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

Jahanzeb Sherwani[1], Sooraj Palijo[2], Sarwat Mirza[2], Tanveer Ahmed[2], Nosheen Ali[3], & Roni Rosenfeld[1]
[1]Carnegie Mellon University
[2]Health and Nutrition Development Society
[3]Cornell University

Abstract—**Information access by low literate users is a difficult task. Critical information, such as in the field of healthcare, can often mean the difference between life and death. We have developed spoken language interface prototypes aimed at low literate users, and tested them with community health workers in Pakistan. We present results showing that 1) in contrast to previous reports in the literature, well-designed speech interfaces significantly outperform touch-tone equivalents for both low-literate and literate users, and that 2) literacy significantly impacts task success for both modalities.**

Index Terms—**Speech technology, spoken language interfaces, low literate, illiteracy, 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. Telephony-based spoken language interfaces offer a unique solution to information access problems for such users, where cell phones are quickly become ubiquitous, and low literacy renders text-based modalities nonviable. In our research on spoken language interfaces for low-literate users in the developing world, we have chosen to focus on a specific domain where information access is especially important: 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 rates in the world, and, not surprisingly, also have the world's largest unmet need for health service providers. 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 [1]. These CHWs vary greatly in literacy levels and receive little refresher training [2]. 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" [3].

Over the past three years, we have been researching spoken language interfaces for information access by low literate community health workers in Pakistan. In the first phase of our work we have established the importance of telephony-based information access systems, and pilot-tested one interface with one group of community health workers [4].

Since then, we have conducted a number of 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 multiple lessons and findings from these studies:

- While reinforcement of existing CHW skills is often cited by public health officials as their most important educational goal, we show that usability testing should *not* focus on skills reinforcement (Section III).
- We describe novel improvements to the Poor Man's Speech Recognizer, for rapidly achieving accurate speech recognition in a new language using a standard US English speech recognizer (Section V).
- We describe a novel mobile setup (hardware and software) for carrying out field research in spoken interfaces, and argue that having such a modifiable system available at the field site is essential for rapid iterative development with participatory design (Section VI).
- We present lessons learnt from three pilot experiments in various urban and rural field sites (Section VII).
- We describe a novel method for quickly and effectively teaching novice, technology-shy participants how to use spoken interfaces, and show that an effective tutorial is crucial when conducting user studies on such interfaces. (Section VIII).
- We present both qualitative and quantitative results from a comparative user study showing that a well designed speech interface significantly

Manuscript received August 22, 2008.

outperforms a touch-tone equivalent for both low-literate and literate users; and that literacy significantly impacts task success for both interfaces (Section IX).

- Finally, we discuss the implications of these results and situate them in the larger context of research in ICT4D and SLT4D (Spoken Language Technologies for Development) (Section X).

II. Related Work

There have been a number of approaches to GUI design for low-literate users. [5] presents design recommendations for non-literate users of a proposed PDA-like device, with many recommendations involving speech. However, these recommendations are not based on empirical evidence from evaluations with actual semi- or non-literate users – they are derived from a literature review of research on Western users. [6] 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 recommendations for keyword search, audio-based help, and limiting the information set to lessen the cognitive load on users during navigation. [7] describes interface design guidelines, and a text-free interface that performed well in a usability test. [8] describes a PDA-based GUI 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. [9] 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. [10] 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.

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 [11] and MIT's Jupiter weather information system [12] 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 easier to build, although exceptions do exist, such as Amtrak's "Julie" system which is unusually flexible. Contrasted to the above are call routing applications, which are used to direct a caller to a specific operator, given a few utterances [13]. 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 still 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 [14, 15, 16].

The Tamil Market project undertaken by Berkeley's TIER group 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 [17]. Results from a usability study of their speech interface suggest a difference in task success rates as well as in task completion times between groups of literate and non-literate users, though differences were not statistically significant. Further, [18] gives a strong indication that there are differences in skills and abilities between these two user groups, describes the linguistic differences in 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.

More recently, researchers at the Meraka Institute [19] have been working on speech and touch-tone interfaces for health information services in South Africa. Preliminary results suggest that touch-tone interfaces may be preferable to speech interfaces. A study by IBM Research India comparing speech and touch-tone interfaces reached a similar conclusion [20]. Taken together, these studies appear to suggest that speech interfaces may not be very useful for low-literate users in the developing world. Based on the results reported in this paper, we strongly disagree.

[21] 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.

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

Finally, [23] describes the difficulties low literate respondents face when asked questions that require abstract thought.

III. 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 [4]. 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 our focus to low literate, rural community health workers. In collaboration with our partner NGO in Pakistan, the Health and Nutrition Development Organization (HANDS), we initially opted to work with reinforcing the material that the health workers were trained on (maternal and reproductive health), which is what a deployed system would eventually need to provide. Additionally, this seemed the prudent choice, as it is preferable to reinforce existing systems and practices 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 rarely 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 considerations 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.

IV. 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.

V. Improved "Poor Man's Speech Recognizer"

For speech recognition, we previously described a "poor man's speech recognizer" [4], 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 in our case). Thus a word such as 'wadheek maaloomaat' (transliterated Sindhi for "more information") 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 our approach significantly by incorporating a novel data-driven method, which we call the "Poor Man's Speech Recognizer++". The basic idea is to enable the developer and/or language expert to quickly generate new pronunciation definitions for words by varying any subset of phonemes in a given word's pronunciation definition, and then testing these variants with limited amounts of data to empirically find the pronunciation definitions that would lead to the highest recognition accuracy. For instance, if the developer is unsure of the optimal choice for the last consonant in the word "maaloomaat", she could specify a wildcard definition of "M AA L U M AA C?", where the "C?" denotes an "any consonant" wildcard. Similarly, if the developer wants to test the optimal phoneme choice for the final consonant-and-vowel combination in the word "bachaao", she may specify "B AX C? V?", where "V?" denotes an "any vowel" wildcard. These pronunciation entries are automatically expanded to a speech recognition grammar consisting of all possible pronunciations based on the wildcards. Thus, if there are a total of 20 consonants in the phonetic dictionary for the source language (in this example, US English), "M AA L U M AA C?" would be transformed into a list of 20 pronunciations, each with a unique final consonant. This speech recognition grammar is then used to run a re-recognition pass over any sample utterance(s) of the given word, and the best matched pronunciations are then manually chosen by the user to be used as the optimized pronunciations in the final system.

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?" (the forward slash denotes an arbitrary word boundary) would result in a 20*20 + 20*20 = 800 word grammar, which is much quicker to compute. While the final result may lose some accuracy with the introduction of an arbitrary word boundary, it is a useful heuristic that works significantly faster (less than a few seconds for a recognition result with MSS). Using this heuristic, a narrower set of pronunciations can be derived, which can then be tested without the arbitrary word boundary. Preliminary results using this improved approach are described in Section VIII.

VI. 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.

VII. Pilot Studies

A. Description

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

Month	Place	Avg. Education	Language	Sample Size
Jan	Memon Goth (town)	5-10 years	Urdu	10
Mar	Umarkot (rural)	<5 years	Urdu	10
Jun	Dadu (rural)	<5 years	Sindhi	10

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 an Urdu or 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 against which further work could be measured.

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 national language of Pakistan) 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 received at least some schooling had any knowledge of Sindhi (7 of the 10 participants). The remaining 3 participants did not understand Sindhi at all.

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 regarding triangulation, 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 complications 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.

VIII. 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 HANDS, and came from Umarkot and a nearby town, Samarro. A day before the actual study began we conducted a pre-study pilot with 3 participants

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 be 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 flavor of the 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 [23]. 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 practice tasks instead of two. The "free-style" 5 minutes were removed. Further, each of the tasks was carried out by the participant, while the facilitator listened in on each dialog, and provided successively less assistance. Specifically, the facilitator gave explicit instructions on every step for the first task, less help on the second task, and almost no help (unless the participant was 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 [24]. 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 were 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, and after two weeks, followed these with 11 more participants over 3 more days. The order of presentation of the two flavors 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.

IX. Results

Of the 20 participants, two were not able to speak Sindhi at all, and were unable to complete any of the tasks successfully – their data were removed from the final analysis.

A. *Personal Information*

Language: Of the remaining 18 participants, it is difficult to classify what language they spoke natively: not only is the local language (Thari) very similar to Sindhi, but there is also significant inconsistency in language and dialog naming. Many participants said they were native speakers of Sindhi, yet their Sindhi was very different from the Sindhi dialect used in the system. The fluidity of local dialects means that it is very difficult to tell with a high degree of certainty what dialect a particular person speaks by simply asking them.

Age: The average age was 23 years (SD = 5.3), with a maximum of 32 and a minimum of 17.

Years in School & Reading Ability: The average number of years in school was 6.3 (SD = 3.3), with a minimum of 0 and a maximum of 12. 3 participants were completely unable to read Sindhi, 5 were able to read with great difficulty, 7 were able to read with some difficulty, and 3 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 comprise the 'literate' participants. Thus, there were 8 low literate participants, and 10 literate ones.

Telephone use: 15 participants had used telephones prior to the study, with 10 participants reporting using a phone at least once every two days.

B. *Quantitative and Qualitative Results*

Task success in the speech interface was significantly higher than in the touch-tone interface. There was a significant main effect for the interface type, F(1,68) = 6.79, $p < 0.05$, with 31 of 36 tasks (86%) successfully completed in the speech condition, and 22 of 36 (61%) in the touch-tone condition. These results are shown in Figure 1.

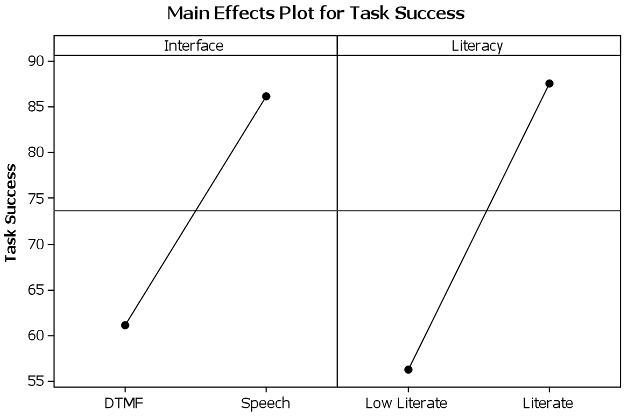

Fig. 1: Main Effects Plot for Task Success. There were main effects for both interface and literacy on task success.

Task success for literate participants was significantly higher than for low-literate participants. There was a significant main effect for literacy, F(1,68) = 10.61, $p < 0.01$, with 18 of 32 tasks (56%) successfully completed by low literate participants, and 35 of 40 tasks (86%) successfully completed by literate participants. These results are also shown in Figure 1.

Literate participants had a perfect task success rate when using the speech interface. There were no interaction effects of literacy and interface. There was a difference of 25% in task success for both literacy groups between the touch-tone interface and the speech interface. Similarly, there was a difference of 32% in task success between low literate and literate participants. It is striking to note, however, that literate participants using the speech interface had a 100% task success rate (20 of 20 tasks), as shown in Figure 2.

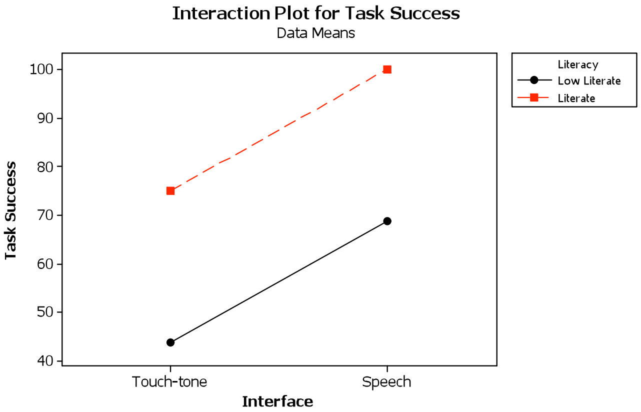

Fig. 2: Interaction Plot for Task Success. Literate participants using the speech interface had 100% task success.

There was no strong consensus on which interface was subjectively preferred. 10 users preferred the speech interface, while 8 preferred the touch-tone system. A sentiment echoed by a number of participants was "I don't use the phone that often, and I am not used to using numbers – I prefer speaking instead". However, other participants said "I am afraid I will say the wrong thing", and that "it is hard to speak to it, because I say too much". These participants understood what they were expected to say, but had a hard time saying it.

Some participants said that speech might be problematic if they're in a crowded area, since the system might end up hearing the sounds around them and get confused.

The 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 said that she would prefer the speech interface if both interfaces had been in her language. This sentiment was shared by other low literate participants for whom the system's language was difficult to understand.

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 portion of the data that was transcribed. Specifically, this portion consisted of 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 can respond with "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.

X. Discussion

A. *Task Success vs. Preference for Interface and Literacy*

While earlier studies have suggested an effect of literacy on interface use [17, 18], this study clearly demonstrates that literacy is a statistically significant determinant of task success. Moreover, the speech interface enjoyed a significantly higher task success rate than the touch-tone interface both for low-literate participants and for literate participants. Literate participants were able to solve every single task successfully using the speech interface, suggesting that lack of literacy constitutes a serious barrier to performance of these tasks, irrespective of the interface used.

One of the surprising, and seemingly contradictory findings in the above results is that literate participants reported that they had to remember less with the speech interface and preferred it, yet low literate participants said that the speech interface was harder, and preferred the touch-tone one, *even though they performed better with the speech interface on average*. This is a known effect with evaluations of speech interfaces [16] although with continued use, it is expected that user preferences conform to match task success [25].

B. *Orality and Literacy*

One of the frequently 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, learning how to answer tests and exams (similar to a user study), and even learning to make sense of other languages, dialects and accents. We have found Ong's framework of Orality and Literacy [26] to be a useful lens through which to analyze these issues.

Ong proposes orality as a way of describing how people think, communicate, and learn in a culture where writing has not become internalized. Orality theory argues that writing has so fundamentally transformed consciousness in literate cultures that we (literate researchers) are unable to grasp how oral cultures and people operate. [27] summarizes Ong's work on orality and discusses its implications for both interface design and user study design in developing regions. One of the key recommendations is the need to fundamentally redesign any content that was originally created for consumption by literate readers, since oral consumers require content with very different organization, presentation and context. Thus, orality provides a rich framework to understand why literacy makes a significant impact on task success, suggests ways to improve performance by low literate users, and also highlights the importance of localization.

C. Localization

Localization refers to adaptation of content to local culture, dialect, terminology and language usage patterns. Although crucial, localization is quite tricky. We have found that even communicating about languages and dialects is non-trivial: a rural participant may self-identify as a “Sindhi” speaker, yet may be unable to understand a “Sindhi” pamphlet recorded by a “Sindhi” speaker from the city. The pamphlet may contain words from other languages (e.g. Urdu), the accent of the city speaker may be unfamiliar, and the dialects of the languages may be substantially different.

Low literate participants are less likely to be exposed to alternative dialects, or to other languages, and find the urban-Sindhi-accented system’s output more challenging than the literate participants. Even one unintelligible word can throw off a low-literate listener completely [27]. When participants found the system’s Sindhi difficult to understand, they were hesitant to speak at all after many prompts with the speech interface, though when given the touch-tone interface, they did attempt to press buttons – this may be because speech interfaces require the user to expose their confusion more publicly by verbalizing something potentially incorrect, versus pressing a button, which is less open to scrutiny (and social ridicule) than speech.

The lesson, then, is that when designing a system for low literate users, it is crucial to choose both the language content and the system 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 or dialect 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. This is substantially different than in the developed world, where one can often expect uniformity in language in a given region, given the conforming effect of schools and of universal access to mass media.

Finally, the choice between speech and touch-tone may be a false dichotomy, as it may be optimal to provide both options, and let the user choose which option to use based on their current situation (e.g., when in a noisy environment, users may prefer to use touch-tone, but may switch to speech in a quiet place). This is common practice in the developed world.

D. Literacy and User Study Design

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 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 [24]), these methods have yet to be rigorously evaluated through multi-site experiments. The need to develop and improve methods for such research is urgent, and much work is needed in this direction.

E. Significant Design Decisions

The system described in this study is the outcome of various design decisions made over the course of more than year of testing various interface prototypes. Most notably, we attempted to optimize each flavor of the interface (touch-tone and speech) as much as possible. Thus, while touch-tone interfaces require a mapping between the specific choice (e.g. diarrhea) and the button to choose that option (e.g. “2”), speech interfaces do not have this requirement. This can be seen when comparing the equivalent prompts from both flavors:

Speech: *“What would you like to hear about? Malaria, Diarrhea, or Hepatitis?”*

Touch-tone: *“For information on Malaria, press 2, for information on Diarrhea, press 3, and for information on Hepatitis, press 4.”*

These optimizations meant that both interfaces were optimized in their own right – it would not make sense to “cripple” the speech interface by forcing it to match the less natural phrasing of the touch-tone interface, since this is one of the very advantages of speech interfaces that we wished to test.

Similarly, the system was chosen to have a persona (e.g., “Dr Marvi”), since it is much more natural for a low literate person to interact with a (virtual) person, than with an abstract system (e.g. “Health-Line”).

Also, the information presented in both flavors was carefully adapted to make the content more conducive for hearing, e.g. simplifying sentence structures, and replacing difficult words with easier phrases. Additionally, the content was designed to give a high-level summary in the beginning, and to successively give both greater detail and larger amounts of content, as the user progressed through the hierarchy. This required a substantial redesign of the content by a content expert (a medical doctor who supervises health worker training), as the existing healthcare material we started with did not follow this paradigm.

Finally, we chose a modified form of implicit confirmation strategy for dialog error-handling instead of explicit confirmation. In our approach, the system repeats what was recognized, and asks the user to take action only if recognition was incorrect (e.g. “Diarrhea. If this isn’t the topic you want, say ‘other topic’”). Explicit confirmations, on the other hand, force the user to state whether the recognition was correct or not (e.g. “Did you say Diarrhea? Please say ‘yes’ or ‘no’.”). Explicit confirmation is preferred when recognition accuracy is low, but is too tedious and distracting when accuracy is high, as is the case here. [28].

While not rigorously quantified, we believe that each of these design decisions was significant in making the user interface more usable by our end-users.

F. Speech Recognition Quality

While speech recognition accuracy has been a persistent problem in our previous work, based on the improvements described in this paper, the system's recognition accuracy (91%) was comparable to commercial systems deployed in the West that use robust recognition models trained on the language they are used for. We believe that robust speech recognition is a necessary (though far from sufficient) condition for the success of a speech system, and great care needs to be taken to improve speech recognition accuracy when conducting such research.

G. 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 in which users learn to use the system with a human mentor, and have shown that it worked successfully. Compared with our prior work using video tutorials, the interactivity and individually-tailored nature of the cooperative human-guided tutorial make it a better fit for both low literate and literate users. Further work is needed to rigorously prove it as a formal method for speech interface usability research.

H. Rapid Iterative Development

In our most recent study, we used our mobile user study infrastructure, 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.

I. Comparison with Similar Research

Our results contradict similar work in the field, most notably the study by Patel et al. [20] testing speech and touch-tone interfaces for listening to pre-recorded radio shows and recording audio content for a radio talk show. In comparing our work with theirs, a number of factors need to be considered.

First, in our system, the speech-input flavor was more conversational (e.g. "What would you like to hear more information about, diarrhea, malaria, or hepatitis?") as compared to theirs (e.g., "To ask a question, say 'question'; to listen to announcements, say 'announcements'; to listen to the radio program, say 'radio'"). It is this mapping of keyword to semantics that touch-tone interfaces are forced to use (e.g. "For information about diarrhea, press 1"), though spoken interaction can avoid this requirement, making the interface more natural. We believe this difference in the interface is very significant for low literate and other technologically inexperienced users.

Next, in their study with 45 participants, the only task that showed a significant benefit of touch-tone over speech was the one that required users to record their voice as the goal of the interaction. Speech interfaces that combine restrictive keyword-based grammars with open-ended "say anything" recording segments are very difficult for users [29], since it is not obvious when (or even why) it is not possible to speak in sentences in one part of the interaction, but it is required to do so in another part.

Finally, based on our goals (a system for community health workers that can be trained), we were able to spend a considerable amount of time training participants in the user study on both the touch-tone and speech interfaces. Their system was designed and tested for users without any training, which is why their user study did not involve any training beyond a brief introduction. This difference is noteworthy, as even a limited amount of training can make a significant difference to the usability of an interface, as we saw during our pilot study.

Thus, when comparing one study with another, it is important to keep the specifics of the design of the interface, study and tasks in mind, as well as of the larger goals of the system involved. Their study is an important and significant contribution insofar as it warns against the design of speech interfaces for tasks involving recording a spoken message in the context of untrained users. However, this should not be extrapolated to mean that touch-tone interfaces are inferior to speech interfaces in the developing world in general. Our study shows that speech interfaces can be significantly better than touch-tone interfaces for a different design of the interface, the task, and the user study.

Finally, the study on the OpenPhone interface for HIV caregivers [19] suggests that users express preference for touch-tone interfaces when privacy is an issue. Privacy was never expressed as an important factor by participants in our study, and it is clear that such issues largely depend on the cultural context involved, as well as the specifics of the system's domain (e.g., HIV vs. neonatal health).

Thus, more work is needed to identify exactly where speech interfaces work well and where they do not.

XI. Conclusion

We draw two main conclusions from this work. The first is that *the ability to perform information access tasks is strongly hampered by low literacy, and not just because of the inability to read*. We derived empirical confirmation that literacy is a significant determinant of success in information access tasks. This by itself is not surprising, but our results further suggest that the problems associated with low literacy go far beyond the inability to read, since they also affect task performance using the speech interface, where no reading is necessary.

Our second conclusion is that, at least for some target populations, *a well-designed speech interface can significantly outperform a touch-tone interface for both low-literate and higher literate users*. Given the potential utility of information access and the pervasiveness of low literacy throughout many parts of the world, we hope to see many spoken dialog systems developed, evaluated and deployed in the near future.

Acknowledgment

This project was supported by Microsoft Research through its Digital Inclusion Initiative and by the U.S. Agency for International Development (USAID) under the Pakistan-U.S. Science and Technology Cooperation Program. We are grateful to both organizations for making this work possible.

We would also like to thank all the members of the HANDS staff that facilitated this research, especially Dr Abdul Aziz, Bansi Malhi, Laal Jaan, and Dr Anjum Fatma. We are very grateful to Safieh Shah, Samia, Arbab, and Lata for facilitating the various user studies. Finally, we would like to thank the community health workers at Wahi Pandi, Memon Goth, and Umarkot who gave their time and feedback.

References

[1] H. M. Kahssay, M. E. Taylor, P. A. Berman. *Community Health Workers: The Way Forward*. World Health Organization, 1998.

[2] S. Hunt. *Evaluation of the Prime Minister's Lady Health Worker Program*. Oxford Policy Management Institute. http://www.opml.co.uk/go.rm?id=380. Accessed Feb 1st, 2009.

[3] 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.

[4] J. Sherwani, N. Ali, S. Mirza, A. Fatma, Y. Memon, M. Karim, R. Tongia, R. Rosenfeld. *HealthLine: Speech-based Access to Health Information by Low-literate Users*. In Proc. IEEE/ACM Int'l Conference on Information and Communication Technologies and Development, Bangalore, India, December 2007.

[5] 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.

[6] 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

[7] 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.

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

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

[10] 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).

[11] 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

[12] 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.

[13] A. L. Gorin, B. A. Parker, R. M. Sachs, J. G. Wilson. *How May I Help You*. Speech Communications, Vol. 23, pp. 113-127, 1997.

[14] J. Sherwani, R. Rosenfeld. *The Case for Speech and Language Technologies for Developing Regions*. In Proc. Human-Computer Interaction for Community and International Development workshop, ACM CHI, Florence, Italy, April 2008.

[15] F. Weber, K. Bali, R. Rosenfeld, K. Toyama. *Unexplored Directions in Spoken Language Technology for Development*. In Proc. Spoken Language Technology for Development workshop, SLT, Goa, India, 2008.

[16] E. Barnard, M. Plauche, M. Davel. *The Utility of Spoken Dialog Systems*. Proc. Spoken Language Technology for Development workshop, SLT, Goa, India, 2008.

[17] 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.

[18] 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.

[19] C Kuun, OpenPhone project piloted in Botswana http://www.csir.co.za/enews/2008_july/ic_05.html, accessed Feb 1st, 2009.

[20] N. Patel, S. Agarwal, N. Rajput, A. Nanavati, P. Dave, T. Parikh, *A Comparative Study of Speech and Dialed Input Voice Interfaces in Rural India*. ACM CHI 2009.

[21] 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.

[22] 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

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

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

[25] A. I. Rudnicky. *Mode Preference in a Simple Data-Retrieval Task*. Proceedings of the ARPA Workshop on Human Language Technology. San Mateo: Morgan Kaufmann, 1993, 364-369.

[26] W. Ong. *Orality and Literacy: The Technologizing of the Word*. London: Routledge, 2002.

[27] J. Sherwani, N. Ali, R. Rosenfeld. *Orality-grounded HCID: Understanding the Oral User*. Submitted to the Information Technology and International Development journal.

[28] D. Jurafsky, J. H. Martin. *Speech and Language Processing*. Prentice Hall, 2008.

[29] J. Sherwani, S. Tomko, R. Rosenfeld. *Sublime: A Speech- and Language-based Information Management Environment*. In Proc. IEEE Int.l Conference on Acoustics, Speech and Signal Processing, Toulouse, France, May 2006.