1. For a different topic, press 0.

IV. Improved "Poor Man's Speech Recognizer"

For speech recognition, we previously described a "poor man's speech recognizer" [anonymized], using a robust speech recognizer trained on US English speech. The basic principle of the approach is to map between phonemes in the desired language (Sindhi in our case) and the trained language (US English). Thus a word such as 'wadheek maaloomaat' (transliterated Sindhi) would be given the following US English phonetic pronunciation: W AH D I K M AA L U M AA DH. In our initially described approach, the choice of phonemes was left solely to the discretion of a language expert. We tested this approach with Microsoft Speech Server (MSS), although the principle would work with any modern

speech recognition system. This approach led to reasonable recognition rates, although it was not very robust, and prone to error when tested in the field.

We have improved upon this approach significantly by incorporating a data-driven approach. While the details of this approach are outside the scope of this paper, the basic idea is to enable the developer to test recognition accuracy based on varying any subset of phonemes in a given word's pronunciation definition. For instance, if the developer is unsure of the optimal choice for the last consonant in "maaloomaat", she could specify a wildcard definition of "M AA L U M AA C?", where the "C?" denotes an "any consontant" wildcard. Similarly, if the developer wants to test the optimal phoneme choice for the final consonant-and-vowel combination in "bachao", she may specify "B AX C? V?". An intermediate processing step would convert these pronunciation entries into a speech recognition grammar consisting of all possible pronunciations with that wildcard. Thus, if there are a total of 20 consonants, "M AA L U M AA C?" would be transformed into a list of 20 words, each with a unique final consonant. This grammar is then used to run a re-recognition pass over any sample(s) utterances of the given word, and the best matched pronunciations would then be manually chosen by the user to be used as the optimal pronunciations in the final system.

This is a simple problem if there are a few wildcards. However, if there are multiple wildcards in the same entry, the combinatorial explosion would make it difficult for the speech recognizer to work with such a large grammar. For instance, if the developer was to try the entry "M V? L V? M V? C?", if there are 20 total vowels and 20 total consonants, this would result in a 20*20*20*20 = 160,000 word grammar, which might be computationally intractable to run recognition on. In our experiments with MSS, such large grammars did not return recognition results even after 10 minutes on one word. A heuristic to solve this problem is to allow the developer to create arbitrary word boundaries, which would reduce the number of combinations in the final grammar. For instance, "M V? L V? / M V? C?" would result in a 20*20 + 20*20 = 800 word grammar, which is much quicker to compute. While the final result isn't as accurate as with the full grammar, it is close to the optimal answer, and works significantly faster (less than a few seconds for a recognition result with MSS). Preliminary results using this improved approach are described in Section VIII.

V. Mobile User Studies

In our initial work, our prototype interface was running on a server physically located in Karachi, accessible over the telephone line connected to a separate telephony server. Physically, this consisted of:

- Windows server running Microsoft Speech Server, containing all the logic for the information access interfaces, also running a Voice-over-IP gateway
- Linux server running Asterisk/Trixbox for Voice-over-IP support
- Uninterrupted Power Supply (UPS) unit as backup in case of power failure
- Monitors, keyboards, mice, routers, and network/power cables

While this worked to some extent, it had the following problems:

- Any power outage lasting longer than the maximum UPS backup time could potentially bring the system down. Running a Windows server for the speech components, and a Linux server for the telephony interface meant a high electrical load.
- Any modifications to the system could not be made at the field site (often a health center) – they would have to be made in the city, away from the actual users. This did not facilitate iterative design with short feedback loops, nor did it enable participatory design.
- Any software/hardware failure would require trained and available personnel at the server site. This was not always possible.
- For extended field research, the above problems were compounded, and it became very unlikely for there *not* to be a problem
- The phone line was also prone to temporary blackouts, sometimes for days on end
- It was difficult to physically move the entire infrastructure to a remote field site, and such a move would not solve the power problems, nor the phone problem – in fact, a new phone line would have had to be provisioned, which could have taken months

Based on the above observations, experiences and constraints, we realized the need for a mobile user study setup, where the actual system would be physically accessible in the field, without the power and telephony issues. This led to the following setup:

- Laptop running Windows with Microsoft Speech Server, along with the Voice-over-IP gateway
- Linksys SPA3102 device (around the size of a 4-port network hub) connected to the laptop through one network cable, and connected to a telephone set through a standard phone cable
- Power for the two devices

Given the low power requirements for these two devices, we were able to get much longer backup times using the same UPS. Further, the portability of the setup meant it was simple to take it to any field site. Finally, interoperating with an actual telephone set meant that we maintained the same physical interface as before, but removed all the intermediary components that were prone to failure. We tested this system in our final user study, and it worked without a problem.

VI. Pilot Studies

A. Description

We conducted a number of pilot user studies over the past year, as described below:

Month	Place	Avg. Education	Language
Jan	Outskirts of Karachi	5-10 years	Urdu
Mar	Umarkot (rural)	<5 years	Urdu
Jun	Dadu (rural)	<5 years	Sindhi

In these studies, we tested the relative effectiveness of printed text against the baseline speech system as described in Section III. The system would only play back audio comprising of a Sindhi speaker reading out the text material verbatim. Users were given an information access task (e.g. name any one danger sign during pregnancy), and were then either given the relevant page (e.g. containing a list of danger signs during pregnancy) or played back the relevant audio clip on the telephone, to answer the question.

These experiments were meant primarily to validate the content we had chosen (including the choice of language), as well as to provide a baseline on which further work could be measured against.

B. Findings

Information presented orally needs to be short. Both low literate and literate users found it hard to hear long passages of text with the purpose of extracting small nuggets of information. When the length of passages were varied (a few sentences, to a page, to a pamphlet), the task became progressively more difficult.

Low literate users were less likely to have ever used a phone. Also, low literate users were more hesitant when picking up the phone (more likely to ask for permission), and were more likely to hold the phone with the mouthpiece too high or too low.

The national language is not always optimal. Initially, our partners had told us that Urdu (the official language) was a language that "most" of the target users would be familiar with and that it would be an acceptable choice for the system. The pilot studies showed that Urdu was not understood at all by 50% of the participants in Umarkot, and 66% of those in Dadu. Of the remaining participants, many still had difficulty since they were not completely familiar with Urdu.

The regional language is also not always optimal. Based on our prior experience, we tested Sindhi content (text and speech) in a rural health center in Dadu district (part of the Sindh province). However, our participants all belonged to migrant communities from Balochistan, and were native speakers of a minority dialect of Balochi without any written form. Thus, only those participants who had been to school had any knowledge of Sindhi (30% of the participants). The remaining participants understood Sindhi to varying degrees, with the more educated ones having a better grasp.

Subjective feedback needs triangulation. When the non-Sindhi speaking participants were asked if they would prefer a system in Balochi, none of them replied that they would – instead saying that the Sindhi system was fine the way it was. This was surprising, as they had not succeeded in any of the given tasks. Further probing and questioning showed that each had a different reason (however valid) for saying this – one said it due to peer pressure, thinking that the others would "blame" her as the reason why the system was not made in Sindhi. Another participant said that she assumed we were talking about official Balochi (unintelligible to speakers of their minority dialect), and said she would prefer a system if it were in *her* Balochi. This reinforces the need to triangulate all subjective feedback in ICTD research, as the sociocultural complexities inherent in such work are impossible to predict and account for in advance.

Speech may be preferable to text, even for a baseline system. 60% of the participants in the Dadu study said they preferred the speech system, while 40% said that both speech and text were equal. No participant expressed a preference for text. Based on the previous point, we must take this with a grain of salt – however, it is expected that users with limited literacy would prefer a system which doesn't require reading. Also, there was no statistically significant difference in task success for these conditions in any of the studies – but it is important to note that the speech system was purposefully poorly designed as it was a baseline system without any interactivity.

Training and working with local facilitators is essential. Over the course of these studies, we worked with user study conductors from the city as well as from the locality in which the research was conducted. While the local facilitators took more of an effort to train (requiring personalized attention, instead of assigned readings), they were much more effective in the user study process. Primarily, they were able to communicate very effectively with participants throughout the study, and were able to understand and translate their issues and feedback clearly to the research team. Additionally, they had deep knowledge of the community, the local context, and of the specific participants as well – so were able to think of issues before they happened, and were also able to provide extra information on past events when needed. Finally, the linguistic diversity (Sindhi and Balochi) that was required for the Dadu study meant that anyone other than a local community resident would not have been able to communicate effectively with all participants. Thus, we strongly recommend training and working with local facilitators for user studies.

VII. Formal User Study Design

In September 2008, we conducted a within-subjects user study testing the speech and touch-tone flavors of the menu-based system described in Section III. The user study was conducted in Umarkot, Sindh, at a training center for

community health workers. Participants were recruited through our partner NGO, and came from Umarkot and a nearby town, Samarro. Prior to the actual user study, we conducted a pre-study pilot with 3 participants a day before the study began.

A. Pre-study Pilot

Our initial design was as follows. Participants would be introduced to the broad goals of the study, and the steps involved. Their verbal consent would be requested. Personal information would first be collected, including telephone use, educational history, and a short literacy test where the participant would read out a standard passage and subjectively rated by the facilitator. They would then be verbally introduced to either flavor of the system (touch-tone or speech), and given a *tutorial*. After the tutorial, they would be given three *tasks*, with increasing complexity, on one disease. After this they would be introduced and taught the other system, and would then be given three similar *tasks* on another disease. At the end of the tasks, they would be given a series of Likert scale[1] questions to subjectively rate the systems on their own and in comparison with one another. Finally, the researcher and facilitator would conduct a short unstructured interview based on the participants' experience in the user study.

The *tutorial* for both flavors of the system consisted of three steps. In the first step, the participant would listen in (using earphones connected to an audio-tap [2]) on the facilitator using the system to complete a task. The facilitator would purposefully make a mistake (choosing the wrong disease) and would then correct it, and successfully complete the task. In the second step, the participant would be given a task to complete, while the facilitator would listen in, giving advice if the participant had any trouble. In the third and final step, the participant would be given 5 minutes to use the system as she pleased.

The three *tasks* were roughly equivalent for both systems. The first task was general: "name any of the signs of disease X". The second task was specific: "how many people are affected by disease X every year?" The third task was very complex, e.g., "is coughing a sign of Hepatitis?" – note that the answer for the third task was always no, meaning that the user would have to listen through all the signs for the disease, and would then need to deduce that since they did not hear it, it is not a sign.

Our findings from this pre-study pilot, covering three participants, were as follows:

- **An effective tutorial is essential**. Our tutorial did not teach participants how to use either system well. They were not able to complete the second task (on their own) effectively, and the 5 minute free-form practice was not helpful either. Thus, their performance on the actual tasks was abysmal, as they were not able to even navigate through the system effectively on the given tasks, much less answer the questions correctly. It was evident that we needed a better tutorial.
- **The tasks were possibly too difficult**. Although it is uncertain whether this was due to the problematic tutorial, participants in the pilot were not able to succeed in any of the given tasks, being especially unprepared for the second and third tasks (the moderately difficult and difficult tasks).
- **The tasks were possibly too abstract**. It is well known that low literate users have difficulty with abstract thinking [16]. Even the task of asking a question without any context (e.g. naming any symptom of a disease) is an abstract task.

B. Changes to the Study Design

Based on the above observations, we made some modifications to the user study design.

The tutorial process was increased to three steps instead of two. The "free-style" 5 minutes were removed, and each step was focused around a specific task. Further, each of the tasks was carried out by the participant, with the facilitator listening in on each task. We used a "training-wheels" learning style, with the training wheels gradually coming off after each task. The participant was also informed that the facilitator would give explicit instructions on all actions for the first task, less help on the second task, and almost no help (unless they were stuck) on the third task.

The tasks themselves were shortened (to make up for the lengthened tutorial step) to two instead of three. These two were also made easier – with both tasks asking a "name any X of disease Y" form question, where X was one of: sign, prevention method, treatment method, cause, and Y was either Malaria or Hepatitis.

Finally, we thought it may be pertinent to concretize the tasks by using the Bollywood Method [2]. In the Bollywood Method, user study tasks are given a dramatic and exaggerated back-story to excite the user into believing the urgency of the problem. We decided to apply this method to only the first of each pair of tasks. Thus, the tasks wwere given a back-story along the lines of: "Disease X has become prevalent in Khatoon's neighborhood. She is worried about catching the disease and wants to know of any one method to prevent the disease. Using the system, find out any one method for prevention of disease X".

After making the above design changes, we conducted the formal study. We requested Sindhi-speaking participants, and worked with 9 participants over 3 days. The order of presentation of either flavor of the system was counterbalanced.

[1] A standard tool used to elicit subjective feedback from participants. Participants are asked how strongly they agree or disagree with a given statement, by choosing a number, say 1 through 5, to represent their level of agreement. In our work, we adapted this tool for verbal presentation, and used a 3-point scale.

[2] Also known as a Telephone Handset Audio Tap, or THAT.

VIII. Results

Of the 9 participants, one was not able to speak Sindhi at all, and was unable to complete any of the tasks successfully – her data were removed from the final analysis.

A. Personal Information

Language: Of the remaining 8 participants, 7 self-identified as either speaking Sindhi natively, or speaking it in their home as a second language. The other participant was a native speaker of Seraiki, but had 8 years of education which had taught her Sindhi well.

Age: The average age was 21.7 years, with a maximum of 32 and a minimum of 17.

Years in School & Reading Ability: The average number of years in school was 7, with a minimum of 3 and a maximum of 12. 2 participants were completely unable to read Sindhi, 1 was able to read with great difficulty, 3 were able to read with some difficulty, and 2 were able to read fluently. For the purpose of the analysis, the first two categories will collectively be referred to as 'low literate' participants, while the last two are the 'literate' participants. Thus, there were 3 low literate participants, and 5 literate ones.

Telephone use: All participants had used telephones prior to the study, although they varied in how frequently they used them. The minimum number of self-reported uses of a telephone was 0 for one user, with a maximum of "daily" for three users.

B. Quantitative and Qualitative Results

Task success in the speech interface was significantly better than in the touch-tone interface. There was a significant effect for the interface type, $t(35) = 4.86$, $p < 0.05$, where the mean number of tasks successfully completed was 83% in the speech condition, and 50% in the touch-tone condition.

More users preferred speech to touch-tone; however, low literate users preferred touch-tone to speech. 5 users preferred the speech interface, while 3 preferred the touch-tone system. It is very significant that all three low literate users were the ones who stated a preference for touch-tone, while all five literate users preferred the speech interface. Their ratings for ease-of-use followed a similar pattern – the low literate participants were the only ones who rated the touch-tone interface easier than the speech interface.

An improved tutorial method worked well. All users were able to complete all of the tutorial steps, even though some took up to 3 tries on one task to get the correct answer. The problems they faced in initial practice tasks were successively corrected over the course of the three practice tasks, such that by the time they began the actual tasks, they were much better prepared to use the interfaces than in the pilot.

Low-literate users expressed difficulty understanding the spoken language output from both interfaces. This was expressed only in the semi-structured interview at the end, when asked what main difficulties they faced. P9, for instance, said she understood the facilitator perfectly well, but didn't understand the majority of what the system said. During her tasks, it was evident that she wasn't able to understand the instructions given to her by either system – as she was waiting without giving any input on certain prompts for up to 20 seconds at a time before hanging up. On further inquiry, it turned out that while P9 was a native speaker of Sindhi, her dialect of Sindhi (and in fact, the Umarkot dialect of Sindhi) is different from the "official" Sindhi that the system's voice was recorded in. This includes both the accent as well as the word content – some words are significantly different in the local dialect. Additionally, the content included some Urdu words, which completely threw off the low literate participants. However, it was difficult to get the participants to explain what they found problematic, as they tended to blame themselves for the problems they faced, rather than blaming the system, or the designers of the system, for creating a system that didn't match her language skills. Finally, it is important to note that when asked if her preference would change if the system was made in her language, P9 (the only low literate user to be asked this question) said that she would prefer the speech interface if both interfaces had been in her language.

Literate users said that the speech system required them to remember less. When asked why they preferred the speech system, the literate users responded that with the button system, they had to remember both the topic they were looking for, as well as the number they would need to press to get it. In some tasks they weren't sure what the correct label was (e.g., when hearing the list of options in the task for naming a preventative method for Hepatitis, there was an initial topic titled "methods of transmission", with the title "methods of prevention" coming later – the first topic was a potentially correct choice), and so they would have to remember two discrete bits of information for any option in the touch-tone case.

Speech recognition accuracy was very high. While earlier experiments with the "poor man's speech recognizer approach" had mediocre accuracy (around 50%), with the improvements described in Section III, the recognizer's accuracy was 91% for the first 6 participants (the other 3 are currently being transcribed). Specifically, there were 150 total utterances, of which 133 were in-grammar (i.e., the user said something that the recognizer should have been able to recognize), and 17 were out-of-grammar. For the in-grammar utterances, 121 were correctly recognized, giving an accuracy of 121/133 = 91%. Further, of the 12 errors, only 2 were misrecognitions, while 10 were non-recognitions. Non-recognitions are significantly easier to deal with, as the system says "I didn't understand what you said, please repeat that…" followed by a list of valid options. Misrecognitions are harder to recover from, as they result in the system confidently (yet incorrectly) assuming that the user said something else, and moves the dialog in the wrong direction (e.g., the user says "Diarrhea", but the system hears "Malaria", and takes the user to information on Malaria). Finally, of the 10 non-recognition errors, 4 were due to acoustic issues caused by the telephony interface, which may be solved by tuning the parameters of the

telephony interface device.

IX. Discussion

A. Literacy and Preference for Speech vs. Touch-Tone

One of the surprising, and seemingly contradictory findings in the above results is that literate users said that they had to remember less in the speech interface and preferred it, yet low literate users said that the speech interface was harder, and preferred the touch-tone one. Did the low literate users have to remember more when they were using the speech interface? Or were there other factors at play in the difficulties they faced in their use of both systems?

One of the frequently re-occurring themes in our research is that low literacy involves more than just the inability to read and write. Low literacy is the result of less schooling, and the experience of schooling imparts various skills beyond the mechanics of parsing written text, such as learning how to learn, learning the process of abstract thinking, learning to trust forms of knowledge other than experience-based knowledge, and even learning other languages, dialects and accents.

Thus, low literate participants are less likely to be exposed to alternative dialects, or to other languages, and would find the linguistically-diverse system's output more challenging than the literate participants. Anecdotally, when an interface's output is not very intelligible, it seems preferable to use a touch-tone input modality, than a speech input, as speech requires the user to expose their confusion more publicly. When faced with the same prompt in both systems, P9 tried different buttons in the touch-tone interface, but did not speak at all in the speech interface.

The lesson, then, is that when designing a system for low literate users, it is essential to choose both the language content and the speaker (whose voice will speak that content) based on the local spoken dialect of the target user population. If there are multiple languages and dialects within the group of intended users, the system may need to be designed with multiple language support if low literate users are part of the user group. Further, any testing of the system must ensure that low literate users are adequately represented, as their experience of any system is qualitatively and quantitatively different from that of literate users, as shown by our research.

Finally, this also suggests that low literate users' preference for touch-tone systems doesn't necessarily imply that such systems are better in all cases – it may be that by improving the speech output of the system, that the preferred modality changes, as a result of interaction effects between literacy, comprehension, and preference between the two modalities.

B. Literacy and User Study Design

Building on the previous point, it is important to note that user study methodologies have been developed primarily with Western, literate participants in mind. Likert scales require the respondent to visually read and respond to the questions. User study instructions are recommended to be given uniformly, by reading aloud from a script – which is very foreboding and artificial sounding for a low literate user. Finally, the act of asking an abstract question (e.g., name any one sign of Diarrhea) and expecting an answer is also abstract, and would be harder for a low literate participant than a literate, schooled participant. While some work has been done in this space (Likert mood knobs, Bollywood Method for task specification [2]), these methods have yet to be rigorously evaluated through cross-site experiments. The need to develop and improve methods for such research is urgent, and much work is needed in this direction.

C. Speech Recognition Quality

In our previous work, we reported that users expressed a preference for speech interfaces with a proviso stating "only if it works well". In most cases when a user says this, she really means that the system's speech recognition accuracy needs to be high, as successive non-recognition prompts can be extremely frustrating. While speech recognition has been a consistent issue in our previous work, based on the improvements described in this paper, the system's recognition accuracy was almost at the level of commercial systems deployed in the West that use robust recognition models trained on the language they are used for. It is a significant finding that not a single participant reported the system's inability to recognize them correctly as a problem – while this was a common complaint in our previous work. Thus, it is clear that robust speech recognition is a core requirement for the success of a speech system, and that great care needs to be taken to improve speech recognition accuracy and quality.

D. The Importance of Effective Tutorials

Through the pre-study pilot, we saw that the initial tutorial strategy we made was not at all effective. By improving the strategy, we saw large improvements in users' ability to access information successfully. With an ineffective tutorial strategy, both interfaces may have been harder to comprehend for all participants, and this might have shifted their reported preference towards touch-tone, based on our earlier hypothesis.

In this paper, we have proposed human-guided instruction, and have shown that it worked successfully with low literate users. Compared with our prior work using video tutorials, the interactivity and individually-tailored nature of the human-guided tutorial make it a better fit for low literate users (as well as literate ones). Further work is needed to rigorously prove it as a formal method for speech interface usability research.

E. Rapid Iterative Development

In our most recent study, we used our mobile usability lab setup, which enabled rapid development and modification of the speech system while in the field. This meant that the feedback of local facilitators was used to make both major and minor modifications to the dialog flow of the system. Additionally, it meant that speech recognition tuning could be done locally and quickly. Finally, it was also possible to make minor changes after the pilot, as there were some issues that

became obvious only when new users started to use the system. All of this underscores the need for having a system development setup that enables field-based modification of the system. We aim to use this method in all our future work.

Related Work

There have been a number of approaches to GUI design for low-literate users. [3] presents design recommendations for non-literate users of a proposed PDA-like device, with many recommendations involving speech. However, these recommendations are not derived from empirical evidence from evaluations with actual semi- or non-literate users – they are derived from a literature review of research on Western users. [4] focuses on extending access to digital libraries by non-literate users, and also gives a short list of recommendations for such interfaces. However, usability tests reveal that users were not able to navigate information effectively, and result in recommendation for keyword search, audio-based help, and limiting the information set to lessen the cognitive load on users during navigation. [5] describes interface design guidelines, and a text-free interface that performed well in a usability test.

Speech interface research has resulted in a number of systems in various domains. While the most well known speech application is probably desktop dictation, this is just one point on a large multi-dimensional space of potential applications that can be made using speech. These dimensions include: choice of device (e.g., desktop, telephony, smartphone), task (e.g., information access, information entry), length of user training (often zero for commercial applications), vertical domain (e.g., stock prices, news, weather), acceptable user input (constrained, open-ended), interaction style (system initiative, user initiative, mixed initiative) and many others. For instance, Carnegie Mellon University's Communicator travel information system [6] and MIT's Jupiter weather information system [7] are two often-cited examples of speech-based information access systems usable over the telephone – these are mixed initiative systems that require zero user training, and accept a large range of user inputs, although as in all speech interfaces, acceptable user input is limited at each step. Most commercial systems tend to be more constrained, since these are cheaper to build, although exceptions do exist, such as Amtrak's "Julie" system which is much more flexible. Contrasted to the above are call routing applications, which are used to direct a caller to a specific operator, given a few utterances. The major push for speech interfaces in the developed world has come from the call center market, and that is what most research has focused on. However, since the needs of the populations that such systems serve are very different, there are entire domains that are unexplored (e.g., access to books through speech). Thus, there is a need for research in domains relevant to emerging regions, targeted towards the specific needs and abilities of users in these regions.

The Berkeley's TIER group's Tamil Market project was the first to design, develop and test a spoken language system with low-literate users in a domain (crop information access) relevant to them [8]. Results from a usability study of the speech interface suggest a difference in task success rates as well as in task completion times between groups of literate and non-literate users. Further, [15] gives a strong indication that there are differences in skills and abilities between these two user groups, and describes the linguistic differences to some detail, and suggests that further research is required to understand the nature of this difference, and to derive principles of dialog design targeted towards such users.

[9] describes a PDA-based interface designed for rural community health workers in India. While this may appear to have similarities to our work, their focus is on information entry, while ours is on information access. Furthermore, their interface is entirely GUI-based – ours is entirely speech-based.

[10] describes a system for data entry as well as access to decision support by community health workers in India. This is in the same domain as our project, and has many similarities to our work. However, our focus is on speech interfaces in this domain, while their approach is GUI-based.

[11] describes the iterative & collaborative design process for and evaluation of a GUI targeted to low-literate users for managing community-based financial institutions in rural India. While the principles of GUI design do not carry across well to speech interface design, the collaborative design process described has lessons highly relevant to all interface design in such contexts.

[12] describes VoicePedia, a purely telephone-based speech interface for searching, navigating and accessing the entire Wikipedia web-site. An evaluation comparing VoicePedia with a GUI-based smartphone equivalent shows comparable task success across interface conditions, although the (highly literate) users in the evaluation invariably preferred the GUI alternative.

[13] gives an excellent review of the potential contributions of CHWs in the developing world.

[14] provides a qualitative look at the informational needs and challenges of the homeless, in an urban site in the US.

[16] describes the difficulties low literate respondents face when asked questions that require abstract thought.

Acknowledgment

We would like to thank our sponsors and our collaborators [anonymized].

References

[1] N. Pakenham-Walsh, C. Priestley, and R. Smith. *Meeting the Information Needs of Health Workers in Developing Countries*. British Medical Journal, 314:90, January 1997.

[2] A. Chavan. 2007. *Around the World with 14 Methods*. http:// humanfactors.com/downloads/whitepapers.asp#CIwhitepaper. Accessed on August 22, 2008.

[3] M. Huenerfauth. 2002. *Developing Design Recommendations for Computer Interfaces Accessible to Illiterate Users*. Thesis. Master of Science (MSc). Department of Computer Science. National University of Ireland: University College Dublin.

[4] S. Deo, D. Nichols, S. Cunningham, I. Witten, 2004. *Digital Library Access For Illiterate Users*. Proc. 2004 International Research Conference on Innovations in Information Technology

[5] I. Medhi, A. Sagar, K. Toyama. *Text-Free User Interfaces for Illiterate and Semi-Literate Users*. Proc. International Conference on Information and Communications Technologies and Development, 2006.

[6] A. Rudnicky, E. Thayer, P. Constantinides, C. Tchou, R. Stern, K. Lenzo, W. Xu, A. Oh. *Creating natural dialogs in the Carnegie Mellon Communicator System*, in Proceedings of Eurospeech, 1999

[7] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T.J. Hazen, L. Hetherington, 2000 – JUPITER: *A Telephone-Based Conversational Interface for Weather Information*, in IEEE Transactions on Speech and Audio Processing, vol. 8, no. 1, January 2000.

[8] M. Plauche, U. Nallasamy, J. Pal, C. Wooters, and D. Ramachandran. Speech Recognition for Illiterate Access to Information and Technology. Proc. International Conference on Information and Communications Technologies and Development, 2006.

[9] S. Grisedale, M. Graves, A. Grunsteidl, 1997. Designing a Graphical User Interface for Healthcare Workers in Rural India, ACM CHI 1997

[10] V. Anantaraman, et al. *Handheld computers for rural healthcare, experiences in a large scale implementation*. In Proceedings of Development By Design, 2002.

[11] T. Parikh, G. Kaushik, and A. Chavan, *Design studies for a financial management system for micro-credit groups in rural India*. Proc. of the ACM Conference on Universal Usability, ACM Press (2003).

[12] J Sherwani, Dong Yu, Tim, Paek, Mary Czerwinski, Yun-Cheng Ju, Alex Acero, *VoicePedia: Towards Speech-based Access to Unstructured Information*, Interspeech 2007, Antwerp, Belgium.

[13] A. Haines, D. Sanders, U. Lehmann, AK Rowe, JE Lawn, S. Jan, DG Walker and Z Bhutta. *Achieving child survival goals: potential contribution of community health workers*. The Lancet 369(9579): 2121-2131. 2007

[14] C.A.L. Dantec, W.K. Edwards. *Designs on Dignity: Perceptions of Technology among Homeless*. In *CHI 08*.

[15] E. Brewer, M. Demmer, M. Ho, R.J. Honicky, J. Pal, M. Plauché, and S. Surana. *The Challenges of Technology Research for Developing Regions*. IEEE Pervasive Computing. Volume 5, Number 2, pp. 15-23, April-June 2006.

[16] A.R. Luria. *Cognitive Development: Its Cultural and Social Foundations*. Harvard University Press, Cambridge, MA. 1976.