Speech Interfaces for Information Access by Low Literate Users

Jahanzeb Sherwani

CMU-CS-09-131

May 2009

School of Computer Science
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee
Roni Rosenfeld, Co-chair
Alex Rudnicky, Co-chair
Alan Black
Raj Reddy
Alex Acero, Microsoft Research

Submitted in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy

This research was sponsored by Microsoft Research, the Siebel Scholars Foundation, the National Science Foundation under grant number EIA-0225656 and SRI International under grant number 55-000691. Partial support for the project was provided by the U.S. Agency for International Development under the Pakistan-U.S. Science and Technology Cooperation Program. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: Speech recognition, speech interfaces, developing countries, emerging markets, ICTD, ICT4D.

For Adil, Omar, and Faraz,
Ammi and Abboo,
and Nosheen

Abstract

In the developing world, critical information, such as in the field of healthcare, can often mean the difference between life and death. While information and communications technologies enable multiple mechanisms for information access by literate users, there are limited options for information access by low literate users.

In this thesis, I investigate the use of spoken language interfaces by low literate users in the developing world, specifically health information access by community health workers in Pakistan. I present results from five user studies comparing a variety of information access interfaces for these users. I first present a comparison of audio and text comprehension by users of varying literacy levels and with diverse linguistic backgrounds. I also present a comparison of two telephony-based interfaces with different input modalities: touch-tone and speech. Based on these studies, I show that speech interfaces outperform equivalent touch-tone interfaces for both low literate and literate users, and that speech interfaces outperform text interfaces for low literate users.

A further contribution of the thesis is a novel approach for the rapid generation of speech recognition capability in resource-poor languages. Since most languages spoken in the developing world have limited speech resources, it is difficult to create speech recognizers for such languages. My approach leverages existing off-the-shelf technology to create robust, speaker-independent, small-vocabulary speech recognition capability with minimal training data requirements. I empirically show that this method is able to reach recognition accuracies of greater than 90% with very little effort and, even more importantly, little speech technology skill.

The thesis concludes with an exploration of orality as a lens with which to analyze and understand low literate users, as well as recommendations on the design and testing of user interfaces for such users, such as an appreciation for the role of dramatic narrative in content creation for information access systems.

Acknowledgments

When I started in 2003, as a young, optimistic computer science graduate from Pakistan looking to work on applying technology to the problems of international development, I really wondered whether I would be able to combine my interests in a way that would be relevant to the developing world, be intellectually stimulating, and also be considered worthy of a PhD in Computer Science, all at the same time. Even if I was to find such a project, how would I find a professor who would be interested in supervising such work, and also be actively engaged with it? Luckily for me, I did. Roni Rosenfeld has been the best advisor I could ever have imagined – a constant source of encouragement, mentorship, sage advice, and a great friend, and I feel extremely fortunate to have had the opportunity to work with him over the past 6 years.

Alex Rudnicky, my other advisor, and Alan Black have also always been a constant source of support, and I am grateful for the many exchanges we have had over the years. Both times after I returned from a year's worth of fieldwork in Pakistan, Alex and Alan's encouragement and ideas gave me renewed strength to continue the work, and constantly led me to consider issues that I hadn't thought of before. I'm also grateful to Raj Reddy, and Alex Acero, for being on my committee, and for providing many helpful comments and suggestions throughout the process. I'd also like to thank Rahul Tongia and Bernardine Dias for all the help and support, as well as for creating an ICTD space within CMU. And a special thanks to Roger Dannenberg for helping me learn how to teach, and for enabling the musician in me.

I'd also like to thank a number of friends and colleagues at CMU, many of whom were part of the Sphinx Group, the Dialogs on Dialogs student reading group, and the Young Researchers Roundtable on Spoken Dialog Systems. Dan Bohus, Antoine Raux, Stefanie Tomko, Thomas Harris, Satanjeev Banerjee, Udhay Kumar, and Matt Marge made my time at CMU both fun and intellectually stimulating, and helped shape my work considerably. I'm also grateful to Betty Cheng, Ulas Bardak, and Vasco Pedro for being great friends, and also startup partners. Also to my undergraduate friends and bandmates, for keeping me in touch with my music: Mustafa Khan, Amarpal Banger, Anisha Anantapadmanabhan, and Latika Kirtane. A special thanks to Baber for being an inspiration and a great friend, and an awesome bassist.

I'd also like to thank Randy Pausch and Caitlin Kelleher, for meeting me during my chance visit to CMU before I'd applied, and for giving me a few words of wisdom that enabled me to do so much.

I'd also like to thank a number of non-CMU researchers I've met in the ICTD space, who have helped me incredibly: Yaw Anokwa, Brian DeRenzi, Divya Ramachandran, Thomas Smyth, Rowena Luk, Neema Moraveji, Melissa Ho, Matt Kam, Bill Thies, Aditi Grover, and Ilda Ladeira. I'm amazed and humbled by the work they do, and am grateful for all the exchanges we have had. I'm also grateful to pioneers in the ICTD field, who have enabled a generation of researchers to do this work, especially Eric Brewer, Kentaro Toyama, and Tapan Parikh.

None of my work would have been possible without the help and support of countless people in Pakistan. First of all, Dr Mehtab Karim and Dr Yousuf Memon at Aga Khan University's Community Health Sciences Department, who helped get the project started, and provided significant support in the first year. Dr Anita Zaidi, Dr Omrana Pasha, Dr Babar Shaikh, Dr Gregory Pappas, who always gave me an ear when I needed it, along with lots of practical advice. The CHS field team, especially Khadim, Rubina, and Afzal, whose tireless efforts to improve health outcomes in economically challenged areas of Karachi has always been a source of inspiration for me.

More recently, working with the HANDS organization in Karachi has shaped me both intellectually and personally. I'd like to thank a number of people at HANDS. Dr Sarwat Mirza for taking the risk and agreeing to work with me since our first trans-Atlantic phone call, and for all the thoughtful contributions to the project since then. Dr Tanveer Ahmed, who has been instrumental in taking the partnership forward, and Dr Sooraj Palijo, who gave both her time and her voice to the system. Dr Anjum Fatma, who gave considerable time to guiding the initial phases of the research. Additionally, I'd like to thank Dr Abdul Aziz and Bansi Malhi, the HANDS managers in Johi and Umarkot – their support was instrumental in enabling the user studies detailed in this thesis. Moreover, I'd like to thank Arbab and Saima in Wahi Pandhi, and especially Lata, in Umarkot, who learnt the basics of HCI user study practice in a manner of hours, and helped facilitate user studies with an expertise and motivation that was incredible. Moreover, it has been an immense privilege and honor for me to have been able to work alongside the HANDS team, and to see a grass-roots organization where everyone from the top-level manager to the field worker does whatever it takes to get the job done – traveling for hours on often broken roads to distant field sites, and creating, nurturing, and building human relationships in these areas, as well as institutionalizing organizations for the betterment of the communities they work in. Working with HANDS and Aga Khan University, and seeing their incredible hard work, and its impact in the realm of public health, has been inspiring and humbling, and has at many times made me want to become the *other* kind of doctor.

Additionally, I cannot give enough thanks to the hundred community health workers and mothers that were interviewed for the needs assessment, and who participated in the user studies. With huge resource constraints, incredible family pressures, and an immense task of caring for their families and communities, they gave their time to participate in our research. These community health workers are the backbone of the health infrastructure in most of the developing world, and the work they do is truly inspirational.

I'd also like to thank my alma mater, the Lahore University of Management Sciences (LUMS), for teaching me two very different sciences – the technical and the social – without which this work would not have been possible. I'd like to thank the founders of the University who had the foresight to extend its vision beyond management to a broad-based undergraduate education: Babar Ali, Razzak Dawood, Nasim Saigol. In the faculty, I'd like to thank all the teachers that shaped my intellectual growth, especially Dr Arif Zaman, Dr Zahoor Hassan, Dr Zafar Rashid, Dr Ghazala Irfan, Osama Siddique, Amelie Blom, Hasan Karrar, Maliha Khan-Tirmizi, and Khurram Husain, for enlightening me on both aspects of my undergraduate education. I'd like to give special thanks to Syed Umair Ahmed Shah, for teaching the most social computer science course ever, and for recommending I apply to Carnegie Mellon. I'd also like to thank and acknowledge Yasser Hashmi, a great friend and mentor, for showing me the exciting world of psychology, and for giving me a helping hand in the world of research, and for encouraging me to chase my dreams – I know I would not have been here today without your help, support and encouragement.

Finally, I'd like to thank my family. My elder brother Adil, for always being there, and to Umaimah for all the help in Sindhi. My younger brothers Omar and Faraz, for hanging out and jamming whenever I needed a break, and for helping beta test the speech interfaces and transcribe and translate pages worth of Urdu text. My father Khalid and my mother Gaiti, for providing more encouragement and support than a son could possibly imagine, and for helping me in every possible way every single time I needed it, and also for helping create both the voice and soul of the early HealthLine prototypes. While it's a rare thing that a child joins a PhD program (especially in Pakistan), it's even rarer when his parents can engage with the work and the process the way you both did. I am eternally grateful to you for everything!

And of course, I'd like to thank Nosheen. For seeing me through everything. For being a part of my research since the beginning – and especially for conducting the needs assessment interviews and analysis, and a number of the user studies. But mostly, for always being there, for sharing in the roller-coaster that is (grad school) life, and for helping me in countless ways. This would not have been possible without you. Thank you so much!

Table of Contents

1. Introduction

In the developed world, speech interfaces for information access have successfully made the transition from the academic research lab to mainstream use. Today, such interfaces are routinely used for tasks such as flight reservation, cinema timetable lookup, directory assistance, and even driving directions.

Most research and development work in speech interfaces has tended to focus on users in the so-called "developed world" – primarily North America and Western Europe – where income and literacy levels are much higher than in the so-called "developing world" – mainly parts of Africa, Asia, and Latin America.

Lower literacy levels in the developing world, coupled with lower incomes, means that traditional models of computing – usually involving an expensive, energy-intensive screen with a high reliance on literacy – may be unfeasible and inappropriate. Spoken language interfaces using telephones, on the other hand, hold significant potential for such users: there is already large-scale deployment of telephones in developing countries, and speech interfaces overcome the cost and literacy barriers posed by regular computers. Unfortunately, there has been very little research done on speech interfaces for low-literate users in developing countries – and initial evidence suggests that there may be significant differences in the way interfaces need to be designed for such users.

In this thesis, I present original research on speech interfaces for low-literate users, with contributions spanning technical aspects of the technology, interface design, as well as a deeper understanding of the end-users.

The main research questions that formed the starting point of this thesis were as follows:

1) For low-literate users in the developing world, how do speech interfaces compare to existing alternatives (text material and touch-tone interfaces), both in terms of objective and subjective metrics?

2) Is literacy a significant factor in the ability of a person to be able to effectively use a telephone-based interface?

3) Is it possible to repurpose existing speech resources (created for users in the developed world) to build robust platforms for spoken interfaces in the developing world?

Additionally, a final goal of this research was to understand how existing methods for conducting speech interface research need to be expanded for the specifics of working with low-literate users in the developing world.

We begin with a review of background material in which our work is situated.

2. Background Material

2.1. Information & Communications Technologies and Development

Just under half the world's population, or around 2.8 billion people, currently live on less than 2 dollars a day. There exist a number of domains where Information and Communication Technologies (ICTs) can provide real value to such populations, in a way that is both sustainable and appropriate (Manseel & When, 1998; Tongia, Subrahmanian & Arunachalam, 2005). There are hundreds of such projects (as cataloged by the World Bank[1]), most of which use existing, off-the-shelf technology (Brewer et al., 2005). However, the direct transfer of "First World" technology has not been successful in most cases, primarily because of the mismatch between the intended environment the technology was designed for, and the ground realities of the environments in which they are deployed (WRI, 2004; Brand & Schwittay, 2006). Brewer et al. describe this situation concisely: "Although it is clear that there are large differences in assumptions related to cost, power, and usage, there has been little work on how technology needs in developing regions differ from those of industrialized nations. We argue that Western market forces will continue to meet the needs of developing regions accidentally at best". Research on technology design tailored to the specific needs of emerging regions is needed to address this issue (Brewer et al., 2005; Tongia & Subrahmanian, 2006; Brand & Schwittay, 2006].

Domains such as health, education, agriculture, e-governance, and commerce have great potential for the application of appropriate technology (Mansell & Wehn, 1998). Within these domains, there are multiple avenues where technology can play a role, such as information access (Geertz, 1978), information entry (Donner, 2004; Prahalad, 2005), (decentralized) information sharing, access to services, and access to automated decision-making systems. These applications in the aforementioned domains can lead to higher efficiency, lower costs, greater reliability, more accessibility, greater localization and better quality of information & content, decentralized communication, which in turn directly lead to better income, health, education, and ultimately, in the quality of life (Abraham, 2006).

Many ICT-in-development (ICTD) initiatives involve the use of standard PCs as the form factor, and standard web-based forms or Windows-based GUIs as the primary interface, and the Internet for connectivity. However, PCs and current GUIs were designed with a specific ("First World") user in mind: a user who can afford a roughly $500 machine as well as Internet connectivity, has access to a stable electricity supply, is literate, uses a language that has a written form, and finally can access and afford technical support when something fails or needs to be upgraded. These requirements are unrealistic for major parts of the developing world, where users in many cases cannot afford such costly technology, do not have access to continuous electricity, are not literate, may be fluent only in a language without a written form, and do not have access to any ongoing support for using unintuitive technology. There are successful ICT projects using PCs: for example, access to market price information (Kumar, 2004), a database for answering questions from rural

[1] http://www.infodev.org

communities (Ramamritham et al., 2006), e-government services through telecenters (Rajalekshmi, 2007). However, there is a large part of the developing world for which such design is not viable. It is no surprise that rates of PC use in the developing world are dwarfed by those in the developed world.

Cell phones, on the other hand, are a huge ICT success (Tongia & Subrahmanian, 2005). Cell phone use across the developing world is increasing remarkably – according to a report by the International Telecommunications Union titled "Mobile overtakes Fixed":

"The greatest impact of mobile communications on access to communication services – in other words, increasing the number of people who are in reach of a telephone connection of any kind – can be seen in developing countries." (ITU, 2003)

Across Sub-Saharan Africa, access to cell phones surpassed access to fixed-line phones in 2000 – the same occurred in South Asia in 2002 (Sinha, 2005). Given that cell phones in the developing world often follows a shared model of usage, the actual number of people with access to telephony is much higher than the number of subscriptions suggest (Vodafone, 2006). The extensive use of cell phones suggests that this is a fundamental mechanism through which underdeveloped regions are benefiting from ICTs, because they are easy to use, affordable, and suitable for non-literate populations. Furthermore, the sustained use of cell phones in these regions also implies the existence of widespread ecosystems of supply, maintenance and technical support, which do not exist for other types of devices. For these reasons, it is believed that cell phones have great potential for facilitating ICT projects of a wide variety (Tongia, Subrahmanian & Arunachalam, 2005).

As of March 2009, cell phones are currently the most widely used consumer electronics device in the world, with 61.1 cell phone subscriptions per 100 people, as shown in the figure below (ITU, 2009).

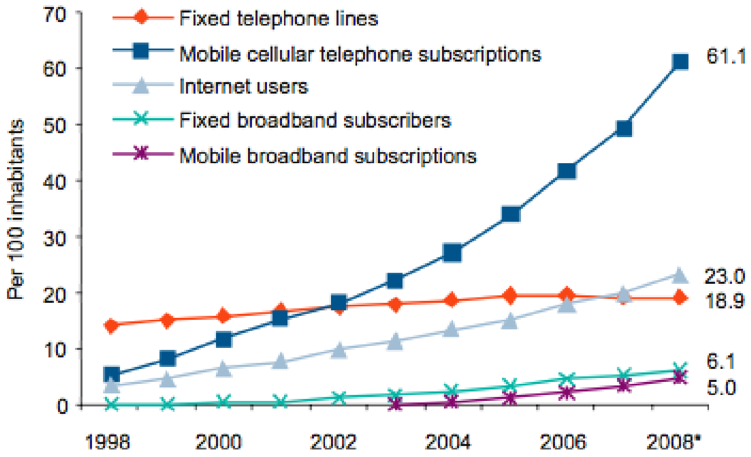


Figure 1: Global ICT Developments, 1998-2008. Source: ITU (2009).

A growing number of ICTD researchers and practitioners have begun to use cell phones as platforms for ICTD initiatives. Recently, there have been many such applications spanning many domains: language literacy (Kam et al., 2009), mobile money transfer (Medhi et al., 2009), remote medical consultation (Luk et al., 2009), post-conflict reconciliation (Best et al., 2009), data entry by rural micro-credit groups (Parikh, 2005), medical decision support (DeRenzi et al., 2008), sharing stories (Jones et al., 2008) and many others. The vast majority of these applications, however, are graphical interfaces that require specific handsets that support them. Unfortunately, the types of handsets that are fueling the growth in cellular penetration are mostly not smartphones with the ability to render complex interfaces and perform computation; rather they are devices that can carry out the basic function of a cell phone: the ability to make a telephone call. Additionally, GUIs largely depend on literacy, and with literacy rates of less than 50% in many developing regions, this is not a mechanism that can work for all (UNESCO, 2008).

The core technologies of speech recognition and speech synthesis, on the other hand, do not require literacy and in fact even work for languages that have no written form. Thus, interfaces that use speech as the underlying modality – otherwise known as spoken dialog systems – hold great promise as an interface choice for such users.

These users are not one homogenous group. Income, literacy, and other factors vary widely within regions, although in general, it is the case that the affluent and literate are the minority, while the poor and semi-literate are the majority. SLTs may not be the answer for those at either extreme of the income & literacy spectra. For the resource-rich, the realities are similar to those in the West, for whom speech has not been appealing, and for whom other technologies such as Internet-through-the-PC may be more affordable and accessible, and so are less motivated to use SLTs. For the extremely resource-starved the situation is completely the opposite: they may not be able to easily learn to use SLTs, and might have more pressing needs, such as food and water, instead of information access (although it has been argued that information can reduce the price paid for such commodities greatly (Kirkman, 2001; Geertz, 1978)). Additionally, some have argued that for any technological intervention in a rural community, it is advised to have “human access points” who introduce and mediate the technology for a given community (Marsden et al., 2008). Community health workers are ideal candidates to function as such human access points, and we now turn to a description of community health.

2.1.1. Community Health Programs in Developing Countries

Healthcare is a fundamental, yet under-serviced need of citizens in developing countries. In addition to having the lowest health indicators (e.g. maternal mortality, neonatal mortality, HIV and Tuberculosis prevalence), these regions 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.

These CHWs vary greatly in literacy levels and receive little refresher training, if any. 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” (Pakenham-Walsh et al., 1997).

The Pakistani government, for example, has initiated a community health worker program with the same logic – called the “Lady Health Worker Programme” (LHWP). This program employs 100,000 LHWs across Pakistan (a country with a population of around 160 million). These LHWs receive 3 months’ training, with no refresher courses in most cases. A recent evaluation of the LHWP gave a strong recommendation for the improvement in the quality of knowledge of the LHWs (OPML, 2002; Afsar & Younus, 2005). Many other countries have similar programs (WHO, 2006; Kahssay et al., 1998).

Traditional mechanisms for health information access by LHWs have not been adequate. The easiest such mechanism for health workers is to ask someone who is better-informed: a doctor, a nurse, or even the health worker’s supervisor. Unfortunately, there are not enough doctors and nurses to satisfy the information demands of the health workers. Furthermore, there are interpersonal dynamics that limit the effectiveness of supervisor-worker training: some supervisors have the same training as the health workers they supervise, and are afraid of losing their job to a well-performing health worker (Afsar, 2005).

Another mechanism used for information access is that of written text: handbooks and training manuals are difficult to distribute, costly to update, and are seldom referred to by health workers (see Section 3.2 for details). Solutions involving information access through PCs (using the Internet or even CD-ROMs) are not viable for reasons of cost, literacy, and access. Thus, there is a strong need for information that is not being addressed through existing mechanisms.

2.1.2. Range of Health Worker types

There are many different levels of health workers, varying from the two extremes of the highly educated MDs to the completely untrained Traditional Birth Assistants (TBAs). TBAs are women whose family’s women have traditionally been the midwives of their community, and their only source of knowledge is what has been passed down through the oral tradition. While much of their practices are safe and effective, their knowledge can actually be harmful to their clients.

For an MD, accessing a dialog system for health information may be needless and cumbersome – MDs are very well educated, can easily refer to other doctors and professional journals, and in many cases have access to the Internet as well. For a completely non-literate TBA, it may be prohibitively difficult to learn the use of the dialog system, even though it may provide real value. My hypothesis is that there exists a group of users in between both these extremes in education and income levels, who are semi-literate and are able to learn the use of a dialog system without much difficulty, but at the same time cannot access, afford or use other information mechanisms, and hence have an optimal mix of ability and motivation to use such a system.

Trained community health workers (such as LHWs) described previously fit this description quite well. They have some limited education, and have limited health training. They clearly have the motivation to access health information (and have expressed this desire unequivocally). Yet, they are unable to use existing mechanisms well (e.g., books) and are not provided the support they can use most easily (e.g., doctors). A well-designed dialog system, tailored to their needs, has the potential to give them the information they want, and their motivation levels and lack of reasonable alternatives should enable them to be more forgiving of the system's flaws.

Furthermore, books are costly to print and distribute, difficult to carry around, and impossible to implement usage monitoring with. Updates to such books are also very expensive. Speech systems, on the other hand, have a much lower cost of scale-up and maintenance. Only a cell phone needs to be carried around, and an update to the system is reflected in all interactions thereafter by all health workers. Also, it is easy to monitor usage of the system, and to use this information to improve and/or localize specific content that is most used. Thus, if speech systems are even comparable to handbooks (and not significantly better) in usability, there are large external advantages they bring as well.

Community Health Worker Roles

The range of activities of community health workers is varied and complex. Each country can have multiple CHW programs (both governmental and non-governmental), and the design of these programs can place very different requirements on CHWs' roles, which in turn decides what information they need. A sample of the activities of CHWs in various countries is given below.

Task Summary	Benin	Botswana	Colombia	India	Liberia	Phillippines	Sudan	Thailand	Yemen
1 First aid, treat accident and simple illness	✓	✓	✓	✓	✓	✓	✓	✓	✓
2 Dispense drugs	✓	✓	✓	✓	✓	✓	✓	✓	✓
3 Pre- and post-natal advice, motivation	✓	✓	✓	✓	✓	✓	✓	✓	✓
4 Deliver babies	✓		✓						
5 Child-care advice, motivation	✓	✓	✓	✓	✓	✓	✓	✓	✓
6 Nutrition motivation, demonstration	✓	✓	✓	✓	✓	✓	✓	✓	✓
7a Nutrition action, weigh children		✓	✓			✓		✓	
7b Distribute supplements	✓					✓	✓	✓	
8 Immunization motivation, clinic assistance	✓	✓	✓	✓	✓	✓	✓	✓	✓
9 Immunization-give injections			✓				✓		
10 Family planning motivation	✓	✓	✓	✓	✓	✓	✓	✓	✓
11 Family planning-distribute supplies		✓	✓	✓		✓	✓	✓	
12 Environmental sanitation, personal hygiene,	✓	✓	✓	✓	✓	✓	✓	✓	✓
13 Communicable disease screening, referral	✓	✓	✓	✓	✓	✓	✓	✓	✓
14 Communicable disease follow-up, motivation	✓	✓	✓	✓	✓	✓	✓	✓	✓
15 Communicable disease action		✓	✓	✓		✓	✓		✓
16 Assist health centre clinic activities	✓	✓	✓				✓		
17 Refer difficult cases to health centre	✓	✓	✓	✓	✓	✓	✓	✓	✓
18 Perform school health activities regularly		✓					✓		✓
19 Collect vital statistics		✓	✓	✓	✓	✓	✓	✓	✓
20 Maintain records, reports	✓	✓	✓	✓	✓	✓	✓	✓	✓
21 Visit homes on a regular basis	✓	✓	✓	✓	✓	✓	✓	✓	
22 Perform tasks outside health sector (e.g.. agriculture)	✓	✓		✓	✓	✓	✓	✓	✓
23 Participate in community meetings	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Roles of Community Health Workers in Various Countries (Kahssay et al., 1998)

Given the range of CHW activities, the information needs would be similarly diverse. However, it is clear that maternal and child health are key domains in which CHWs operate. It is these domains that our work focuses on.

2.2. Interfaces for Low-literate Users

There has been a modest amount of research on principles of design for semi-literate users in the developing world. Hofstede (1997) developed a framework of five fundamental cultural dimensions based on anthropological research on IBM employees from 53 countries:

- Power distance: the extent to which people accept power inequality and social hierarchies
- Individualism vs. collectivism: the orientation towards individual or group achievements
- Masculinity vs. femininity: the degree to which a culture separates gender roles
- Uncertainty avoidance: the degree to which uncertainty is considered uncomfortable
- Long-term time orientation: the degree of devotion to long term traditions

There have been suggested mechanisms for concretizing such higher-level theories by mapping each dimension to specific user interface components (Marcus, 2003;, Ackerman, 2002). For instance, Ackerman suggests that web interface design for cultures with high power distance (i.e., where power inequalities are more accepted) should use more formal text. These recommendations are based on the authors' experience in designing web-sites in different countries, using Hofstede's underlying theoretical framework.

Huenerfauth (2002) presents a large list of design recommendations for non-literate users of a proposed PDA-like device, with many recommendations for speech interfaces. However, these recommendations are not derived from empirical evidence from evaluations with actual semi- or non-literate users – they are more the result of a literature review of interface issues with Western users. Most of the given design recommendations assume that such users would respond very similarly to speech interfaces as Western users do (e.g., "[users] of speech-only interfaces like to feel that they are in control of the interface, and allowing users to interrupt the system's speech output is an important way to convey this sense of control"). However, according to Huenerfauth, users with high power distance *prefer* interfaces where the system maintains control.

A project by the creators of the Greenstone Digital Library platform focused on extending access to digital libraries by non-literate users (Deo et al., 2004). The authors enumerate a list of user requirements for the interface, including high learnability, habitability, usability, robustness to errors, and non-textual as far as possible. Testing paper prototypes provided the further guidelines of the need for an easily accessible main menu, as well as audio-based help. Usability studies of the actual system showed that users (non-literate entrants to an adult school in New Zealand) were not able to adequately navigate the information, mostly because of the high memory demands of the text-free browsing interface. The authors suggested that keyword search and/or browsing of a limited information set could be a potentially more feasible approach.

Medhi et al. at Microsoft Research, India, have done significant work on text-free graphical interfaces (Medhi, Sagar & Toyama, 2006). A number of GUI-specific design principles are given, including the need for audio feedback on all functional units on the display, and the importance of providing help at any time during the interaction. Usability tests on the actual applications showed that users (non-literate sweepers in an Indian urban slum) were generally successful (task success between 50-100%) in completing the given tasks using the text-free interfaces, and were completely unable to do so using the text-based one. More recent work (Medhi & Toyama, 2007) shows the importance of providing real-world context to low-literate users when training them on an interface.

There are a number of other projects investigating interfaces for non-literate users, such as Livestock Guru (Heffernan, 2006), which provides livestock information to rural users in India and Bolivia, and PCtvt, which seeks to create a compelling interface that enables access to television, video, telephony and regular PC use, for semi-literate and non-literate users (Reddy, 2004).

There has been some work on speech interfaces for low-literate users, which is described in the next section.

2.3. Speech Technologies and Applications

Speech interface research has resulted in a number of systems in various domains. There is 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 (Rudnicky et al., 1999) and MIT’s Jupiter weather information system (Zue et al., 2000) 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 (Williams & Witt, 2004). Systems such as Speech Graffiti (Rosenfeld, Olsen & Rudnicky, 2000) require training for users to learn a small set of standardized keywords and interaction methods, which can then be used for any number of speech applications.

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 than the populations in developing-world countries, there are entire domains that are still unexplored (e.g., access to large amounts of text information 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 (Sherwani & Rosenfeld, 2008; Barnard, Plauche & Davel, 2008; Weber, Bali, Rosenfed & Toyama, 2008).

The Tamil Market project by UC Berkeley’s TIER group was the first to design, develop and test a spoken language system with low literate users in a domain (agricultural information) relevant to them (Plauche et al., 2006). 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, Brewer et al (2006) give 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. No research study to date has shown a significant difference between the use of spoken interfaces by literate and low-literate users.

Shortly after the Tamil Market project, the Local Language Speech Technology Initiative conducted similar research in rural Kenya (Nasfors, 2007; Gakuru & Tucker, 2009), though it used touch-tone inputs exclusively, without any use of speech recognition for input. While their project was targeted towards low-literate users, their evaluation was only able to get one such person (out of ten user study participants), highlighting the difficulty of accessing low-literate individuals in user studies.

More recently, and coincidentally with the research described in this thesis, researchers at the Meraka Institute (Grover et al., 2009) have been working on speech and touch-tone interfaces for health information services in South Africa. Their evaluation shows that touch-tone interfaces may be preferable to speech interfaces, even though task success is comparable. A study by IBM Research India comparing speech and touch-tone interfaces suggests that touch-tone is preferable both in terms of subjective and objective metrics (Patel et al., 2009). Taken together, these studies appear to suggest that speech interfaces may not be very useful for low-literate users in the developing world, and that touch-tone interfaces are be better. The research presented in this thesis strongly contradicts these findings.

In terms of user preference for speech versus other modalities, Rudnicky (1993) suggests that literate users prefer speech even when it is less optimal than other modalities. However, a study by Sherwani et al. (2007) comparing VoicePedia (a purely telephone-based speech interface for searching, navigating and accessing the entire Wikipedia) with a GUI-based smartphone equivalent, shows that highly literate users invariably preferred the GUI alternative, even when task success is comparable.

3. HealthLine

3.1. Overview of HealthLine

All the research presented in this thesis was conducted as part of the HealthLine project, the goal of which was to design, develop and evaluate a spoken interface for health information access by low-literate community health workers. The site of the research was initially the metropolitan city of Karachi, Pakistan, although the focus later shifted to two field sites in rural Sindh province in Pakistan.

The following section presents results from a need assessment that was conducted to understand the health information needs of community health workers in Karachi.

3.2. Assessment of the Informational Needs of Health Workers

Prior to the development of an information access system prototype, a needs assessment was conducted between February and May 2007, to understand the information needs of community health workers in Karachi, Pakistan. A total of 35 health workers were interviewed, both in urban and peri-urban parts of Karachi. The goals of the study were to understand gaps in health workers' knowledge, to tabulate the topics they wish to learn more about, and their supervisors' perceptions of health workers' knowledge gaps. The needs assessment was guided by an evaluation of the government's Lady Health Worker program (OPML, 2002), which was conducted between 1999 and 2002.

The specific goals of the needs assessment were as follows:

- To understand the range of health workers operating at the community-level in Karachi, and the nature of the health services that they provide
- To grasp the perceptions of health workers regarding the prominent health problems facing their community
- To identify the specific areas of health in which information is most needed by community health workers
- To assess the Urdu reading and comprehension abilities of health workers, as a means to evaluate their literacy skills
- To explore the preferences of health workers regarding the design and use of HealthLine, and gauge its overall need and suitability in the local context

Laypeople were also considered potential users of the system, and so a total of 15 mothers from the target communities were included in the study as well.

3.2.1. Methodology

A cross-sectional survey based on a sample of convenience was conducted, using two structured questionnaires with both close and open-ended questions – one was a detailed questionnaire used in the initial in-depth interviews with 12 health workers to get a richer sense of the lives and practice of health workers, while one was a shorter questionnaire administered to 23 health workers which focused mainly on the key goals of the study. Hence, a total of 35 community health workers were interviewed.

The interviewed health workers belonged to the low-income settlements of Rehri Goth, Sultanabad, Hijrat Colony, Jamkanda, and Shedi Goth, all of which are located in or near Karachi, Pakistan.

Different types of health workers were interviewed, details of which are depicted below:

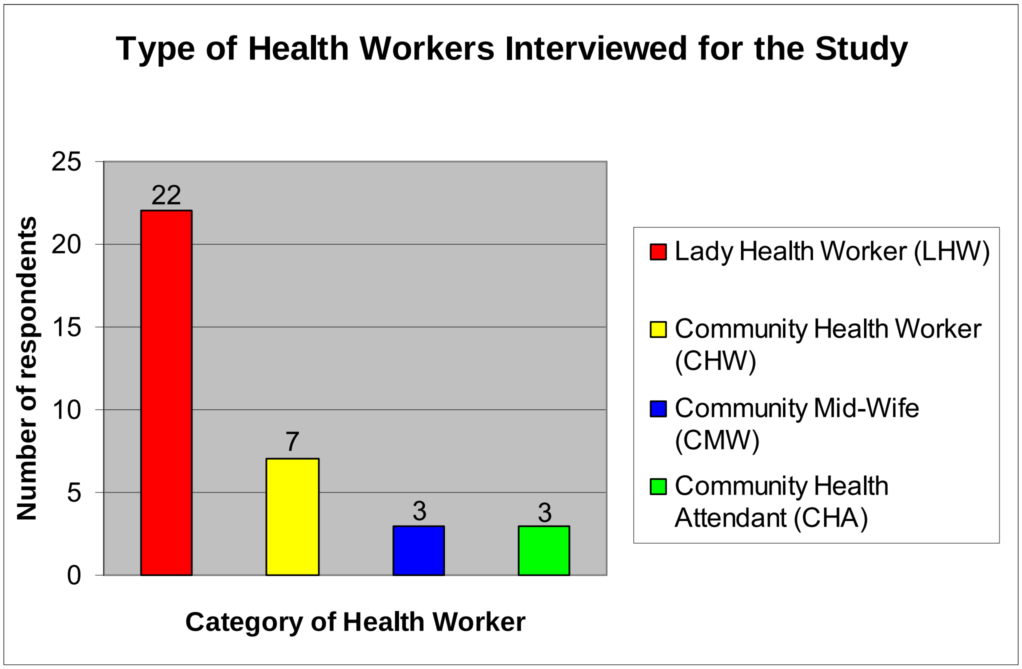


Figure 2: Type of Health Workers Interviewed

The LHWs are government-funded female health workers, and formed the largest group in the study. They have an extensive network in Pakistan with more than 90,000 workers, serving more than 80%of Pakistan's rural population.[2]LHWs receive a basic training of 3 months, in which they learn about various aspects of primary health care and family planning. They constitute the first-level health care givers in the communities in which they operate.

The CHWs are employed by the private organization, Aga Khan University-Hospital, Karachi. Their main role is to facilitate health-related studies and trials conducted by AKU doctors, apart from monitoring health conditions in their communities. Their training varies in its length and content, and depends on the health project that they are required to work on.

Community mid-wives are trained and funded by the non-governmental organization HANDS (Health and Nutrition Development Society). Their focus is on monitoring pregnancies and conducting deliveries. They have received one year of theoretical and practical training by HANDS.

Community health attendants are male health workers funded by HANDS, and focus primarily on child health. They have also received one year of training by HANDS.

The interview questionnaire covered topics including:

- The basic profile of the health worker, including age, marital status, and education

[2] http://www.procor.org/research/research_show.htm?doc_id=729543.

- The nature of their health service e.g. number of years in practice, number of clients served
- Their perception of common complaints and illnesses in their communities
- Their evaluation of how difficult key health topics are for them to understand and address
- Their health information needs
- Their access to phones
- Their preferences regarding HealthLine e.g. would they prefer it as a learning tool or as a service for immediate patient assistance
- Their literacy and comprehension skills

Apart from structured interviews with health workers, interviews were also conducted with key informants including medical doctors involved in research related to public health, town health officers, and LHW supervisors.

3.2.2. Profile of Interviewed Health Workers

Sex

The emphasis of the study was on female health workers, as women tend to be the most common health service providers at the community level. 32 out of the 35 health workers who were interviewed were female, while the remaining 3 were male.

Age

The age profile of the interviewed health workers is presented in the chart below:

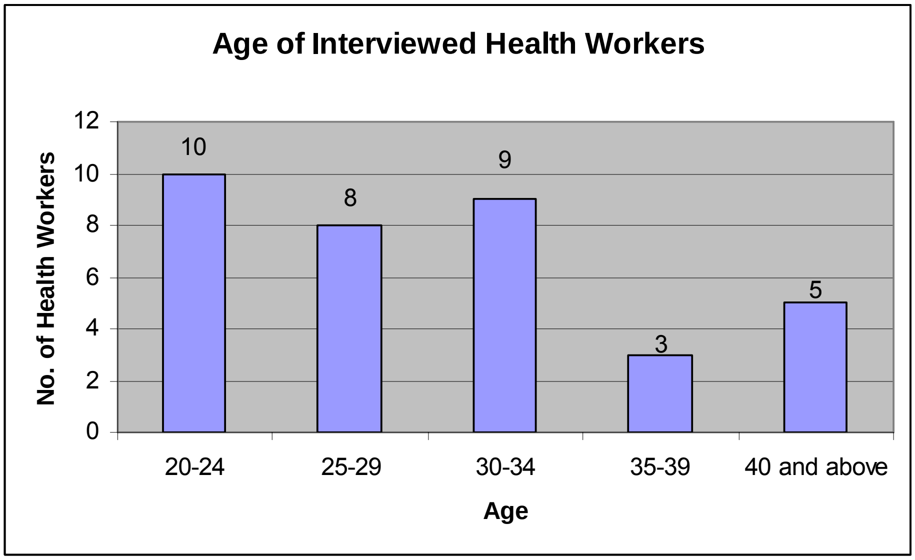


Figure 3: Age Distribution of Interviewed Health Workers

The mean age was 30.3 years.

Marital Status

Out of the 35 health workers who were interviewed, 45.7% were married, 40% were single, while 14.3% were widowed, divorced, or separated.

Level of Schooling

The chart below depicts the years of schooling completed by the interviewed health workers:

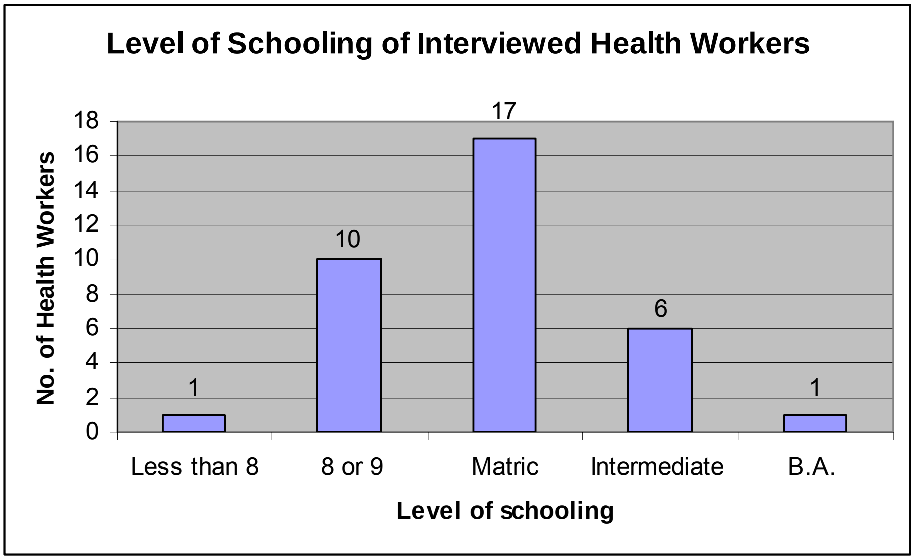


Figure 4: Level of Schooling of Interviewed Health Workers

Almost 70% have completed schooling till 10th grade or above, pointing to a potentially high level of literacy amongst health workers.

Nature of Service

On average, the health workers interviewed for the study have been serving in the area of community health for 6.1 years. The mean number of people whom they provide health services to is 1100, and the mean number of hours that they work in a day is 4.2. According to health workers' perceptions, on average 10 people visit their house for advice or treatment in a week.

3.2.3. Use of Written Material on Health

Leaving out the AKU-affiliated CHWs, the 28 other health workers who were interviewed for the study had received a written manual with their training which serves as the primary reference guide for their practice. All of them said that they still have a copy of their training manual, but that they do not use it frequently. Nevertheless, we felt it was important to understand the information needs for which health workers did use their training material, as it would indicate areas in which HealthLine could also be a potential source of knowledge reinforcement.

The table below shows the kinds of health topics for which the interviewed health workers most recently consulted their manual:

Health Issue/Question	No. of Respondents
Management of Diarrhoea - e.g. dose of O.R.S. according to child age and weight	5
Management of Pneumonia - e.g. dose of cotrimaxazole according to child age and weight	3
Pregnancy care - e.g. what to do if a pregnant woman's BP goes high and she has asthma?	3
Family Planning - e.g. under what conditions should a woman abstain from taking injections for birth control	3
Hepatitis	1
Treatment of worms	1
Eclampsia	1
Delivery complications	1

Figure 5: Use of Training Manual by Health Workers

The table helps to outline the specific areas of child health (diarrhea and pneumonia) and maternal health (pregnancy care and family planning) in which health workers feel the need to revisit their training material. These areas constitute key aspects of the mandate of most health workers, and hence the reinforcement of knowledge in these areas is especially important. Additionally, from the above table we also get a sense of the type of information within these areas that health workers find most difficult to remember and address e.g. dosage information.

3.2.4. Access to Phones

The nature of phone access was also investigated through the interview process, as it is important to understand a) whether health workers have the facility of phones to be able to use HealthLine, and b) the type of phone (mobile or land line) that they are most likely to use to call HealthLine. The charts below present the details regarding the access to phones for 35 health workers:

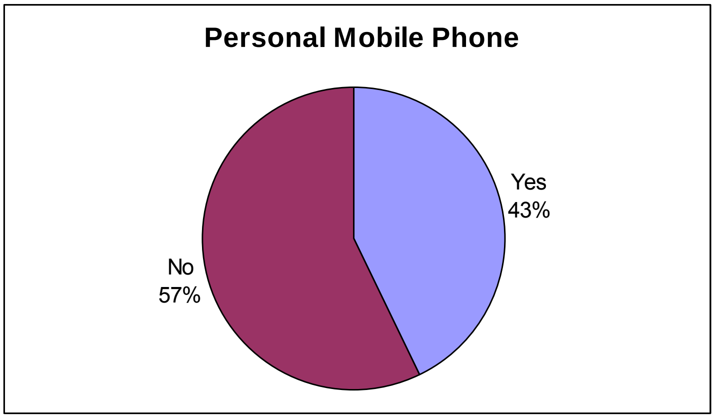


Figure 6: Health Workers' Access to a Personal Mobile Phone

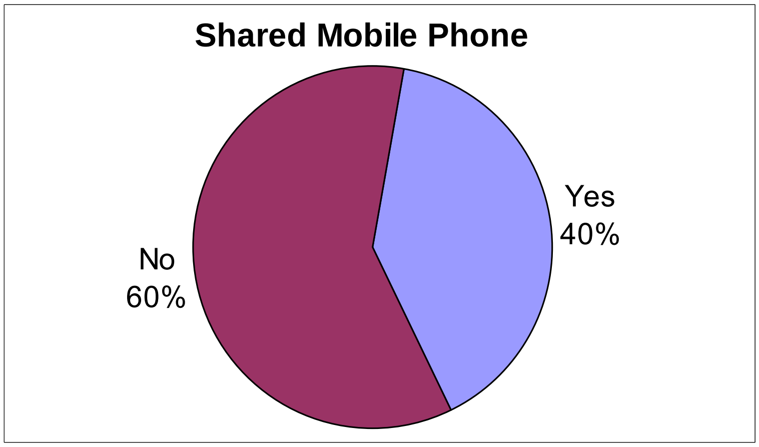


Figure 7: Access to a Shared Mobile Phone

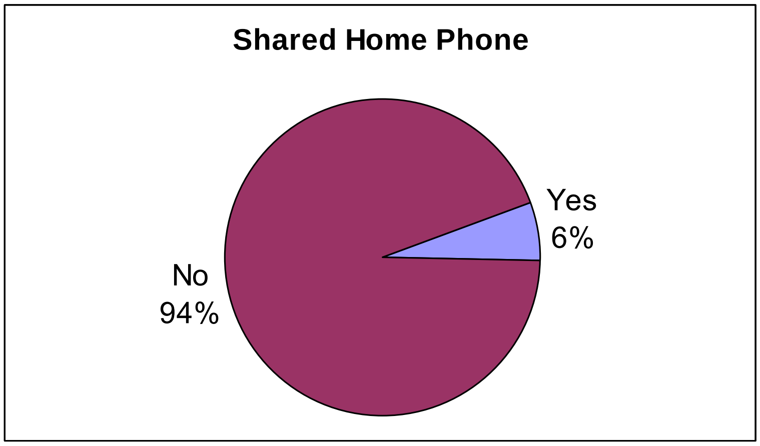


Figure 8: Access to a Shared Home Phone (Land Line)

22.9% of health workers i.e. 8 out of the 35 health workers who were interviewed said that they did not have access to a personal cell phone, shared cell phone, or a shared home phone (land line). Hence, 77.1% had access to some form of phone service, which is more likely to be a cell phone.

Health workers can also have access to a phone service by using a public call office or PCO. The pattern for PCO use is depicted below:

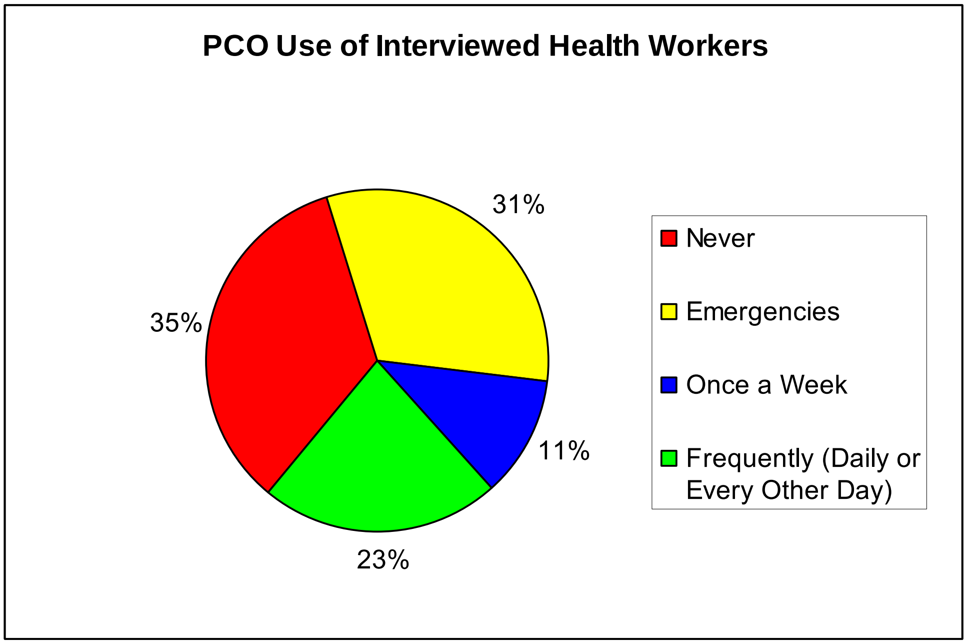


Figure 9: Health Workers' Use of PCO phone

Hence, 65.7% of health workers either never use the PCO, or use it only in the case of emergencies. One might expect that the PCO use of the 8 workers who had no access to a phone is higher than that of the rest of health workers, but this is not the case; only 1 of the 8 uses the PCO frequently while the rest never use it or use it in emergencies.

3.2.5. Attitude towards HealthLine

All the 35 interviewed health workers said that HealthLinewould be a useful medium for addressing their information needs. These health workers were also asked whether they would prefer a system which can provide immediate help while dealing with patients, one that they can call for enhancing their own knowledge, or one that would allow them to do both. The responses to this question are summarized in the chart below:

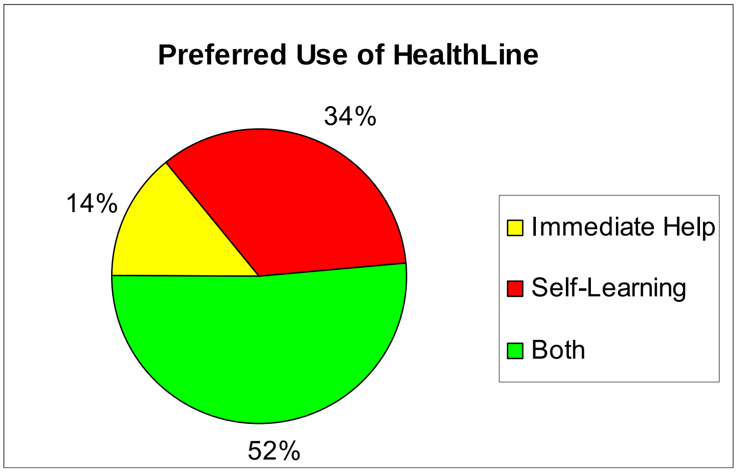


Figure 10: Health Worker Preferences regarding Type of Use

As is evident, a little more than half the respondents would prefer a system that provides the dual facility of immediate help and self-learning. At the same time, those who made a choice between the two options tended to opt predominantly for a HealthLine that focuses on self-learning.

Health workers who preferred a system that would provide on-the-spot information for addressing patient queries and emergencies cited a number of reasons for their choice. One health worker remarked:

"In many circumstances, facing patients can be overwhelming. Sometimes, my mind goes blank and I can't think properly. I can't even remember the information that I already know." - Lady Health Worker, Hijrat Colony

Another one pointed out:

"Information that can help us deal with practical situations promptly is most valuable for us."- Lady Health Worker, Rehri Goth

During detailed interviews with 12 health workers, further preferences regarding HealthLine were investigated. Health workers were asked whether they would prefer to call HealthLine in front of the patient, or in her/his absence by going to a separate room. 75% of the respondents said that they would not prefer to call HealthLine in front of the patient, citing reasons such as:

- The patient would think that the health worker is incompetent
- The information might be troubling or not relevant, making the patient all the more anxious

Amongst the 25% who prefer to call the system in front of the patient, the reasons that were cited include:

- Dealing with the situation might require certain questions to be asked of the patient, and hence the patient's presence in the same room is necessary

- The patient would feel grateful that the health worker is double-checking her/his knowledge

Generally, health workers who expressed no discomfort with calling HealthLine in front of the patient tended to come across as confident of their abilities, and had also been in the profession of community health service for a long time.

During detailed interviews, health workers were also asked whether they would still use HealthLine if it made mistakes in understanding their spoken query, and all but one said this would not hinder them from using the system. With regard to the time of day when health workers would like to use HealthLine, all but one said that night time would be most suitable.

3.2.6. Urdu Reading & Comprehension Ability

Reading Ability

A key purpose of this study was to investigate the ease with which health workers can read and understand written text in Urdu. To explore this, 20 health workers were asked to read a paragraph in Urdu as part of the interview. The paragraph was on the topic of "shock" and was taken from the LHW manual. The reading skills of all the 35 health workers could not be personally observed, as some health workers could not be individually interviewed (due to their busy schedules and lack of cooperation from their supervisors) and were hence administered shorter questionnaires in a group setting.

5 out of the 20 health workers (25%) who were asked to read a paragraph in Urdu during the interview had difficulty in reading the text, which meant that one or more of the following factors applied to them:

- Took a much longer time to read the text as well as making mistakes

- Took reasonable time to read the text but skipping 1-3 words at a time, while making mistakes.

- Inability to read slightly difficult Urdu words in the text, e.g. ("manfi" which means negative, and "radd-e-amal" which means reaction). Two health workers who faced this difficulty proceeded to ask me the meaning of these words.

- Replaced the end of sentences (where the verb occurs in Urdu) with their best guess, suggesting that these health workers try to get the gist of the sentence by concentrating on the middle and fill in the rest.

The 5 health workers who faced such difficulties had completed varying levels of schooling: 1 had schooling below 8th grade,[3] 1 had finished 8th or 9th grade, while 3 had studied up till matric (grade 10th). This demonstrates that even those with higher level of schooling can face difficulty in reading written material in Urdu.

[3] This is below the requirement for qualifying to be a Lady Health Worker, for which a candidate has to complete at least 8th grade.

Comprehension Ability

35 health workers were administered a comprehension test as part of their interview, in which they had to read a paragraph and answer two questions based on it. The test was changed after interviewing the first 12 workers – since the reading passage was on a topic (weaning) that workers were very familiar with, health workers (10 out of 12) strongly tended to give generic answers which were not based on the text, but rather, on health workers' prior knowledge.[4] Hence, it was not possible to make any assessment about their level of comprehension. Consequently, a new comprehension test was designed which was on the topic of "shock", and administered to the remaining 23 health workers (16 LHWs, 7 CHWs). This text was also taken from the LHW manual, but was chosen because it covered an under-emphasized aspect of the LHW curriculum. Indeed, 13 out of the 16 LHWs who were tested on their Urdu comprehension were unable to recognize that the text was from their own training manual. This indicates that a potentially large number of LHWs are not deeply familiar with their basic training material.

Based on the given text about emotional shock, health workers were asked to answer two questions:

1. According to the text, what are two physical symptoms which can be present in the body of a person suffering from shock? (2 points).

2. According to the text, should a person suffering from shock be given warm water or cold water? (1 point).

In the instructions for the test as well, it was emphasized that answers had to be based on the text, not on health workers' prior knowledge. The answer to question 2 was a counterintuitive one, as it seems to be commonly believed that a person in shock should be given colder fluids to drink. When it was observed that some health workers were instinctively giving the answer, they were asked to re-read the passage carefully and then give the answer. *It might still be the case that some health workers got this question wrong because they were answering from prior knowledge, not because of lack of comprehension.*

A total of 3 points could be gained in the comprehension test. All the health workers scored either a "3", "2", "1.5", or "1" on the test. Importantly, no health worker scored a 0. The scores have been classified according to the following criteria:

- 3 points: "Very Good"
- 2 points: "Good"
- 1.5 points: "Average"
- 1 point: "Poor"

[4] This could also be due to methodological difficulties in administering a comprehension test. Several interviewees were nervous while reading in front of me, and tended to answer questions *promptly* rather than *thoughtfully*.

The following chart demonstrates how health workers performed in the comprehension test:

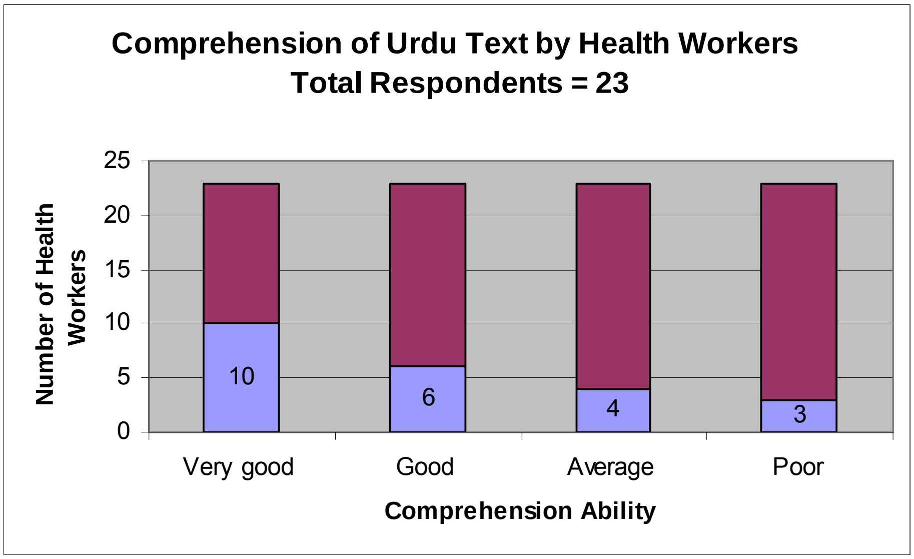


Figure 11: Comprehension Ability of Interviewed Health Workers

As is evident, 10/23 i.e. 43.5% health workers scored a full 3 points, demonstrating very good comprehension ability. If the health workers who achieved between 2 and 3 points are added up, then almost 70% of health workers who were interviewed can be said to have reasonably good comprehension ability.

Hence, the 30% who face difficulties in comprehending Urdu text could potentially benefit more from HealthLine than those who can comprehend written Urdu with relative ease.

It also needs to be noted that those health workers who had average-poor comprehension ability in Urdu did not necessarily have lower levels of schooling. As the chart below shows, there were some workers who had completed 10 and 12 years of schooling but still performed poorly on the comprehension test:

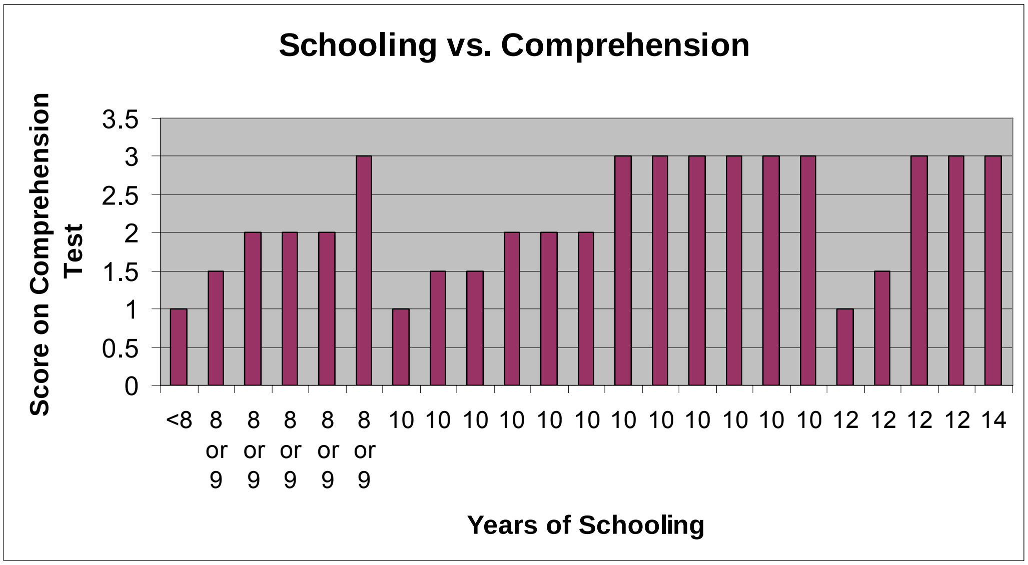


Figure 12: Comprehension Ability for Health Workers with varying Levels of Schooling

3.2.7. Self-Perceived Information Needs of Health Workers

35 health workers were asked to identify the three most important areas of health about which they lack information, and would like to have more training. This was an open-ended question. Respondents suggested between one and four topics, often specifying those health concerns for which they are regularly contacted by their clients, but feel unable to help them deal with these concerns. Hence, this question also gives a glimpse of health issues that are prevalent in the community, and yet go unaddressed by community health workers.

The results of this question are encapsulated in the following chart:

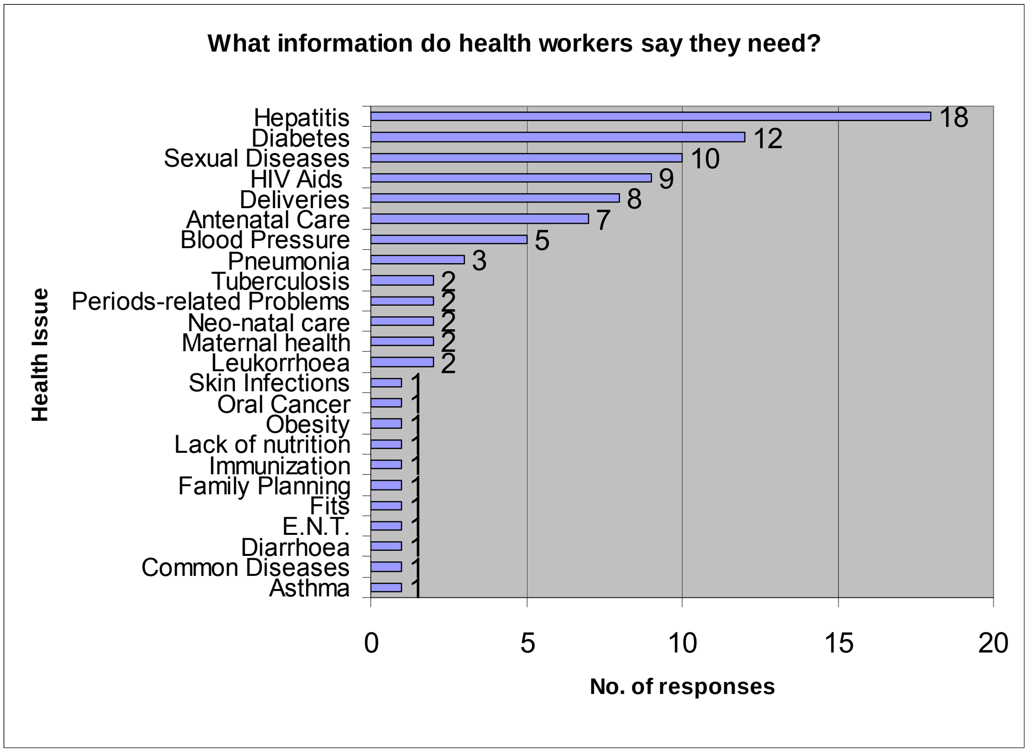


Figure 13: Information Needs of Community Health Workers

As is evident, there is a clear and compelling need for more information on hepatitis. The second topic on which health workers need more information on is diabetes, followed closely by sexual diseases. The data was analyzed to check if a particular information need was limited to a specific place or type of health worker. However, it was concluded that the information on the above topics was required across the board.

While the above chart gives the whole range of topics on which health information is needed, it might also be useful to combine similar topics together to get a broader perspective. Two overarching categories of health information that can be developed are:

- Sexual Health, which includes information on sexually transmitted diseases such as HIV AIDS, vaginal infections such as leukorrhoea, periods-related problems, and family planning
- Pregnancy and Deliveries, which includes information on antenatal care, complications regarding deliveries, as well as maternal health[5]

When the health information needs are reorganized using these two categories, the following picture emerges to give us the four main areas of health on which information is needed:

[5] In the context of the LHW manual as well as this study, “maternal health” most often denotes the care of a pregnant woman.

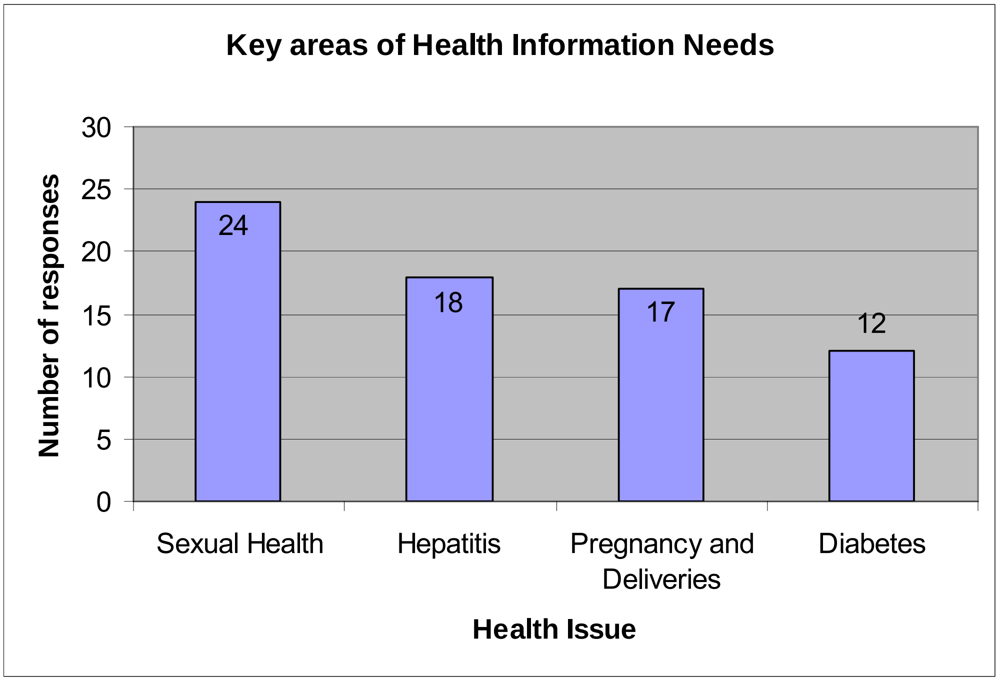


Figure 14: Three main areas of Information Needs for Health Workers

What is striking about the above chart is that sexual health and pregnancy care already constitute key areas in which LHWs and other community health workers provide their services, and yet they feel that they do not have adequate knowledge about these topics. While LHWs have basic knowledge about deliveries, what they strongly desire is more detailed practical training on the topic so that they may be able to undertake deliveries themselves and handle its complications. However, this is an aspect of maternal health which is not included in the LHW mandate. Diabetes is also a topic which is not part of the mandate of any health worker interviewed for this study, and hence community health workers generally have no information or training on this subject. Yet it seems to be a growing concern in the communities they serve, leading them to desire more information on it. I even met a health worker herself who was suffering from diabetes, and was distressed about how to deal with it.

It is interesting to see how the health information needs revealed by this study are covered in the LHW manual. Diabetes is completely absent from it, as it is not part of the LHW curriculum. Hepatitis – which constitutes the single largest area in which health workers say they need information – is devoted less than half a page in a 361-page manual. A wide variety of sexual diseases are covered hastily in two pages, with the exception of AIDS which gets 5 pages by itself. *This lack of written material and coverage might be one reason why health workers most need information on these topics.*

3.2.8. Self-Perceived Difficulty Level of Key Health Topics

35 health workers were asked to rate the difficulty level of key health topics i.e. maternal health, nutrition, family planning, immunization, diarrhoea, pneumonia, and sexual diseases, on a Likert scale of 1 -5 where 1 represented "Very Easy" while 5 represented "Very Difficult". These topics were pre-determined as "key" because of their importance to LHW practice, and hence, this was a close-ended question. "Difficulty level" was defined in terms of how tough and challenging different topics are for health workers to understand, remember, and practice.

In the following charts, the difficulty level is first depicted in terms of total responses for each rating, and then in terms of the average difficulty rating for each topic:

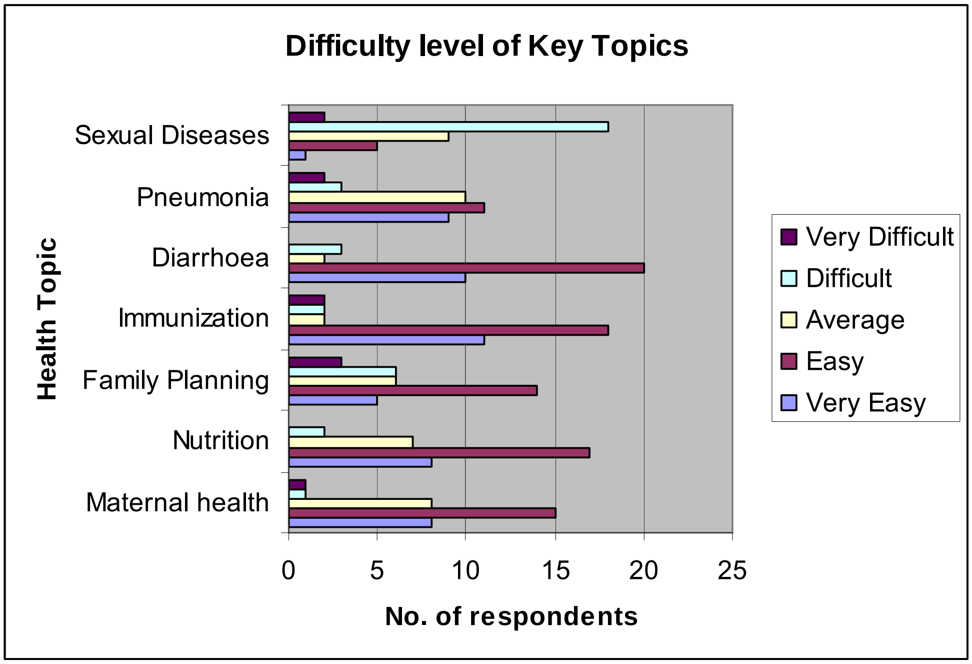
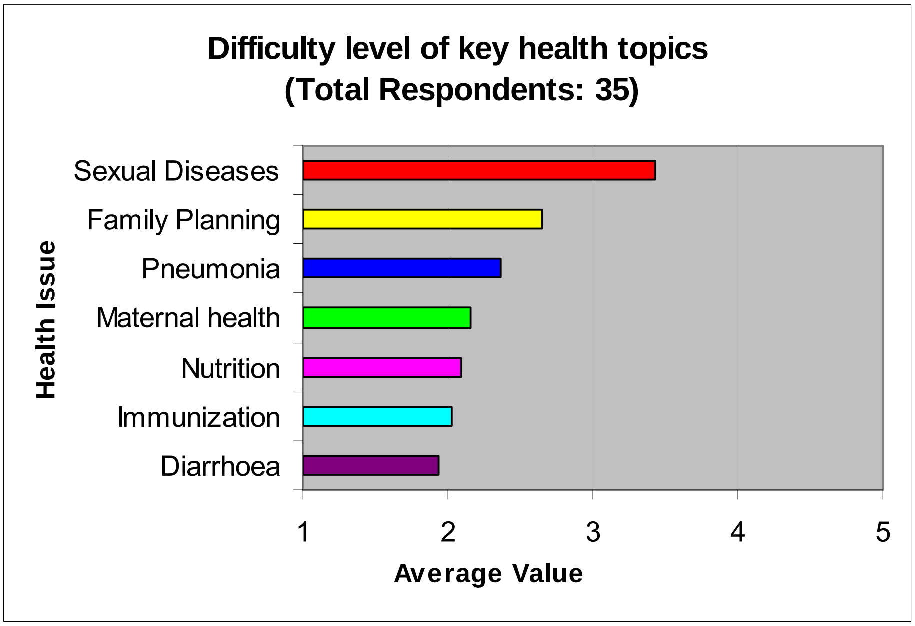


Figure 15: The Difficulty Level of Key Health Topics for Health Workers



Topic	Mean	Standard Deviation
Sexual Diseases	3.43	0.92
Family Planning	2.65	1.20
Pneumonia	2.37	1.14
Maternal Health	2.15	0.94
Nutrition	2.09	0.83
Immunization	2.03	1.07
Diarrhoea	1.94	0.84

Figure 16: Average Difficulty Value for Specific Health Topics

In terms of difficulty level, sexual diseases emerge as the most difficult topic with a mean value of 3.43 on the Likert scale. Put differently, 82.9% of health workers said that the difficulty level of sexual diseases was average, difficult, or very difficult. This corroborates the results of the previous section, in which sexual diseases emerged as a key area in which health workers need more information. The kinds of questions within the area of sexual health that health workers find difficult include:

- How to deal with vaginal discharge that has a foul smell
- How to deal with rashes, itching, and burning in the vaginal area
- What to do if periods are irregular, too heavy, or too painful

Because the above queries are quite common, some health workers have started prescribing their own treatments for them such as antibiotics and homeopathic medicine instead of referring patients to a health center (as they are currently required to do).[6] This practice might well be contributing to ill-health of patients and

[6] LHWs in particularly resent the practice of "referrals" as this makes patients less likely to come to an LHW for advice and treatment.

misinformation about sexual diseases, making the provision of training in this area all the more necessary.

After sexual diseases, the topic of family planning is the most difficult for health workers as 44.1% of them rated its difficulty level as average, difficult, or very difficult. The challenging concerns with respect to this topic are:

- Under what conditions should women abstain from using hormonal pills or injections for birth control

- How to address women's concerns about the side effects of birth control methods e.g. community women feel that the use of IUCD leads to a high risk of miscarriages after being discontinued

While pneumonia is a topic that most health workers have had considerable training in – and hence, many do not perceive a strong need for information on it as demonstrated in Figure 1 – it is still a topic which health workers find difficult to address in their daily practice as 42.9% of them rated its difficulty level as average, difficult, or very difficult. Hence, pneumonia should also be seen as a topic in which there is a need for information access and knowledge reinforcement. The types of difficulty faced with respect to this topic are:

- Mistakes made in applying the IMCI protocol for ARI management e.g. applying the classification algorithm for a three month old child to one who is six months old.

- Inability to remember and administer the dose of cotrimaxazole according to severity of condition, child age, and child weight

3.2.9. Performance-based Difficulties in Addressing Health Issues

Knowledge Test Administered by Oxford Policy Management

There might of course be a difference in the kinds of topics that health workers say they find difficult, and those in which health workers fall short in practice i.e. the actual difficulty in performance. The performance-based difficulties faced by LHWs can be gauged from the comprehensive evaluation of the LHW programme that Oxford Policy Management conducted in 2002. As part of this study, a knowledge test was administered to 500 LHWs from across Pakistan. Through this test, specific areas of weakness in the knowledge base of LHWs were identified, mainly:

- Inability to remember the immunization schedule – less than half the LHWs interviewed could name the four EPI vaccines (BCG, DPT, polio, measles), give the correct number of doses, and identify the correct age at which the doses are given.

- Lack of knowledge about the correct doses of medicines for common illnesses. For example, less than one quarter of LHWs interviewed could state the correct dose and duration for a course of Cotrimoxazole for a child with pneumonia, even though they were encouraged to use their training manual and medicine box to answer the question.

- Inability to correctly conduct case-based analysis of diarrhoea and pneumonia using the classification and management protocols that LHWs have been trained in. For example, 79% could not identify simple pneumonia. 35% could not identify a case of severe dehydration, while 29% could not identify severe pneumonia.

Knowledge Test Administered as part of HealthLine Needs Assessment Study

Knowledge-based questions were also posed during in-depth interviews of 12 health workers. They were asked about danger signs during pregnancy, danger signs of newborns, vaccination schedules, and case-based management of diarrhea and pneumonia. Case-based analysis was most difficult for health workers, with 66% of health workers unable to diagnose a case of severe pneumonia and 66% unable to diagnose a case of simple pneumonia. Health workers tended to classify severe pneumonia as simple, and simple pneumonia as a case of cough/cold. Interviewed health workers fared better in diarrhoea management, as a significantly less percentage of 16.7% were unable to classify a case of mild to moderate dehydration, and the same percentage of 16.7% were unable to identify a case of severe dehydration. However, it was not the case that health workers who could not diagnose simple pneumonia or diarrhea were the same as those who could not diagnose severe cases of these illnesses.

3.2.10. Health Workers' Perceptions of Common Illnesses amongst Adults

Health workers interviewed for the study were asked to identify the three main complaints and illnesses that adults in their communities faced. These problems are mostly faced by adult women, as most of the clients that the interviewed health workers serve are women. The question was asked to get a sense of health workers' perception of what health problems are prevalent amongst the communities in which they serve. It might also be another way of establishing what health information is "needed", as presumably, information on common illnesses would be considered useful by health workers and community members alike.

The chart below summarizes the responses to this question:

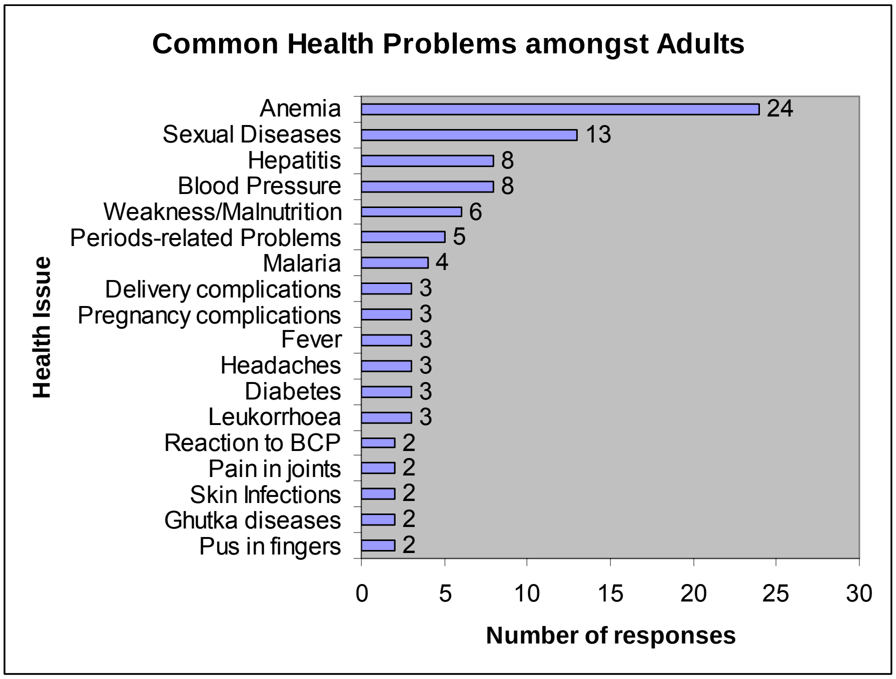


Figure 17: Common Health Concerns of Adults in the communities served by Health Workers

Across the board, health workers identified anemia as a major disease afflicting the health of women, particularly during pregnancy. Anemia was often mentioned alongside weakness and malnutrition, and in fact, expressedly linked to it. Information about addressing anemia seems to be commonly present amongst the health workers who were interviewed, as none of them identified anemia as a topic on which they need more information (see Figure 1). Sexual diseases and hepatitis again emerge as key health concerns, now on the basis of their prevalence in the community. Blood pressure is revealed as another common complaint amongst adult women, and is also a topic that was moderately high in the list of topics that health workers need information on (see Figure 1). The LHW manual does not directly cover information on blood pressure.

The last two health complaints depicted in the chart – “Ghutka diseases” and “Pus in fingers” – were both reported in the fishing village of Rehri Goth. *Ghutka* is a mixture of betel nut and tobacco that is ever-present in the mouths of men and women in this village. Like other forms of tobacco, ghutka is also detrimental for health; a number of health workers I spoke to even linked the presence of anemia and cancer in the community to the pervasive use of ghutka. “Pus in fingers” is a condition affecting women who are in the occupation of shrimp-peeling and de-veining.

3.2.11. Mothers as Potential Users of HealthLine

Rationale

During the data collection period of this study, it was felt that the education levels of LHWs and other community-based health workers do not capture the literacy range needed to test HealthLine, as 97% of the interviewed health workers had achieved schooling between 8th grade and graduation, and 70% of them could comprehend written Urdu sufficiently well. For HealthLine, it is imperative to access a range of users with diverse literacy and comprehension levels as the appropriateness and usability of the system needs to be gauged for users with varying literacy levels. Moreover, users with lower levels of literacy and comprehension are particularly significant as the need and incentive for a phone-based system might be much more for them than for literate users.

After consultation with medical researchers and the field staff at the Community Health Sciences Department, AKU, it was decided that mothers who have had no schooling or very little schooling would be included in the study to ensure that potential users of HealthLine encompass a wider literacy range. From the perspective of public health, it would also be beneficial to use HealthLine to directly give health information to mothers as they are the first providers of care. Indeed, it has been suggested that health education for mothers is indispensable for reducing child mortality due to diarrhea and acute respiratory infections in the slums of Karachi.[7]

Profile of Interviewed Mothers

A total of 15 mothers were interviewed – 5 each from the squatter neighborhoods of Sultanabad, Hijrat Colony, and Rehri Goth – to assess how their health information needs might vary from those of health workers in the same communities. The interviews with mothers were much shorter than the ones that were conducted with health workers, focused on a few key questions pertaining to their health problems and health information needs.

The average age of the mothers who were interviewed was 34. 13 out of the 15 interviewed mothers were homemakers. The other two were volunteers at the AKU-managed Primary Health Care centre in their communities. All mothers except one said that they had access to phone through a landline or mobile service.

The level of schooling of these mothers is depicted in the following figure:

[7] D'Souza RM. "Role of health-seeking behaviour in child mortality in the slums of Karachi, Pakistan." *Journal of Biosocial Science* 2003; 35(1):131-44.

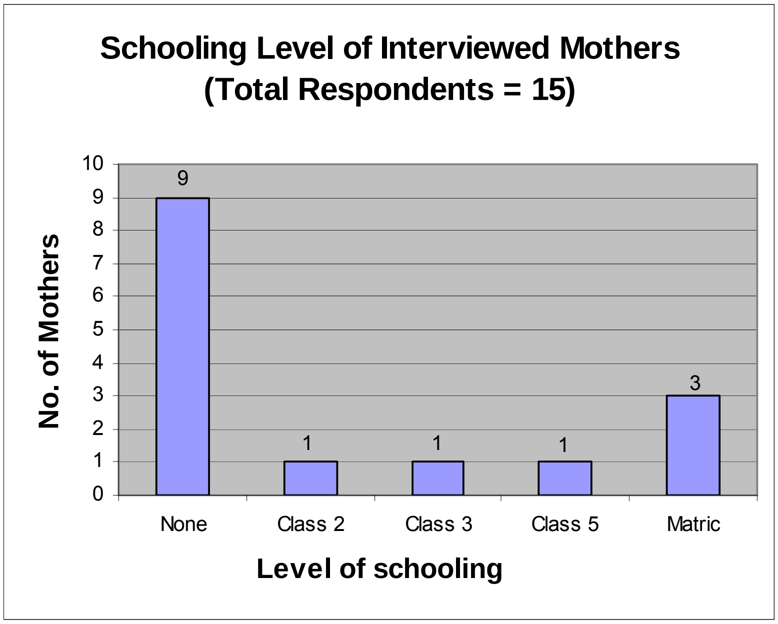


Figure 18: Schooling Level of Interviewed Mothers

Hence, 60% of mothers interviewed for the study have no formal schooling.

For a variety of reasons, a formalized literacy and comprehension test could not be conducted for mothers who were interviewed for the study. To begin with, the interaction with mothers had to be limited in time because most of them were preoccupied with household and child-rearing tasks. Moreover, mothers tended to be more shy in their demeanor as compared to health workers, and hesitant to talk at length. Lastly, given the complete lack of schooling of most of the mothers, it seemed not only pointless but also condescending to administer a literacy test to them.

However, an attempt was made to gauge the literacy and comprehension level of mothers through narrative questioning. 3 out of the 9 mothers who had received no schooling said that they were still able to write their name and read basic numbers, while the rest said that they were completely illiterate. 3 mothers had attended school up till class 5, and said they could write their names, read basic Urdu, and do elementary math. The remaining 3 mothers said that they had completed studies till matriculation (grade 10); however, one of them said that her schooling was in Sindhi so she could barely read Urdu, one was able to read a given Urdu text with significant difficulty, while the last one said that she could read basic Urdu and Sindhi but had not read in a long time due to being completely involved in household duties. In effect, even the mothers who had completed matric did not seem to possess the literacy and comprehension skills associated with this level of schooling.

Overall, it was felt that for home-based mothers, accessing written material is not a viable option for increasing their health-related knowledge.

Threshold for Seeking Healthcare

The 15 mothers interviewed for the study were asked about when they themselves or their families decide to visit a doctor for addressing a health concern. Half of them emphasized that for their child's health, they go to a doctor immediately. As one mother commented,

"I take my child to a doctor on the first day of illness, even if I have to borrow money to pay for the doctor's fees." - Mother, Sultanabad

The remaining half commented that for their child's health, they first focus on home-based remedies and then visit a health center if symptoms continue to persist.

Apart from one respondent, all mothers said that they do not go to a doctor immediately for their own health. They go when they "*are completely unable to move"* or when they "*get a big disease."* Hence, it seemed that almost all mothers try to ignore or suppress their health concerns till they reach an unbearable point. Mothers also tend to feel that their domestic duties are too overwhelming to be disrupted by "feeling sick" and going to the doctor – it is just best to take a painkiller or home-based remedies even if one is very sick. Some mothers also commented that even if they are completely bed-ridden, they avoid going to doctors as they prescribe lab tests which are very expensive. It seemed that mothers deem medical treatment "unaffordable" for their own health, but not for their child or husband's health.

In such a context, it is all the more significant to provide health information to mothers about illnesses that afflict their own health, as these concerns are most likely to remain unaddressed for longer periods.

Self-perceived Health Information Needs of Interviewed Mothers

All the mothers who were interviewed said HealthLine would be a good resource for enhancing their health knowledge. However, when asked about what information they would like, most thought long and hard before saying anything. One said:

"I have no idea. I'm like a blind person. My husband takes me everywhere. If you take me out on the street, I wouldn't even know which direction to go in to find the clinic."
- Mother, Hijrat Colony

Others remarked that any knowledge about health issues faced by their family would be useful. Since the mothers interviewed for the study had not received any training in health, they generally did not have a vivid sense of what information they lack as compared to health workers who were easily able to articulate specific topics in which they need more knowledge. Hence, for mothers, it might be useful to follow an "information push" model that communicates important public health messages and information about common illnesses, versus an "information pull" one that caters to specific knowledge needs.

This is not to suggest that mothers were completely unable to share perceptions about their health information needs. Some mothers were particularly articulate, and pointed out up to 4 topics in which they desired more health knowledge. The figure below captures the range of responses received from mothers on this matter:

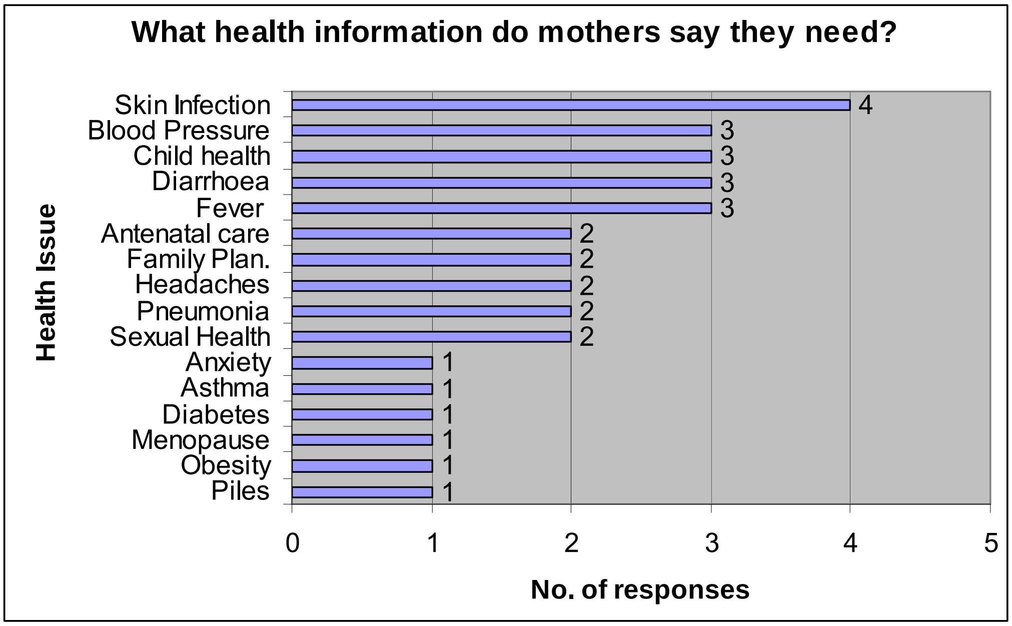


Figure 19: Mothers' Self-perceived Needs for Health Information

No health issue of overriding importance emerges here as compared to the responses of health workers, in which the topics of hepatitis, sexual diseases, pregnancy and deliveries, and diabetes were prominently high in terms of the perceived need for information. This may well be due to the smaller sample size for mothers. Even in this sample of 15, it is striking that not a single mother mentioned the need for more information on hepatitis; this was the leading topic of interest for health workers and the reason that many cited for this interest was the prevalence of hepatitis in the communities in which they serve.

The most important topics for mothers in terms of health information needs are: skin infections, blood pressure, child health, diarrhea, and fever. The highest number of requests for information on skin infections might be an outcome of the fact that they are particularly common among young children at this time of the year, especially in squatter settlements.[8] Generally, child health concerns including skin infections, diarrhea and fever are common concerns of mothers throughout the year, and thus constitute topics on which mothers desire more information. Though not specifically identified in terms of health information needs, the topics of cold/cough and asthma/wheezing can be presumed to be part of the information on child health that mothers desire, as they identify these are particularly common in their families apart from skin infections and fever:

[8] Personal communication with the nursing and health-monitoring staff at AKU-CHS field sites from where mothers were selected for the study.

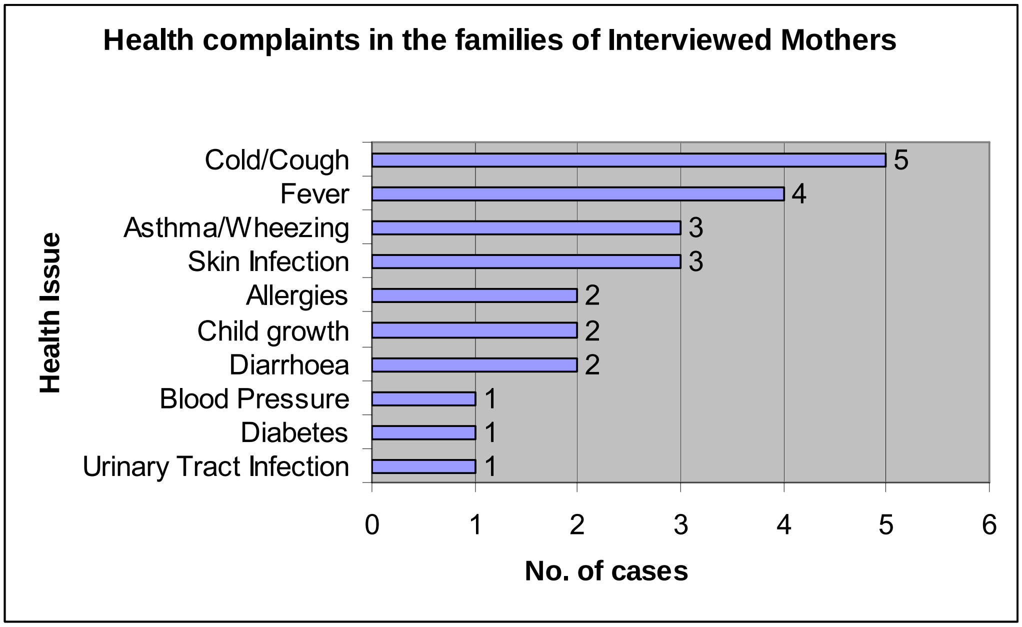


Figure 20: Common Health Problems within the families of Interviewed Mothers

Health Problems faced by Interviewed Mothers

Blood pressure is the only topic in the top five popular topics for health information (see Fig. 18) that mothers requested for their own health needs. This is not surprising given the prevalence of high blood pressure amongst the mothers interviewed, as 6 out of 15 mothers said they were suffering from the condition.[9] The figure below illustrates this, as well as the prevalence of other health complaints faced by the interviewed mothers:

[9] Indeed, hypertension is common in Pakistan in general. According to a pilot study on hypertension being conducted by Dr. Tazeen Jafar (AKU-CHS), one in three persons aged 40 years or above in Pakistan suffers from hypertension.

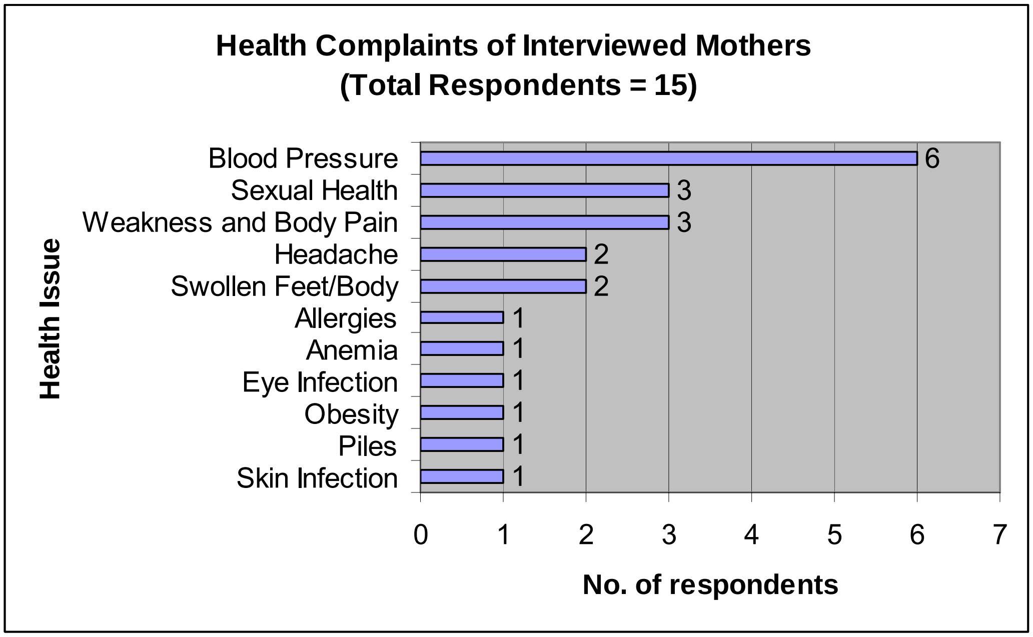


Figure 21: Health Problems faced by Interviewed Mothers

Apart from blood pressure, sexual health as well as generalized weakness/body pain emerge as other prominent health concerns of the interviewed mothers. When health workers were asked about common complaints in their communities, they had also identified blood pressure as a significantly common health concern amongst adult women (8 responses); blood pressure is also a topic that was present in the list of topics that health workers need information on (5 responses).

3.2.12. Key Informants' views on HealthLine

A number of key informants were interviewed to understand their perceptions of HealthLine, and seek their suggestions for making it a valuable intervention. The table below captures their responses to the two principal questions that they were asked:

Informant	Area of High/Unserviced Information Need	Preferred Use of HealthLine
Dr. Gregory Pappas (AKU)	- Reinforcement of the existing LHW curriculum	Self-learning limited to existing training material
Dr. Z.A. Bhutta (AKU)	- Diarrhoea Management - ARI Management	Immediate Help/Hotline
Dr. Anita Zaidi (AKU)	- Diagnosis and triage for common illness - Neo-natal care - Important Contacts e.g. where to refer a drug addict for rehabilitation?	Immediate Help/Hotline; Phone not a good medium for learning

Dr. Sarwat Mirza& Dr. Anjum Fatima (HANDS)	- Standard Protocols Disease Management - Help with case-based analysis	Immediate Help and Self-learning
Zubeida Golani, LHW Supervisor	- Infertility - Pain in knee joints	Self-Learning
Dr. Nausheen Kehar, Town Health Officer	- Dealing with complications of pregnancy	Immediate Help/Hotline

Figure 22: Key Informants' Views on HealthLine

As is evident, the management of common illnesses using standard protocols was emphasized as a key area in which health workers need knowledge reinforcement. There was no agreement as to whether the system should focus on immediate help, or on offline learning. All the key informants stressed that what health workers need most is practical training and experiential learning – something that seems beyond the scope of HealthLine.

3.2.13. Conclusion

To explore the context in which HealthLine will operate and assess the health information needs that it can address, 35 health workers and 15 mothers were interviewed in various low-income neighborhoods of Karachi. A cross-sectional survey based on a sample of convenience was conducted, using structured questionnaires with both close and open-ended questions.

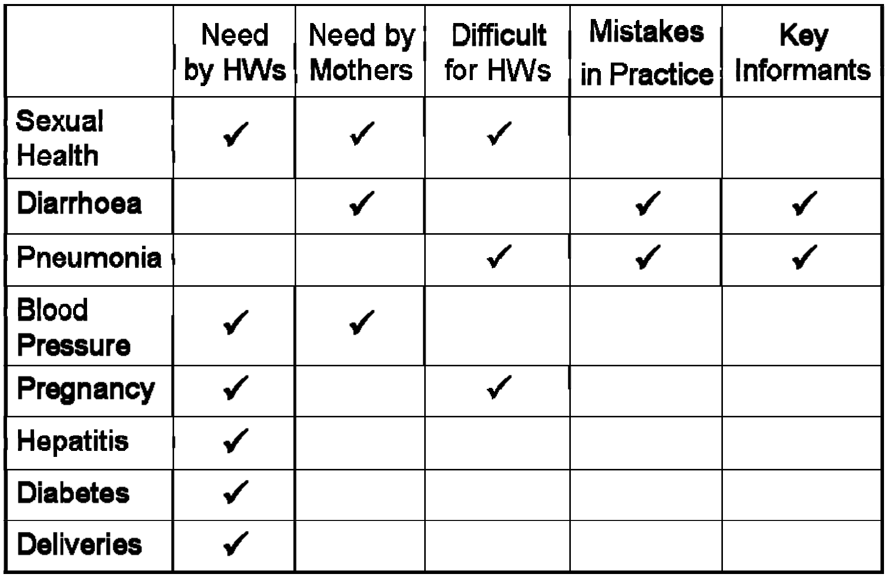
It was discovered that 70% of health workers had achieved schooling up till matric (10th grade) and above. 25% of health workers for whom a personally observed reading test was conducted had significant difficult in reading health-related text in Urdu, whereas 30% of health workers who were asked to answer two questions based on the text exhibited significant difficulty in answering them. This suggests that a significant minority exhibits a lack of comprehension ability; however, the tendency to answer questions based on instinct or prior knowledge could also be a contributing factor.

60% of mothers who were interviewed were completely non-literate; even the 3 who had studied up to matric expressed that they could not read and understand Urdu well. It is thus recommended that mothers are also included as end-users of HealthLine, as they would provide the necessary range to assess the impact of literacy on the use of the speech system.

The focus of HealthLine should be on self-learning, be it the reinforcement of learning material on familiar topics such as on diarrhea and ARI management, or the provision of new material on topics that they lack knowledge in e.g. sexual diseases. It is only with material with relatively new content that we can test whether users are "learning" from HealthLine or not; with familiar material, it will be hard to test whether health workers are learning due to HealthLine or answering based on prior knowledge.

During the reading test, it was discovered that a number of users are unable to read relatively complex Urdu words (used in health manuals for community health workers) and are also unaware of what such words mean. Hence, it is imperative that the content of HealthLine be in simple Urdu, which users are able to understand and comprehend easily as this would facilitate the learning potential of HealthLine.

As part of this study, different criteria was used and investigated to identify the key areas in which HealthLine can provide relevant health information. The chart below summarizes the findings:



	Need by HWs	Need by Mothers	Difficult for HWs	Mistakes in Practice	Key Informants
Sexual Health	✓	✓	✓		
Diarrhoea		✓		✓	✓
Pneumonia			✓	✓	✓
Blood Pressure	✓	✓			
Pregnancy	✓		✓		
Hepatitis	✓				
Diabetes	✓				
Deliveries	✓				

Figure 23: Health Topics according to Different Criteria of Need

Sexual health, blood pressure, and pregnancy are three areas of maternal health that need particular attention, whereas diarrhea and pneumonia are topics in child health that require knowledge reinforcement. Information on hepatitis, diabetes, and deliveries is also strongly needed by health workers.

The final choice of topics will ultimately depend on the quality of material that is available for each. The recommended next step is to identify the material on one of the above topics – preferably a topic like diabetes which health workers have no training on – automate 1-2 pages on it, conduct a pre-test on selected questions, and then examine how health workers/mothers answer these questions after interacting with the system. Moreover, it is equally necessary to investigate the actual process of how health workers hear, understand, interact with, and learn from the system before extending the project to a wider knowledge base and users.

3.3. Health Information Resources & Content Development

The needs assessment conducted presented a clear list of topics on which to focus the development of the system. The primary sources for health material were two Karachi-based public health organizations, the Health and Nutrition Development Society (HANDS) and Aga Khan Health Services Pakistan (AKHSP). The list of material accumulated were as follows:

1. HANDS' pamphlet: "Messages for Community Health Workers" (3 pages)
2. AKHSP pamphlets on diarrhea and pneumonia (4 + 3 pages)
3. The Government Lady Health Worker manual (350 pages)
4. A summary of main messages for Lady Health Workers (20 pages)
5. AKHSP pamphlets on hypertension, diabetes, leucorrhea (8 + 9 + 3 pages)
6. HANDS' Community Midwives' manual (526 pages)
7. Where There Is No Doctor (446 pages)

AKHSP's pamphlets had been designed for literate, urban readers, and had been tested extensively on them. Thus, AKHSP staff cautioned against the use of these materials on audiences that the pamphlets had not been tested on, such as low-literate CHWs. For this reason, changes were made to the text to simplify meaning and sentence construction for oral presentation.

Where There Is No Doctor is a popular health manual created by the Hesperian Foundation, and has been translated into many languages. However, researchers at Aga Khan University, Karachi, expressed concern at using even the Urdu translated material, as it was not fully localized, and hence not completely appropriate to the needs of the target communities. Localization in terms of language, content, socioeconomic and cultural context was seen as a crucial need when choosing a health information source.

4. User Studies

From September 2007 through October 2008, we conducted 5 user studies with community health workers. These are summarized in the following table:

#	Place	Language	Participants	Type of System	Interactive
1	Wahi Pandi	Urdu	9	Menu-driven	✓
2	Memon Goth	Urdu	10	Output-only	
3	Umarkot	Urdu	10	Output-only	
4	Wahi Pandi	Sindhi	10	Output-only	
5	Umarkot	Sindhi	20	Menu-driven	✓

These user studies are described in detail in the following sections.

4.1. Measuring the Baseline: Testing a Menu-Based Dialog System

4.1.1. Prototype System Architecture

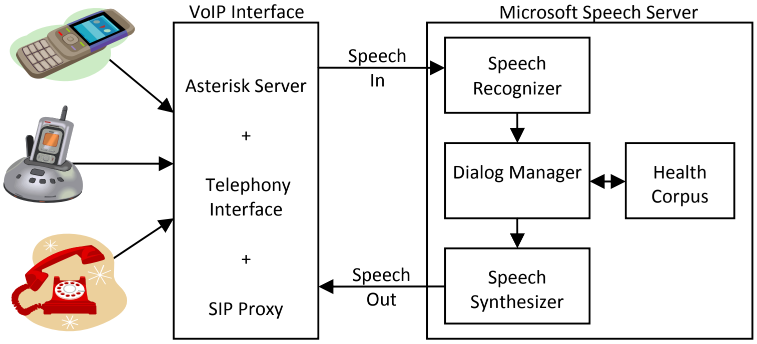
It is noteworthy that in the Tamil Market project (Plauche et al., 2006), considerable effort was spent in creating acoustic models for Tamil, involving recording Tamil speakers speaking out certain words, which proved challenging with low-literate speakers. Additionally, their approach led to a small vocabulary, single-word recognizer – which would limit dialog design choices considerably, especially when using an iterative-design methodology. One significant benefit of their approach was low word error rate (as low as 2%). However, in our work, we decided to try a different approach, one that would minimize the need for training data.

Our solution was to use a commercial package, Microsoft Speech Server 2007 Beta (MSS), which includes a speech recognition engine, a speech synthesis engine, and a dialog management architecture, along with dialog authoring tools. However, MSS supports only a few languages, and did not support Urdu. To solve this problem for speech synthesis, all audio was recorded as individual prompts from an Urdu-speaking voice talent. It was then possible to encode prompt text in Urdu in the code (through Unicode), so that there was no extra layer of mapping required – running an MSS function such as Prompt.AppendText(“سلام“) would work correctly.

For speech recognition, we used our novel Salaam method that leverages the robustness of speech resources in the developed world (e.g. for US English) for recognizing speech in resource-scarce languages. This is described in greater detail in Section 5.

4.1.2. SIP server

To run a dialog system over the phone network, MSS requires a SIP-over-TCP VOIP end-point. All of the MSS-recommended hardware telephony products were digital line-based (e.g., T1/E1), and comparatively expensive ($3000+). Additionally, the costs of setting up an E1 line in Pakistan are significantly higher than setting up an analog line. Asterisk, on the other hand, is a widely used, open-source VOIP platform, which significantly lower hardware costs (<$1000), and also supports analog phone lines. However, Asterisk implements only SIP-over-UDP, and none of the available SIP TCP/UDP proxies we tried were able to connect Asterisk and MSS. Our solution was to write a proxy for this purpose, sending SIP messages over UDP to Asterisk, and over TCP to MSS. Due to complexities in the SIP protocol, it was not adequate to simply relay a message from one end-point to the other – this is why available proxies did not work. Instead, the entire SIP conversation had to be orchestrated as two separate conversations: one between Asterisk and the proxy, and the other between the proxy and MSS. Once the session is initiated through SIP, all audio is passed directly between Asterisk and MSS. The system architecture is shown below.



Other challenges in building the system, relating to Urdu text and speech processing, are described in Sections 7 and 7.2.

4.1.3. Health Information Content

For this study, only the following two sources were chosen from the list of health information: the “Messages for Community Health Workers” manual, and the pamphlets on diarrhea and pneumonia, all of which were in Urdu.

4.1.4. Dialog Design

While the primary research interest was in investigating various alternative dialog strategies, we decided to create a very simple prototype to see how viable it could be without any significant effort. We now present the dialog design used in the prototype. All prompts and valid recognition options have been translated from Urdu to English below. A sample conversation is given below:

Hello, I’m Dr Jameela, and I’m here to give you whatever health information you need. What topic do you want information on: *Diarrhea*, *Pneumonia*, **or** *Important Messages for Health Workers***?**

Diarrhea

Diarrhea: This topic has 7 sections, when you hear the one you want, just say it.
1 What is Diarrhea **[pause]**
2 Causes of Diarrhea **[pause]**
…
*7 The Third Principle of Treatment***.**

What is Diarrhea [could have said “One”]

What is Diarrhea? Diarrhea is a dangerous disease that can… [continues to the end of the section]. To hear this again, say *repeat***; to choose a different section, say** *different section***, or for a different topic, say** *different topic***. If you’re done, say** *goodbye***.**

[Hangs up]

In addition to the above dialog options, there were a limited set of globally active commands and behaviors. First, the user could say “what can I say?” at any time, which the system would respond to by giving a list of currently valid options, as well as tersely explains the current status of the dialog.

At any step in the dialog, if the recognizer had low confidence on what was spoken, the system would ask for a confirmation, such as: **I think you said Pneumonia, am I correct?** – to which a valid response would be **yes** or **no** including a few variations (equivalent to "yeah" and "that's correct").

Finally, if sound input level was too low, the system would reply with **I'm sorry, I didn't hear you**, as well as an escalating silence prompt (i.e., on two successive silences) of **I'm sorry, I still didn't hear you, please speak louder**. Similarly, on instances where the recognizer generates no valid hypothesis, the system would say **I'm sorry, I didn't understand you, please say that again**, and on an escalating non-recognition, **I'm sorry, I still didn't understand you, please say that again clearly**.

4.1.5. Study Design

The goal of the pilot study was to test the system with a handful of low-literate health workers to understand the issues that arise in the use of a relatively simple menu-based speech interface. CHWs that were part of the Pakistan Initiative for Mothers and Newborns (PAIMAN)[10] were chosen for this study. These workers were located in rural areas, and had a lower literacy level than the workers interviewed in and near Karachi as part of the needs assessment. They had been trained on 7 themes, each of which had a number of messages they need to deliver to their communities. Of all the health workers we had the option of working with, these had the least amount of education, and hence were ideal for an initial pilot test.

The study design was as follows: each participant was given an introduction to the study, with an emphasis on the fact that it was the *system* being tested, not the CHWs themselves. Demographic information was taken, and then the participant was asked to read one line from their training manual, as a quick literacy evaluation. CHWs that were able to read were then asked to answer a simple health question using the text pamphlets. Participants were then given a brief verbal explanation of the working of the system, and then asked to find the answer to a health question using the system. Finally, they were asked to rate the system along 6 dimensions on a 5-point Likert scale, based on the SASSI set (Hone & Graham, 2000).

4.1.6. Results

The actual study took place at the Basic Health Unit in WahiPandi, in Dadu District, Sindh. Due to a miscommunication with local coordinators, there was no telephone present with which to dial the system, and so we used a USB audio device (that closely resembled a phone) plugged into a laptop running MSS as the interface mechanism.

9 CHWs participated in the pilot. 3 of them had not received any formal schooling, and hence had minimal fluency in Urdu. These 3 were not able to continue with the study.

[10] More information is available at: http://www.paiman.org.pk

Of the remaining 6 participants, 5 were able to read reasonably well or effectively, while 1 was not able to. When given a question to answer from the pamphlet, these literate 5 were successful in finding the answer within 60 seconds. One participant was a skilled reader, and she skimmed the text, browsed the section headings, and skipped to the appropriate section to find the answer. The remaining 4 participants started by pointing their index finger to the first word on the first page, and then read sequentially until they found the answer in the relevant section (which was in the second section out of seven sections). Thus, in a task with the section placed at the end, these participants would undoubtedly have taken significantly longer.

When these 6 participants used the speech interface, 5 of them were able to successfully hear and report the information they heard. The 1 participant who did not succeed failed mostly due to an echo cancellation problem resulting from a defect in the USB audio device: the microphone picked up the audio emanating from the earpiece speaker, which led to the system ending up in a never-ending non-recognition loop ("I'm sorry, I didn't understand—I'm sorry, I didn't under—I'm sorry…"). Most significantly, the low-literate participant who was not able to read pamphlets at all *was* able to use the speech interface successfully.

Finally, the questions asked after the interaction were as follows: How useful did you find the system? How difficult was it to use? How annoyed were you while using it? How many errors did you feel it made? How well did you know what to say to the system? Was the interaction too long? The mean responses are shown below.

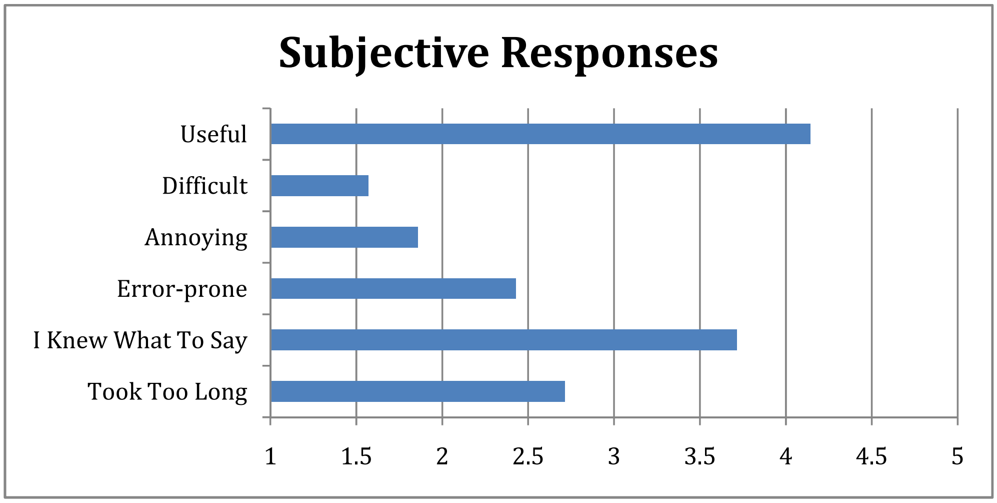


Fig. 4. Subjective responses to Likert-scale questions.

4.1.7. Discussion

In a pilot setting, our inability to hear the conversation was very detrimental, as we couldn't tell until post-hoc analysis whether a conversation was failing because of user confusion, or because the system was hearing extraneous noise (including its own voice). In a more rigorous experiment, this would not be a concern, but for a pilot, suspending an interaction, explaining the issue, and then restarting the dialog could be useful.

Environmental noise was a frequently-occurring problem: the health center had tiled flooring with little furniture, and the group of health workers waiting outside the room to begin the user study were talking to one another, with echoes that reverberated throughout the health center. Many of the rooms were connected without doors, and the windows did not have any covering beyond a wire mesh, so it was not possible to block any of this sound. It was also not possible to ask the health workers to speak softly, as it would create further apprehension on their part for the user study, and it was not possible to ask them to move to a different location either. One method we tried showed promise: by lowering the input volume on the microphone, the system would only hear the loudest of sounds. Many times, this would mean that it would not even hear the user speak, but after triggering the silence prompt even once (**I'm sorry, I couldn't hear that**), CHWs were quickly entrained to speaking louder so as to avoid hearing the prompt again.

We also realized that having a short, implicit confirmation after each step (the repetition of the chosen item to inform the user what was recognized, as in the underlined text in the following dialog: **What topic do you want? Pneuomonia, Diarrhea, or—**[user: **Pneumonia**!]**. Pneumonia*:* this topic has 7 sections**) was counter-productive, as users almost always said an enthusiastic "yes!" when they heard this, which the system did not expect, which led to a series of misunderstandings. In fact, this "feature" was put into the dialog with the assumption that our target audience was somewhat docile and not outspoken, and so might not respond to a prompt unless explicitly asked to do so. Since then we have realized the need to reverse our assumptions in this regard.

Instead, it appeared that **changing the interface metaphor** used and **making the response more concrete**, if slightly lengthier, could be beneficial. Thus instead of saying "**Pneumonia, this topic has 7 sections**", the system could say "**I've opened up the book on Pneumonia, and I see 7 sections**". The concept of a topic having a number of sections is very abstract, and participants had difficulty grasping that concept. Concretizing this aspect of the system may make it easier for users to create a mental model of the interaction, which would improve their understanding the of the prompts.

Also, the audio prompt recordings for topic names (e.g. **Diarrhea, Pneumonia**) were re-used both for questions and for statements, and these were recorded with an ascending prosody. When used in the question (for instance, **Do you want information on Diarrhea, Pneuomonia, or…**), these worked well. However when used as an implicit confirmation, most users believed that the question was being repeated, since the prosody was identical. This shows that re-recording prompts with appropriate prosody for different contexts is important, especially with low-literate users.

We also realized the need for **explaining and demonstrating the system** for a new user. While a verbal explanation of the system's usage was adequate, having new users hear and see an actual conversation would make it much clearer, and might also make it easier to remember global commands such as "what can I say?", and so video-based demonstration may be an effective means of achieving this.

While the initial steps in the dialog were conversational, the final one consists of a one-way lecture, with the system's voice reading the text out for the entire section, which seemed sub-optimal. Three users employed the repeat command to hear the text again. One user tried to echo the system's reading out of a bulleted list, as a mechanism to help her remember it more effectively (presumably as compensation for inability to write), although this resulted in her interrupting the system's speech output twice. There is clearly a need for improvement in this part of the dialog, as **long, non-interactive prompts were very difficult for the participants to follow**.

4.1.8. Conclusion & Lessons

The prototype was the first step towards creating a larger system containing more information, and more functionality, targeted towards more categories of CHWs. Similarly, the pilot was the first step towards testing the system with more categories of health workers, and with different kinds of health information. One major issue became apparent through the above experience:

- **The high correlation between education and language exposure in the developed world.** While our partners had agreed that Urdu would have been an acceptable choice, only those health workers that had been educated (even minimally) were the ones who were able to speak in Urdu. Non-literate health workers were unable to understand Urdu, and were not able to access information from an Urdu-based speech interface.

Additionally, the following issues with the user study design became apparent:

- **The difficulty of getting accurate responses in subjective evaluations with low-literate users**. The system was sub-optimal on many counts, most notably due to the USB audio device's feedback problem. However, the subjective evaluations were mostly positive – which is strikingly different than similar research with participants in the developed world (Sherwani et al., 2007). This suggests a much stronger 'politeness' factor in these participants' responses than in user studies in the West. Additionally, this user study only provided participants with one interface to judge, and it is known that participants provide more critical feedback (at least in the West) when given multiple interfaces.
- **The difficulty of getting any feedback during the user study**. The audio interface problem meant that at some points, participants were hearing a voice interrupt itself time and time again – a behavior that would be perceived as odd by anyone accustomed to human conversation. However, even when faced with this output from the system, not a single participant said anything was wrong, or that the system wasn't making sense – instead, they just calmly listened until the experimenter intervened. This can be explained both by the novelty of the system and of the user study process, and also by the low self-efficacy of users in the developed world, who tend to blame themselves rather than the system for any errors. Since the user study was carried out by researchers that were not from the same community as the participants, this may have caused further hesitation on their part.

- **The difficulty of explaining a system verbally.** In a user study, it is important to enable the participant to learn the interface to a sufficient skill level so that the usability testing of the interface reflects a regular user's usage rather than a novice user. In the West, this is often done by:
    - Providing written material. With low literate users, this would bias the study towards participants with better reading skills.
    - Describing the system declaratively. However, it appears that low literate users are not used to receiving information in declarative forms; instead preferring to learn through experience.
    - Leveraging participants' knowledge of other automated systems. Unfortunately, women from low income rural communities in the developing world have rarely ever interacted with an automated system.

  In this study, we used verbal explanation, and found that it was not optimal. With written material, participants could listen to the system while reading the tutorial; however, by verbal description it was not possible to interject into their conversation with the system to explain what to do. Also, no matter how detailed and complete a description was given to the participants before using the system, it would not be possible to explain a completely novel system just through verbal description.
- **The need to concretize the interaction metaphor.** Abstract concepts such as topics and sections were very difficult for the participants to follow. It was clear that an improved interaction metaphor was needed to concretize the interaction for the participants.

4.1.9. Next Steps

Based on the findings from the previous study, a number of methodological changes were needed. First, the correlation between education and language exposure meant that to continue using the Urdu infrastructure, we would need to shift to users that were more comfortable speaking in Urdu. The lack of Sindhi speakers in the research team made switching to Sindhi unfeasible. While the health workers in Wahi Pandi were less likely to speak Urdu, health workers near the metropolitan city of Karachi were more likely to speak the language, and so the next study, was aimed at community health workers in Memon Goth, a small town on the outskirts of Karachi.

Furthermore, instead of testing a complete end-to-end system, the next study focused on testing individual components separately, specifically the differences in comprehension between speech and text modalities. Thus, the technical component under examination was the system's output on its own.

Finally, the content was chosen to focus on information that participants had already been trained on, based on the recommendation from the literature and from the NGO partner, with the reasoning that since the deployed system would need to reinforce existing training (rather than provide new information), the content used in the user study should reflect this domain.

4.2. A Comparison of Urdu Text & Speech Comprehension in an Urban Context

The focus of the second study was on the comparative comprehensibility of speech and text by low literate users. As a contrast to the previous study, where an entire end-to-end system was tested, this study focused on only the speech output component in isolation, to provide feedback on how this component could be optimized in an end-to-end system.

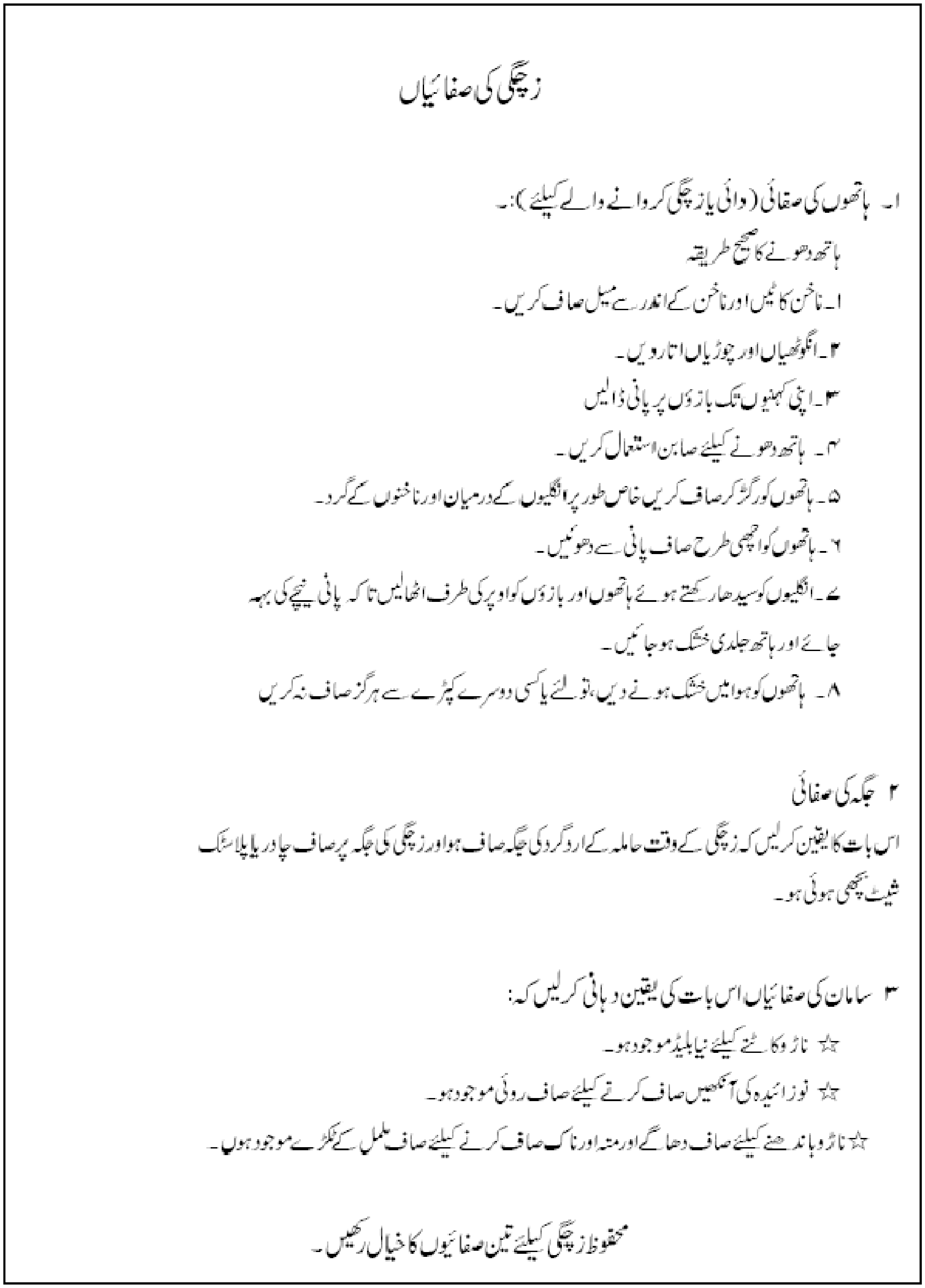
Two interfaces were tested. The first was the verbatim text from the 10 chosen pages. The second was a speech interface created from the 10 pages. The content of the system was the text from the 10 pages adapted slightly, expanding tabular data into a conversational style. Using the Urdu recording environment described in the previous chapter, the same voice talent was recruited to read aloud the text from these adapted 10 pages. The length of the audio ranged between a maximum of 137s and a minimum of 46s (M = 96.2s, SD = 33.8s).

4.2.1. Prototype Design

Since we intended to test the understandability of the system's speech output in isolation from speech recognition issues, we built a prototype optimized for this task. The system was effectively a phone-based audio player, using only touch-tone inputs. The text from each of the ten pages was recorded as separate audio files on the speech server. On dialing the number, any one of the numeric keys (0 through 9) could be used to start the playback of one of the 10 pages of content. During playback, touch-tone keys allowed pause and resume, as well as skipping forward or back by 10 seconds.

4.2.2. User Study Design

For the actual tasks, participants were asked to answer a set of ten questions, one per page of the manual. Participants were asked to use the printed text to answer five questions, and the speech interface for the other five. For text use, the experimenter would locate the page appropriate to the question asked, and hand it to the participant, while in the speech interaction, the experimenter would key in the appropriate page number and then hand the phone to the user study participant, and after a 3 second pause, the system would start playing back the audio for that page. The order of presentation of speech and text was counterbalanced, and the ordering of the questions was also counterbalanced. One page from the manual is shown below as an example:



زچگی کی صفائیاں

۱۔ ہاتھوں کی صقائی (دائی یا زچگی کروانے والے کیلئے):۔

ہاتھ دھونے کا صحیح طریقہ

۱۔ ناخن کاٹیں اور ناخن کے اندر سے میل صاف کریں۔

۲۔ انگوٹھیاں اور چوڑیاں اتار دیں۔

۳۔ اپنی کہنیوں تک بازوں پر پانی ڈالیں

۴۔ ہاتھ دھونے کیلئے صابن استعمال کریں۔

۵۔ ہاتھوں کو رگڑ کر صاف کریں خاص طور پر انگلیوں کے درمیان اور ناخنوں کے گرد۔

۶۔ ہاتھوں کو اچھی طرح صاف پانی سے دھوئیں۔

۷۔ انگلیوں کو سیدھا رکھتے ہوئے ہاتھوں اور بازوں کو اوپر کی طرف اٹھالیں تاکہ پانی نیچے کی بہہ جائے اور ہاتھ جلدی خشک ہو جائیں۔

۸۔ ہاتھوں کو ہوا میں خشک ہونے دیں، تولئے یا کسی دوسرے کپڑے سے ہرگز صاف نہ کریں

۲ جگہ کی صفائی

اس بات کا یقین کرلیں کہ زچگی کے وقت حاملہ کے اردگرد کی جگہ صاف ہو اور زچگی کی جگہ پر صاف چادر یا پلاسٹک شیٹ بچھی ہوئی ہو۔

۳ سامان کی صفائیاں اس بات کی یقین دہانی کرلیں کہ:

☆ ناڑو کاٹنے کیلئے نیا بلیڈ موجود ہو۔

☆ نوزائیدہ کی آنکھیں صاف کرتے کیلئے صاف روئی موجود ہو۔

☆ ناڑو باندھنے کیلئے صاف دھاگے اور منہ اور ناک صاف کرنے کیلئے صاف ململ کے ٹکڑے موجود ہوں۔

محفوظ زچگی کیلئے تین صفائیوں کا خیال رکھیں۔

Figure 24: One of the ten pages used in the study. This page focuses on sanitation preparation before delivering a child.

The audio content consisted of recordings of an Urdu speaker reading aloud the ten pages. Minor adjustments to the text were made to accommodate the oral channel (e.g., tabular information was converted to conversational speech).

Participants were given an introduction to the research, after which demographic information was recorded. Each participant was asked to read a standard paragraph in Urdu, and the time taken to read it was recorded. They were then given an introduction to a system (either speech or text) through a pre-recorded video, along with a practice task. They were then given five information access tasks one at a time, and were asked to use the given system to find the answer. After the set of five tasks, they were then shown the pre-recorded introduction for the other system, a practice task for that system, and five more information access tasks to perform with that system. At the end of this, they were asked to subjectively rate the systems on various dimensions.

4.2.3. Participants and Information Content

A group of 13 government health workers (LHWs) were identified in Memon Goth, a town in the outskirts of Karachi. They had been trained in Urdu, and spoke the language fluently. Their education level ranged between 8th and 12th grade. While all of them had received education in Sindhi, only five had received education in Urdu.

The content chosen for this study was a 10 page subset from the 20 page refresher manual for LHWs, which in turn summarized the important messages from the full 350 page manual.

Of the 13 participants, a pilot study was done with three, while the final user study was done with the remaining ten.

4.2.4. User Study Tasks

In consultation with the NGO partner, it was decided that the user study tasks should be representative of the information needs of health workers, and so these tasks were designed by a public health expert at the NGO. These tasks were designed to ensure that health workers understood crucial elements of their training. Nine of the ten questions asked for more than one piece of information, with some questions asking for as many as seven items (e.g. “name the 7 danger signs in a newborn infant”). The average number of items asked per question was 3.5.

4.2.5. Pilot Study

In a pilot study involving three of the participants, it was apparent that participants were overwhelmed with the use of different buttons for various features (for pause, resume, skipping forwards and backwards by 10 seconds). Thus, in the final study, this was simplified by not revealing the skip functionality, and only teaching the pause/resume features.

Additionally, it was apparent that Likert-based evaluation was problematic with the participants. At many points, participants gave conflicting responses to the questions asked, and also appeared unsure of what the questions were asking. Based on this, extra care was given to how these questions were asked in the final study.

4.2.6. Results

In terms of task success, an identical number of questions were answered correctly in both the speech and text conditions: 35 correct answers out of a total of 50 questions in each condition. Thus, there was no difference in task success between conditions.

Task completion time was significantly different between the two conditions, with speech (*M* = 121s, SD = 29.5s) taking longer than text (*M* = 50.7s, SD = 99.2s), with an analysis of variance (ANOVA) showing F(1, 98) = 23.50, p < 0.001.

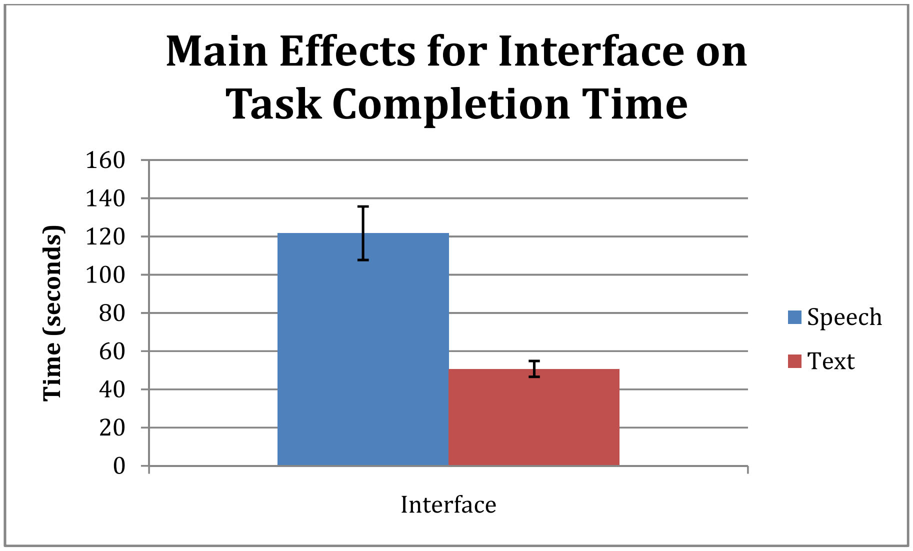


Figure 25: There was a significant difference in task completion time between the speech and text conditions.

Surprisingly, even with the stark difference in task completion time, none of the subjective response dimensions (frustration, ease-of-use, cognitive load, perceived usefulness) had any statistically significant differences. The average responses for both speech and text are shown below:

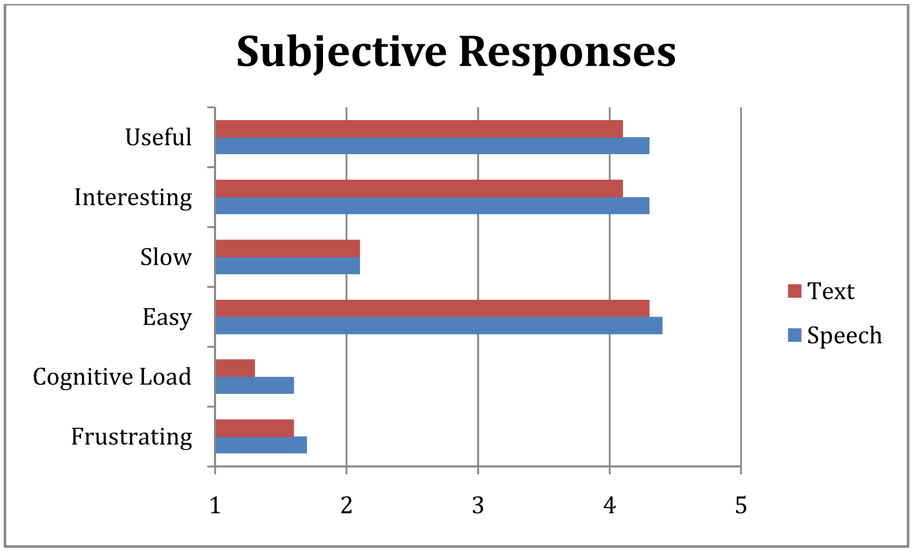


Figure 26: Subjective Responses for Speech and Text conditions. Participants were asked to rate the system on a 5-point scale.

Moreover, when asked to rate the systems against one another after completing the entire user study on a 5-point scale, participants rated them almost exactly identically on all three dimensions. A rating of 5 meant a strong vote for the speech interface, while a rating of 1 meant a strong vote for text, while a median rating of 3 meant the interfaces were the same on that dimension. The average ratings were 3.1 for ease, 3.0 for speed, and 2.9 for preference, showing that these interfaces were regarded as very comparable to one another.

4.2.7. Discussion

The findings in this study were surprising for a number of reasons. First, research on users in the West (Sherwani et al., 2007) suggests that such users are extremely sensitive to the time spent in using different interfaces to access information, when the time difference is significant. This was not the case in this study, since the speech tasks took significantly longer than the text tasks, and yet the subjective responses were comparable, which suggests a potentially fundamental difference in the sensitivity to time difference by low-literate users in the developing world.

The task completion time was consistently higher for every participant (with one exception, described later) in the speech condition than in the text condition. The main reason for the increased task duration in the speech tasks was due to the long audio clips: their average length was 96.2 seconds, which imposed a large cognitive load on participants. Literate users in the West recall only 56% of the information presented when the audio length is 15 seconds long (Langner et al., 2006), and this drops to 13% as more information is presented. While the task completion rate in our study was much higher (70%), it must be noted that participants were allowed to hear the content as often as they wished to answer the question. However, it is clear from our study that long audio segments are intrinsically difficult for any person to parse. Additionally, various other factors compounded to make the long duration even more problematic:

- Participants almost always waited till the end of the audio clip to give a response to the question, even though the training video explained that they could stop listening and give the answer the moment they found it, or press a button to pause the system's voice. Perhaps unsurprisingly, participants stuck to the principles of human-human dialog, and did not put down the phone (or press a button) while the system was talking, nor try to speak over the system's voice.
- Because they waited till the end of the system's utterance to respond, participants often forgot the answer (and even the question), and had to listen to the audio from the beginning, leading to a sharp increase in task completion time.
- 9 of the 10 questions asked for more than one piece of information. Thus, even if participants gave the answer while the system was speaking, they would often miss the remainder of the answer that was played while they were speaking
- The use of the "pause" button that was described in the training video would have solved this, but the pause button was only used by 2 of the 10 participants
- In the above instances, whenever they had any difficulty, the entire clip would be played from the beginning

Another striking feature of this study was that the time taken in the speech condition was always higher for every single participant, except the most literate one: P5 had studied till the 12th grade, and took 51 seconds to find the correct answer via text and 41 seconds via speech. Anecdotally, this counter-intuitively suggests that effectively comprehending text material that is read aloud may require more literacy than when simply reading the text itself. This point is explored further in Appendix B (Section 8).

One significant qualitative finding was the difficulty in administering the subjective evaluation questions. Likert scales are meant to be administered textually and visually, however, given the variable literacy skills of our participants, this was not viable. It was apparent that the subjective evaluation questions were either not understood, or that it was unclear to the participants what the difference between the various dimensions were. This is in stark contrast to user studies in the West, where Likert scale-based questions are routinely administered textually to literate participants.

Another problem arose in the context of the choice of health information content. We had chosen to work with reinforcing the material that the health workers were trained on (maternal and reproductive health), which is what an eventual deployed system would need to provide. 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).

Thus, in subsequent studies, we shifted to working with content that the participants had *not* been trained on before, without any taboo elements in the content.

Finally, given the fact that the speech interface tested in this study was trivial to design – it was playing audio files that read aloud the given text – the fact that task success between both conditions was identical strongly suggested that the role of speech interfaces for such users could potentially be far more compelling and viable than for literate users in the West. Thus, a more optimized interface could potentially make speech even more compelling, as measured both the objective and subjective metrics.

4.2.8. Conclusion & Lessons

The following conclusions arose from this study:

- **A baseline speech interface is comparable to a text interface, for users in the developing world with basic literacy skills**. This is surprising, since literate users in developed countries prefer text interfaces in a comparable context. Additionally, the study suggests that an improved speech interface could outperform the text interface without significant effort. However, given the nature of the participants in the study, it was unclear how the same two interfaces would compare when tested with participants lower down on the literacy continuum.
- **The length of the system's utterances should be kept as short as possible.** It was difficult for participants to remember the question, find the answer, and then remember the answer for the duration of the audio segment, as participants rarely paused playback of their own volition.
- **Literacy is more than just the inability to read and write**. The fact that it was only the most educated health worker that was able to access information from the speech interface faster than the text interface, as well as the difficulty for the participants in understanding and responding to the subjective evaluation questions, suggests there were some other factors that were strongly correlated with literacy, or even the sociocultural context of our study. We return to these issues later in this thesis.

4.2.9. Next Steps

From this user study, it was also clear that changes to the user study design were needed based on the following findings:

- **The choice of content (health information) is crucial**. Specifically, it was sub-optimal to use material that health workers had already been trained in, and that novel, unfamiliar material should be used in such user studies – however, it should still be relevant, so that participants in the user study do not view the task irrelevant and hence trivial.
- **Each user study task should ask for only a single piece of information**. While content experts may prefer asking longer questions, such questions are difficult to work with in the context of a user study.

These changes were incorporated into the subsequent user studies.

Furthermore, given that most of the participants in our study had the ability to read – a reflection of the fact that we conducted the study in the outskirts of a large metropolitan city – it was clear that further research was required to understand the comprehensibility of speech and text for lower-literate users. Thus, in the next study, we aimed to use the same user study design with a group of health workers of lower literacy. Our partner NGO, HANDS, had just started a new community health project, which focused on the training and monitoring of around 400 community health workers, which would provide basic health services as well as a range of other services including microfinance. By the time we began our study, many health workers had been identified from the Umarkot district (a rural area) as part of HANDS' baseline survey, but had not yet been trained. These workers were the target population for the next user study, whose aim was to explore differences in comprehension with participants with lower literacy levels.

4.3. A Comparison of Speech & Text Comprehension in a Rural Context

4.3.1. User Study Design

The content in this study was drawn from the same source as in the previous study. However, considerable portions of the text (and corresponding speech) were removed to make the tasks easier and shorter, given the lower-literate nature of the intended participants. Also, while the previous study has some questions that asked for multiple answers (e.g. "name 3 danger signs in pregnancy"), in this study each question asked for only one piece of information (e.g., "name any one danger sign in pregnancy"). The remainder of the study design was identical to the previous one.

The study was conducted at HANDS' regional office in Umarkot. Participants were women from nearby communities who had been identified for subsequent health worker training. HANDS had suggested that even though many participants may not have been fluent Urdu, they would likely have sufficient exposure to it to participate in the study.

4.3.2. Results

Of the 10 participants, 5 were not able to understand written Urdu at all. However, these 5 were still able to use the speech interface to varying degrees of success. Overall, in terms of the task success rate, there was a statistically significant difference between the speech (M = 76%, SE = 6%) and text (M = 46%, SE = 7%) conditions, F(1, 98) = 7.66, p < 0.01.

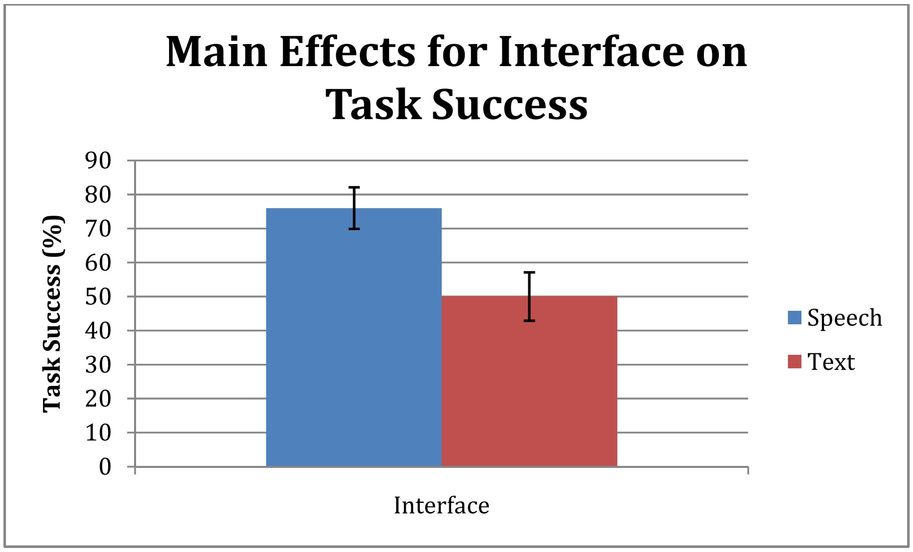


Figure 27: Difference in Task Success between Speech and Text interfaces. The difference was statistically significant.

Since the user study facilitator only spoke English and Urdu, and the participants were not very comfortable with either of these languages, no meaningful subjective evaluation data were collected.

4.3.3. Discussion

Through this study, it became clear that when targeting low-literate users in a developing country, the **national language is often *not* the optimal choice for a speech interface**. The lack of a strong education system meant that people were not uniformly able to learn fluency in the national language, and instead were very likely to be fluent only in the language of their community. In our case, this meant that some participants only spoke Thari, while others spoke a local dialect of Sindhi.

The participants were not yet health workers – they had only been identified to be trained as health workers in the future. Thus, they were not at all familiar with the health content that was part of the system. This avoided the issue from the previous study where participants felt embarrassed when they didn't know an answer they felt they should have known. However, participants still offered answers before looking to the interface to find it. This suggests that the very premise of a user study may not be well suited to low-literate populations in the developing world – since "looking up" answers to questions is not a natural process in these cultures.

4.3.4. Conclusion

This study gave a strong indication that **speech can be a winning interface for low-literate users, since there was a statistically significant difference of 26% in task success between the text and speech interfaces**. However, the confounding factor of language choice meant that it was the least literate participants that were not fluent in the interfaces' language.

Furthermore, **the choice of language is crucial**. National languages may not be the best choice even if claimed to be preferred by the users or their supervisors. In the context of this research, it was decided to use Sindhi as the language for subsequent user studies.

4.4. Comparing Sindhi Speech & Text Comprehensibility in a Rural Context

4.4.1. Participants & Information Content

In consultation with HANDS, health workers at Wahi Pandi (the site of the first user study) were chosen for the next user study. Further, 6 pamphlets were identified that had been previously translated into Sindhi, and were not part of the Wahi Pandi health workers' training. Each pamphlet focused on a specific health issue or disease, specifically: diarrhea, malaria, pneumonia, hepatitis, child nutrition, and child immunization. A female native Sindhi speaker from HANDS' field office from the outskirts of Karachi was recorded reading the text from each pamphlet aloud, which formed the basis of the speech content for this study.

4.4.2. User Study Design

To ensure that the study design was appropriate for the context of the intended users, two HANDS field co-ordinators from the community were enlisted as study co-designers. They were first given a tutorial covering the basics of user study design, and the rationale behind the methods employed in such research. After this, they gave their input on a number of tasks. They translated the user study introduction used in earlier studies from Urdu into Sindhi, and also adapted it to include reference to the specific context of how the study fit into the overall goal of the health workers to provide health services to their community members. Further, the user study design had been kept partially incomplete on purpose, so that the co-ordinators could complete it collaboratively with the research team, so that it was not only made more grounded in the local context, but also so that the co-ordinators felt ownership of the user study process. This proved beneficial in the course of the user study as the facilitators were invested in the study and in its outcomes.

The study design was similar to the previous two studies. The sole independent variable was the interface, which was either text or speech. Each task consisted of a question for which the answer was to be found from the given interface: the text pamphlet or the spoken recording of that text (played back over the phone).

Participants were each given a verbal introduction to the study, after which their demographic information was recorded. For those able to read Sindhi, they were asked to read a standard paragraph aloud (which was timed). They were then given an introduction to one of the two interfaces (order was counterbalanced), along with a practice task. They were then given 3 timed tasks one after the other, and were asked to find the answer using the interface. They were then switched to the other interface, with an introduction and practice task, and 3 more timed tasks. Finally, they were asked to rate each system on various subjective metrics.

4.4.3. Results

Of the 11 participants, 3 were not able to speak Sindhi. All of the participants were ethnically Baloch – migrants from the neighboring province of Balochistan – and were native speakers of a minority dialect of Balochi. The data from these 3 participants was removed in all analysis. The remaining 8 however were able to speak Sindhi to varying degrees of fluency.

The speech condition had a lower task success rate (13/24 correct answers) than the text condition (16/24 correct answers), although the difference was not statistically significant. Task completion time did vary, and the difference was statistically significant, with the time to complete tasks with the speech interface (M = 287s, SD = 115s) higher than with the text interface (M = 177s, SD = 138s), F(1,46) = 8.96, p < 0.01.

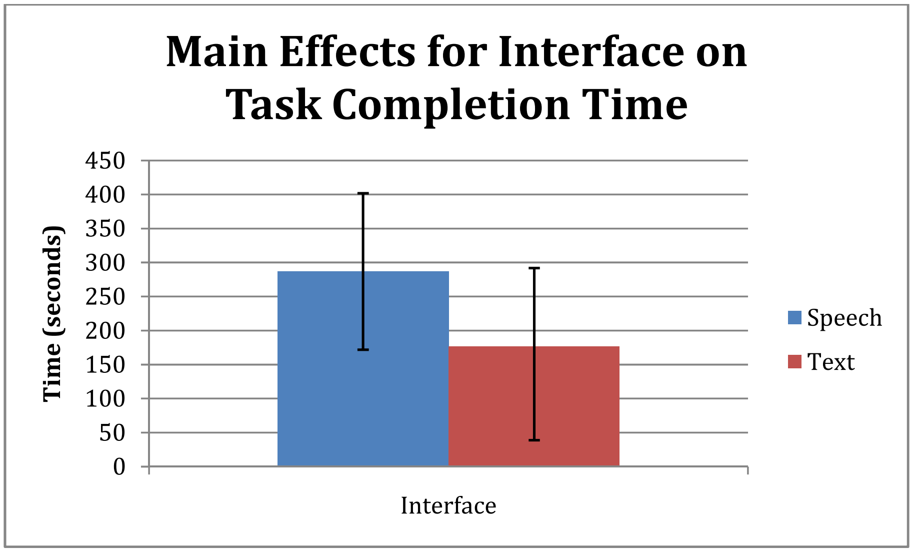


Figure 28: Task completion time was significantly different between the speech and text conditions.

None of the subjective metrics showed any significant difference between the two conditions. However, when asked to choose which system was better, 3 preferred speech, and 5 felt they were both the same, but no participant said that text was better.

For the three participants who were not able to speak Sindhi (and hence were not able to complete any of the tasks successfully), in an open-ended discussion after the study, each participant was asked if they would prefer Balochi as the language of the system. Surprisingly, each of them replied that the system was fine the way it was in Sindhi, and they didn't want it in Balochi. After deep probing, each of these participants said they would prefer Balochi, and each had a different reason for not saying so earlier. One participant felt that we were referring to the official Balochi, which is broadcast on radio, but is unintelligible to them. Another participant felt that if the system was made in Sindhi, she would be blamed by her peers as the reason for this change. Yet another participant said she would prefer a Balochi system only after being told that all the answers she gave were incorrect.

4.4.4. Discussion

Unlike the previous studies, where the information content had an average of 96 seconds, in this study the content was considerably longer (M = 244s, SD = 84s). This increased length of the audio content made it even more difficult for participants to access information effectively. This shows that non-interactive audio content (especially content designed for literate consumption) is not an adequate form of information provision for low-literate users. However, given that three of the participants could not read at all, speech was the only possible interface choice for them, provided the information was in their language. Finally, it is striking that no participant preferred text, even though the speech system was considerably difficult.

The system's language was changed to Sindhi for this experiment with the expectation that it would be more accessible to participants – however, it was learnt that participants were non-native speakers of Sindhi, and their native language (a dialect of Balochi) had no written form, and our partner NGO had no material designed for them.

The responses by the Balochi-speaking participants were striking in how much probing was required before the participants spoke their mind. This contrasts significantly with similar experiences in speaking with user study participants in the US, where participants are happy to express their feelings when asked.

4.4.5. Conclusion

This study showed that **text content designed for literate consumption cannot be used as-is for low literate users**. Instead, **adaptation is essential** to make the content more compelling for the medium, and for the target user population.

Further, the prevalence of minority languages in developing countries is very different than in developed countries. The existence of uniform curricula across countries with strong school systems means that the national language is often the most widely spoken language, and hence an ideal choice as the language for a spoken dialog system. However, in developing countries where there is no strong school system, it is much more likely that numerous local dialects have evolved independently, meaning that **there is no obvious single language for a spoken dialog system that can be expected to be optimal for a large low-literate population**.

The difficulty in eliciting subjective responses from the Balochi-speaking participants is striking, particularly since it is the one situation for which a "gold-standard" existed – the data clearly showed that they were not able to use the system, yet they said that the system was fine the way it was. This suggests that **subjective responses may not be as reliable** as in the West, and that **participant feedback needs to be triangulated** as much as possible.

It is still important to note the preference for speech over text, even when the speech interface is notably difficult to use. Contrasting this finding with literate users in the US again suggests a fundamental difference in the way these different groups of users perceive a system's utility.

4.4.6. Summary of Findings

The less literate the user, the shorter the speech segment needs to be. 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. This may be contrary to people's expectations that low-literate people may compensate for their lack of literacy with a stronger oral memory – but the opposite is true, at least with material that was originally prepared as written text. As we will see later, this depends critically on the difference between material prepared for oral and literate audiences. In the West, research has shown that literate listeners can recall only 56% of the information presented to them orally if the audio segments are 15 seconds long, and this decreases to 13% as the audio extends beyond 30 seconds in length (Langner et al., 2006). In the context of low-literate users, it seems that the need for shorter speech segments is actually greater than with literate users.

The national language is not always optimal, and the regional language is also 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. Our user studies showed that Urdu was understood by 50% of the participants in Umarkot, and 33% of those in Dadu. Even for these participants, many had difficulty since they were usually not fluent with Urdu. Based on this experience, we shifted from Urdu to Sindhi content. 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. Thus, the ability to speak a widely spoken language is often correlated with education, and the lack of resources in local languages makes it more difficult to provide (or even test) information access systems targeted at completely non-literate users.

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 such 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. Even in the study where speech was the least optimal interface, no participant expressed a preference for text. Based on the previous point, this must be taken with a grain of salt – however, it is expected that users from an oral tradition with limited literacy would prefer an oral 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.

In terms of user study design, one finding clearly emanated from the experience in this study:

Training and working with local facilitators is essential. Over the course of the above 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.

4.4.7. Next Steps

Based on the cumulative experience in the previous studies, the next (and final) user study was designed to avoid the various pitfalls described above, mainly:

1. A full end-to-end interactive system was built, with speech recognition and touch-tone inputs for users to be able to navigate the content themselves
2. Improvements to the recognition engine were made to increase accuracy significantly
3. The content was adapted both in terms of the wording as well as organization
4. Considerable effort was put into training participants on the use of the interface

4.5. A Comparison of End-to-End Speech and Touch-tone Interfaces

Based on the previous studies, the final study was designed to test a complete end-to-end system, to compare the use of speech versus touch-tone as input modalities.

4.5.1. Content Adaptation

From the earlier user studies, it was clear that using the content without adaptation was suboptimal for a number of reasons:

- Text content can be organized spatially in tables and lists – these organizations are lost in oral recitation
- Text content can use visual markings, such as bold-face fonts, underlining, as well as special characters to denote noteworthiness (e.g. by using an asterisk)
- Text content does not need to be conversational
- Text content can be very long, as it relies on the reader to set the pace

Speech content, especially speech designed for consumption by low-literate listeners, cannot use these features of text. For instance, the previous user studies showed clearly that listening to long passages was not an easy task for the participants, with smaller audio portions much easier to understand than longer ones. Also, given the lack of exposure of participants to formal texts, and their comfort with conversational speech, the text needed to be adapted for conversational presentation. Finally, the use of spatial and visual cues needed to be replaced by sentences that could express the same information.

Thus, the first design decision in this study was to adapt the health information content for spoken output, instead of recording the content verbatim from start to finish. For this purpose, the content was split up into smaller “nuggets” of information. Specifically, each pamphlet was reworked into a tree. Each node in the tree would have a title, a short nugget of content, and any number of descendants. The root node of each tree would have a short introduction describing the specific health issue (e.g. diarrhea), and would spell out the major sections of the information, along with the higher level sections (e.g. symptoms, methods of prevention, treatment). Each of these was a separate node, with its own descendants. Each node was modeled to give more information than the previous node, both in terms of detail as well as length.

Given the tree-like structure of the information, the interaction was modeled as a conversational menu. However, instead of offering each choice abstractly, the system expressed a persona (“Dr Marvi”), who would provide the caller with the information they needed. The use of the persona made the system more natural and conversational, as speaking to a person closely matched the style of interaction the participants were used to.

The system would first ask the user to select a given topic (e.g., malaria, diarrhea, or hepatitis), after which would be given a short introduction to the topic and then asked to choose from a specific sub-topic (e.g. general information, signs, preventative measures, treatment), after which they would be given detailed content broken down into chunks of three bullet points at a time. The interaction was implemented in two different interface types: one using touch tone input for choosing between the options, and the other using speech input. Here is a sample call for both interface types, 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.

Table 2: Sample interaction for both speech and touch-tone interface types.

4.5.2. Formal User Study Design

In September 2008, a within-subjects user study was conducted testing the speech and touch-tone interface types of the menu-based system described above. 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 a pre-study pilot was conducted with 3 participants

4.5.3. Pre-study Pilot

The design for the pilot 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 interface type (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 interface type, 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 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 interface types consisted of three steps. In the first step, the participant would listen in (using earphones connected to an audio-tap[11]) 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 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 too abstract**. It is well known that low literate users have difficulty with abstract thinking (Luria, 1976). Even the task of asking a question without any context (e.g. naming any symptom of a disease) is an abstract task.

4.5.4. Changes to the Study Design

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

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

The tutorial process was increased to three practice tasks instead of two. The 5 minutes of open-ended interface exploration 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 (Chavan, 2007). 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 interface types was counterbalanced.

4.5.5. 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. The average time for each participant to complete the user study was 77 minutes.

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.

Quantitative and Qualitative Results

Task success in the speech interface was significantly higher than in the touch-tone interface. There were significant main effects 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 below.

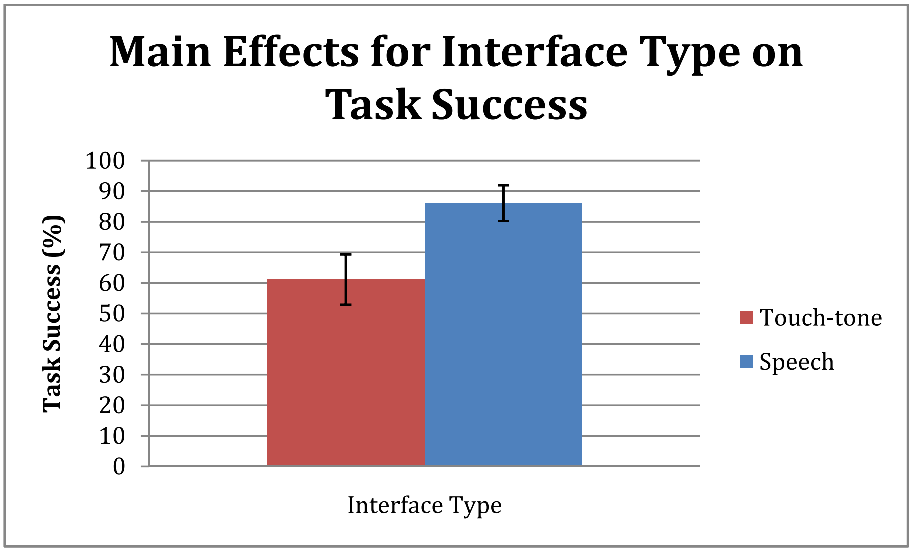


Figure 29: Main Effects for Interface Type on Task Success. There was a statistically significant improvement with the speech interface over the touch-tone interface.

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 shown below.

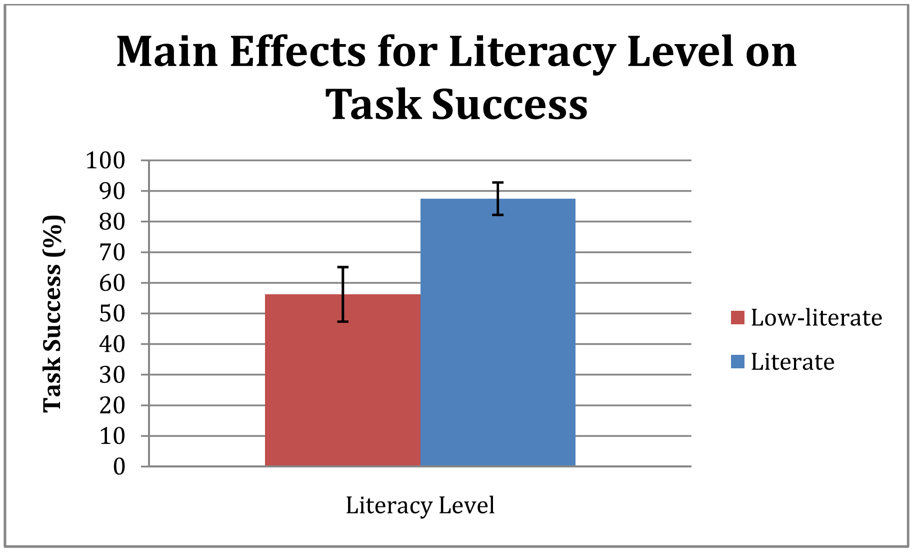


Figure 30: Main Effects for Literacy Level on Task Success. There was a statistically significant in task success between low-literate and literate participants.

Literate participants had a perfect task success rate when using the speech interface. There were no interaction effects of literacy level and interface type. 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 below.

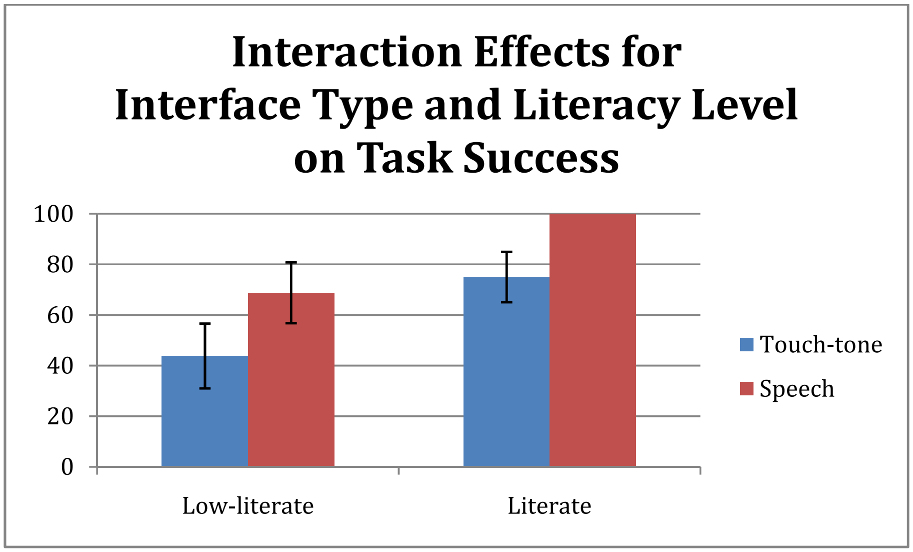


Figure 31: Interaction Plot for Task Success. Literate participants using the speech interface had 100% task success.

There was no strong consensus on which interface type was subjectively preferred. When participants were asked to rate each interface type individually, the ratings were slightly, yet consistently, in favor of speech, though the differences were not statistically significant.

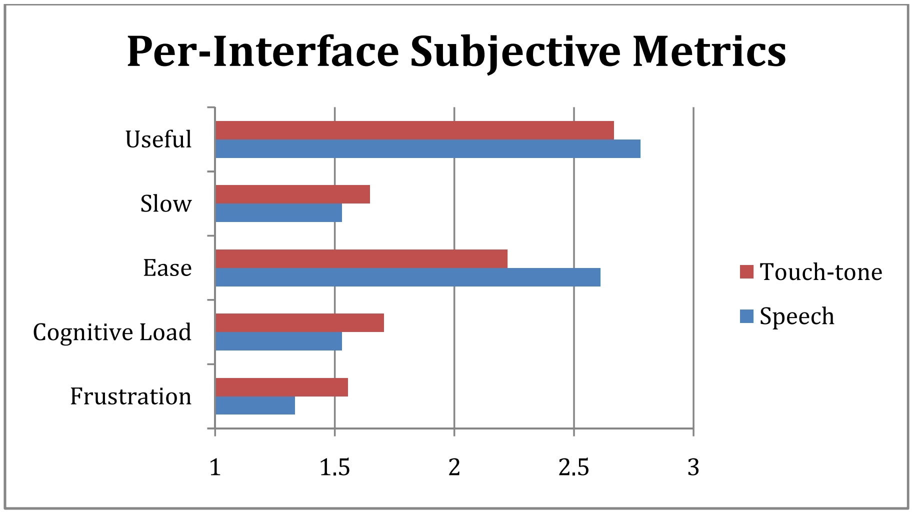
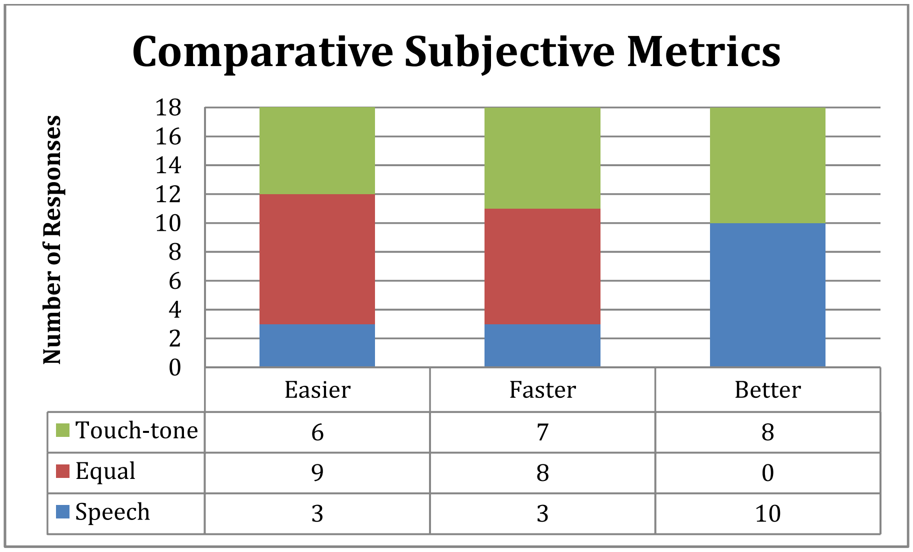


Figure 32: Subjective metrics for each interface type. A 3-point scale was used, where a rating of 1 corresponded to "a little" and 3 corresponded to "a lot". Participants favored the speech system on all counts, although the difference was not statistically significant.

However, when participants were asked three comparative questions at the end of the study, to choose which interface type they found easier, faster, and better, they often chose in favor of the touch-tone interface, as shown below.



	Easier	Faster	Better
🟩 Touch-tone	6	7	8
🟥 Equal	9	8	0
🟦 Speech	3	3	10

Figure 33: Subjective metrics across interface types. Participants were asked to rate which interface they found easier, faster, and better. "Equal" was not given as a valid choice for the final question.

For the final comparative subjective evaluation question, participants were asked which interface type they found better, and “equal” was not accepted as a valid answer. 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 had 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 interface types. 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.

4.6. Discussion

4.6.1. Significance of Objective and Subjective Metrics

Literacy is a significant factor in determining task success in both interface types. While earlier studies have suggested an effect of literacy on interface use in the developing world (Plauche et al., 2006; Brewer et al., 2006), this study clearly demonstrates that literacy is a statistically significant determinant of task success. Moreover, the speech interface had 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. For developed countries, where most research in the West focuses on literate participants, this finding strongly suggests that low-literate segments of the user population are being ignored by current research paradigms, and that **special attention needs to be paid to the needs of low-literate users, even in the developed world**.

4.6.2. Localization

Although crucial, localization is quite tricky. Localization refers to adaptation of content to local culture, dialect, terminology and language usage patterns. Even communicating about languages and dialects is non-trivial: some participants self-identified as being “Sindhi” speakers, yet were unable to understand the “Sindhi” content recorded by a “Sindhi” speaker from the city. A variety of factors can cause difficulties in such cases: the content may mistakenly contain a few words from other languages, the accent of the city speaker may be unfamiliar, and the dialects of the languages may be substantially different.

Literacy is strongly correlated with fluency in a major language. Low literate participants were less likely to be exposed to alternative dialects, or to other languages, and found 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 (see Section 8.1.4). 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.

Thus, 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.

One important question that this discussion raises is: how many languages are involved? In the case of Pakistan, a country of 176 million, the CIA World Factbook[12] lists 9 languages (Punjabi, Sindhi, Siraiki, Pashtu, Urdu, Balochi, Hindko, Brahui, English, Burushaski) plus “others”. The number of dialects within these is clearly greater based on our experience – and most importantly, these dialects are often mutually unintelligible. This could mean that information access systems would need to be localized to even 100 – 1000 dialects in the country, although further work is needed to better estimate this number.

It may appear that localization incurs a one-time, fixed cost of content generation, followed by a marginal cost of translating that content to other dialects, with the collaboration of speakers of the local dialects. However, there are at least two (if not more) issues to keep in mind. First, since it is crucial that the information be correctly translated (as in the field of health information), any translations must be vetted with the organization (governmental or NGO) that created the content, which would introduce an additional cost. More significantly, however, the task is not simply one of translating words, but also of localizing the concepts. For instance, if the treatment for a disease is given as feeding a certain food such as rice, but rice is not commonly found or eaten in a given geographic regions, it is not enough to simply translate the word for ‘rice’ into the necessary dialect – the localization would need to ensure that there are locally-appropriate alternative solutions (e.g. bread) that would be acceptable in terms of both their medical appropriateness and their availability in the socio-cultural context. Clearly, the issue of localization is one that requires significant further effort to fully explore and develop solutions for. Our work shows that it is a crucial element for any information access system.

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.

4.6.3. Literacy and User Study Design

There is a dearth of usability methodologies designed for use in developing-world contexts. 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 (Chavan, 2007)), these methods

[12] https://www.cia.gov/library/publications/the-world-factbook/geos/pk.html

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.

4.6.4. Speech Recognition Quality

While speech recognition accuracy has been a persistent problem in our previous work, based on the improvements described in Section **Error! Reference source not found.**, 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. 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.

4.6.5. The Importance of Effective Tutorials

Novel tutorial techniques are needed for low literate participants in user studies. By improving the tutorial step between the pilot and the final user study, we saw large improvements in users' ability to access information successfully. With an ineffective tutorial, both interfaces may have been harder to comprehend for all participants, and this would have skewed the results.

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.

A corollary of the need for effective tutorials is that **human-mediated interfaces are likely to be the most successful paradigm for applications for low-literate users**. This is in stark contrast to the Internet boom in the West, which has been fueled by the success of direct-to-consumer web applications. The training requirement suggests that either smaller numbers of end-users would be trained to be human mediators, unless novel approaches are worked on that significantly reduce the need for training.

4.6.6. Rapid Iterative Development

This study field tested a mobile user study infrastructure (described in more detail in Section 7.3), which enabled rapid development and modification of the speech system while in the field. This meant that the feedback of local facilitators and pilot study participants was used to make significant modifications to the the system. Additionally, it meant that speech recognition tuning could be done locally and quickly. This underscores the need for having a system development setup that enables field-based modification of the system.

4.6.7. Comparison with Similar Research

The results presented above contradict similar work in the field, most notably the study by Patel et al. (2009) testing speech and touch-tone interfaces for listening to pre-recorded radio shows and recording audio content for a radio talk show, where participants preferred touch-tone over speech interfaces. In comparing our work with theirs, a number of factors need to be considered.

First, in our system, the speech-input interface 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 (Sherwani, Tomko & Rosenfeld, 2006), 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 (Grover et al., 2009) 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. Our work shows that trained users can use speech interfaces more effectively than both text and touch-tone interfaces.

5. Speech Recognition for Resource Scarce Languages

5.1. Motivation

Speech resources are a pre-requisite to building a speech interface in a given language, and the development of such speech resources is an expensive process (Schultz & Kirchhoff, 2006). Acoustic models, language models, pronunciation dictionaries, and letter-to-sound rules need to be derived from large amounts of text and speech data, with the speech data derived from a diverse set of speakers of the target language. In languages without a written form, this becomes even more difficult. While language resources exist for most developed-world languages (e.g. English, French, German, etc.), such resources are extremely limited for the so-called 'resource-scarce' languages. It is no coincidence that most of the languages spoken natively in the developing world fall under this category.

Much research has been done on speech recognition & synthesis for resource-scarce languages. Schultz's GlobalPhone project (Schultz & Waibel, 1998; Schultz, 2002) focuses on the creation of acoustic models in the target language using small amounts of training data in the target language, but with large amounts of training data in source languages. Davel's Default&Refine approach (Davel & Barnard, 2008) provides a novel approach for bootstrapping pronunciation dictionaries for speech synthesis. Each of these approaches still requires a speech expert to perform crucial steps, such as defining the phoneme set for the target language, and all subsequent steps are done in the context of this phoneme set. Additionally, these approaches have not yet resulted in robust speech recognition systems that can be deployed and maintained by laypeople, or even technical people without a background in speech technologies. However, one of the goals of HealthLine was to transition the project to a Pakistan NGO for deployment, and so the system had to require no technical supervision in the long term, with a robust architecture that would scale to thousands of simultaneous calls if necessary. This requirement made the use of research systems unfeasible, and a commercial speech recognizer was needed. However, no commercial speech recognition systems exist for recognition in Pakistan's local languages (e.g. Urdu or Sindhi).

There is promising work on automatically determining pronunciations based on audio examples (Bansal, Nair, Singh & Raj, 2009) using a joint Viterbi decoding method. However, this requires low-level access to the recognition engine, which is not possible with commercial systems. Open source speech recognition engines do allow such access, though the quality of acoustic models that ship with such engines are not ordinarily comparable with the highly tuned acoustic models bundled with commercial speech recognition systems.

In this section, we describe a novel method we call **Speech-based Automated Learning of Accent and Articulation Mapping (Salaam)**. This approach allows a speech interface developer to leverage existing resources already developed for "developed-world" languages, such as US English, for use in recognizing "developing-world" languages.

The HealthLine system had to be designed to initially recognize Urdu, and in later user studies, Sindhi. Urdu and Sindhi were thus the target languages, while the source language used was US English. The speech recognition engine used was Microsoft's Office Communications Server 2007 Speech Server. The method presented here has only been tested with this architecture, with US English as the source, and Urdu and Sindhi as the target languages – however, there are strong indications that this technique would work on most other recognizers, with source and target languages that share a common phonetic space.

5.2. Cross-Language Phoneme Mapping

Most (if not all) speech recognition systems allow the definition of new words in a dictionary or lexicon file, which contains each word's textual representation (e.g. the word "read"), and its phonetic representation(s) (e.g. "R EH D" and "R IY D"). Our approach was to define each of the words that needed to be recognized in the target language, and hand-code pronunciation was defined using US English phonemes. For instance, the word "سلام" ("salaam") would have been described in the lexicon as سلام→S AH L AA M. Once all the relevant words were defined this way, these words could be combined in speech recognition grammars for actual speech recognition.

One major limitation of this approach is that it cannot recognize phonemes that do not exist in the language whose phoneme set is used. For example, the غ sound (a guttural "g" fricative phoneme, called GH) has no equivalent in the US English phoneme set, and so a fallback such as H might be used. Using such fallbacks still leads to correct recognitions, although the recognizer confidence is lower on these words. However, this limitation should primarily prove a problem in situations where the missing phoneme has to be differentiated from the fallback equivalent – for example, if trying to recognize the difference between words such as "GH IH Z AA" and "H IH Z AA", as these would be represented identically using the US English phoneme set. These issues can be side-stepped by carefully designing the dialog to avoid such words.

This approach was tested using the non-tutorial tasks from the final user study (described in Section 4.5). This test set consisted of 78 utterances with a vocabulary of 15 words. On this test set, the Salaam method yielded a word error rate (WER) of 16.7%. Since the grammar consisted solely of single words, the word error rate was identical to the concept error rate.

After publishing results based on this method (Sherwani et al., 2007), it was subsequently used by IBM Research India in a prototype system (Patel et al., 2009), and achieved an accuracy of 94% in their domain and vocabulary.

5.3. Using a Data-Driven Approach

In the expert-based method described above, the choice of phonemes was left solely to the discretion of a language expert who could translate the phoneme space of US English phonemes onto the phonemes of the target language words. This rule-based method can be improved upon by incorporating a data-driven method.

In this method, for each target word’s pronunciation entries, the language expert marks every phoneme for which the expert is not confident. The mark can indicate either that the phoneme may be a vowel, a consonant, or any phoneme. For instance, if the expert 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 consonant” wildcard. Similarly, “v*” denotes an “any vowel” wildcard, and “.*” denotes an “any phoneme” wildcard. Thus, if the developer wants to test the optimal phoneme choice for the final consonant-and-vowel combination in “bacha”, she may specify “B AX c* v*”.

An intermediate processing step expands these wildcard-encoded pronunciation entries into a speech recognition grammar consisting of all possible pronunciations with that wildcard. Thus, if there are a total of 20 consonants in the source language’s phoneme set, the wildcard-encoded pronunciation “M AA L U M AA c*” would be expanded into a list of 21 words, each with a unique final consonant (as well as with no final consonant). This grammar is then used to run a re-recognition pass over any sample(s) utterances of the given word, and the results from this re-recognition would show which pronunciation(s) best matched the training data. These pronunciations would then be manually chosen by the user to be used as the optimal pronunciations in the final system, or iterated upon with other wildcards to continually improve on the chosen pronunciations.

With multiple wildcards in the same pronunciation, the combinatorial explosion quickly makes it difficult for the speech recognizer to work with such a large grammar. For instance, if the expert 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 would be computationally intractable to run recognition on. In our experiments with the Microsoft Speech Server, such large grammars did not return recognition results even after 10 minutes on one word (using a computer with a 2.6GHz Intel Core 2 Duo processor with 4GB RAM). For this reason, we used a heuristic as a workaround to this problem, where the expert creates arbitrary word boundaries in the target word, to reduced 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 it would be with the full grammar, since the full grammar cannot be computed at all, the loss in accuracy is an acceptable compromise. The heuristic works significantly faster (less than a few seconds for a recognition result with MSS).

This method was tested on the same test set as described above. In addition, a training data comprising 50 utterances from 2 native speakers were used. These speakers were not in the final test set.

The earlier expert-only method gave a WER of 16.7%. Using the expert + data method (Expert-seeded Salaam), the WER was reduced to 6.4% – an absolute reduction of 10.3%.

5.4. Speech-Based Automated Learning of Accent and Articulation Mapping

While the expert-based method requires a language expert who can translate between the phoneme space of the source and target languages, the data-driven method further requires the presence of training data which can be used to optimize the hand-coded pronunciations created by the expert. In real-world settings, it is often the case that it is simple to find a few native speakers from whom a small set of training data can be recorded. However, a language expert who understands the phoneme space of both source and target languages is not always available. With this in mind, we developed another method that removes the need for an expert, yet is able to automatically discover pronunciations, given a small amount of training data.

In this method, a native speaker of the target language simply types the target words into the speech recognizer's pronunciation interface using the source language's grapheme set. In our case, this meant expressing Sindhi words in the English alphabet – a common practice for people who have used any modern computer interface (even SMS text messaging), as these are invariably in English. The speech recognizer's pronunciation interface's letter-to-sound rules will automatically create pronunciations for those typed words in the source language's phoneme set. These pronunciations would invariably perform far worse than the hand-coded pronunciations of the expert-based method. However, they provide excellent starting points on which the data-driven method may be applied.

Given the automatically generated pronunciations, the consonants in these pronunciations are kept as-is, but the vowels are replaced with vowel wildcards ("v*"), and word boundaries artificially added after every 3 wildcards. These are then expanded as in the data-driven method, and these expanded pronunciations are used to run a re-recognition pass over the training data. For each wildcard, the most frequently occurring vowels are chosen for that wildcard's location, and through this process, an improved pronunciation for each word is derived. This process can be repeated to hone in on the final choice(s) of pronunciations for each word. More formally, the algorithm may be given as:

1. Unify the pronunciations derived from the LTS rules into as compact a form as possible by converting vowels into vowel wildcards
2. Add a word boundary after every third vowel wildcard
3. Expand the wildcards to create all possible combinations of pronunciations
4. Run a re-recognition on the training set using the expanded pronunciations collectively as the recognition grammar
5. For each wildcard vowel position, choose the most frequently occurring vowel from the recognition hypotheses for each word in the training set

Removing the need for an expert to provide initial pronunciations means that the method is completely automated. We have named this automated approach Salaam: Speech-based Automated Learning of Accent and Articulation Mapping.

As an example, for the word transliterated as “maaloomaat” (information), the letter-to-sound rules generated the following pronunciations:

M AE AX L U M AX AE T
M AE AX L U M AE AE T
M AE AX L U M AA AE T
M AE AX L U M AX AX T
M AE AX L U M AE AX T

These pronunciations were used to generate the wildcard form: M v* v* L v* M v* v* T. This is shown conceptually in the figures below:

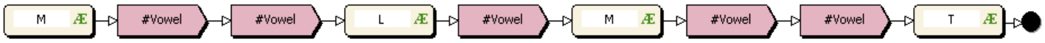


Figure 34: Conceptual representation of the phonemes of the word "maaloomaat"

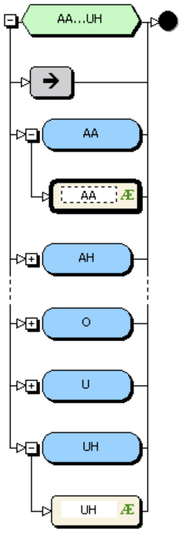


Figure 35: Expansion of the “Vowel” rule. The rule can map to any of the 18 vowel phonemes, or to a null phoneme.

In the actual grammar file, the pronunciations are expanded into words separated by the word boundaries. In this example, there is one word boundary (before the second M phoneme), and so the pronunciation consists of two words w1 and w2. Since there are 18 vowels plus 1 null vowel, w1 (with 3 vowel wildcards) consists of 6859 unique pronunciations, while w2 consists of 361 unique pronunciations. A shortened list of w1’s pronunciations is given below:

M L
M L AA
M L AE
…
M AA AA L AA
…
M UH UH L UH

Using the above configuration, the speech recognizer was used to perform recognition over training samples of the word. The most frequently occurring phonemes for each vowel wildcard position were chosen to create the final data-driven pronunciation for that word.

As an example, a one-pass application of this data-driven method yielded the following recognition results:

M AI L U M AH T
M AI L U M AA T
M AI I L U M AA T
M AI AX RA L M AO T

Thus, by choosing the most frequently-occurring vowel for each slot, the final pronunciation for the word is: M AI L U M AA T. In this case, both consecutive vowel wildcards were mapped into a single phoneme. While this method can be expanded substantially with the use of n-best lists for analyzing the optimal phonemes at any given slot, and also with the use of weights on the final pronunciation, this was not done in the current analysis due to a limitation of the API available in the commercial speech package. Applying this algorithm to each training word sample resulted in one or more pronunciations for the given word. In the cases where there was a tie between competing phonemes for a given slot, both phonemes were used to create multiple pronunciations for the same word.

This method was tested with the same data as before. The initial words were entered by a native speaker who had no experience with speech recognition and phonetics, and these words were used by the Microsoft Speech Server to automatically derive pronunciations. On the test set, these pronunciations gave a WER of 34.6% – as expected, it is far worse than even the expert-based method. After one pass of the LTS-seeded, data-driven Salaam method, however, the WER was reduced to 8.9%.

5.5. Conclusion

The results described above are summarized in the following graph:

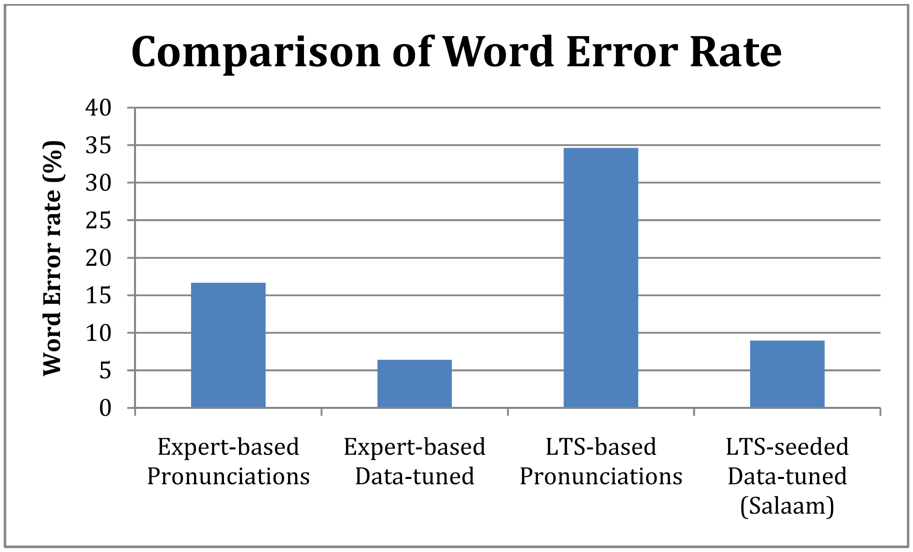


Figure 36: Empirical results from the various mapping methods

These empirical results show that research in speech interfaces for resource-scarce languages can be effectively done today, and that researchers need not wait for the existence of such resources before beginning to build, test, and even deploy such interfaces. Additionally, the high accuracy achievable through automatic methods suggests that such methods could create robust systems that could even be commercially deployed in such contexts.

The approach was demonstrated under the title "Speech Recognition In Any Language… While You Wait" at an international conference[13] held at Carnegie Mellon University in Qatar, Doha, in April 2009. Over a period of three hours, a variety of participants without any prior experience in speech recognition trained the system in the following 14 languages: Afrikaans, Bangla, Chinese, Hebrew, Hindi, Malayalam, Portuguese, Sesotho, Spanish, Swahili, Tamil, Twi, Urdu, and Welsh. The number of words trained ranged between 3 and 10. Participants were asked to type in words in their language as if they were typing them in an English SMS, and then based on their pronunciations, an initial guess of the pronunciations was manually entered by a speech expert. They were then asked to use a phone-based dialog system, which guided them to repeat each training word thrice, based on which the system learned pronunciations based on the Salaam method. The only significant issues seen were for 3 words in Portuguese. For the remainder of the languages and words, without any manual tweaking of the pronunciations, the accuracy was greater than 90%. While this is an informal evaluation, it suggests that there is great potential in the Salaam method for rapidly generating high-accuracy speech recognition capability in new languages.

[13] Information and Communications Technologies and Development: http://www.ictd2009.org

6. Conclusion & Contributions

6.1. Summary of Findings

6.1.1. Speech Recognition and Dialog Research Findings

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, coupled with the high penetration of cell phones, the research presented above proves that speech interfaces may be extremely viable for such users. Moreover, they may be useful even for literate users without access to smartphones and computers.

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.

Well-designed speech interfaces are viable modalities for information access by both low-literate and literate users in the developing world. Our work has focused specifically on workers that are in the network of a governmental or non-governmental organization that provides training, monitoring and ongoing support to them. In developing countries, this model is widely seen in public health, agricultural outreach, microfinance, and many other domains.

The Salaam method described in this thesis leads to highly accurate speech recognition. The automated method enables users without any speech recognition experience to be able to create robust speech recognition capability for their language by leveraging existing commercial speech recognition engines and models.

6.1.2. ICTD Research Findings

Feedback from participants in a user study with low-literate users is problematic. Participants do not speak out even when there are considerable problems in the user study (e.g. a system constantly repeating an error message).

Subjective evaluation from low-literate participants needs to be triangulated. Participants often try to be polite, and moreover, tend to blame themselves for problems with the interface they are testing.

Teaching an interface to low-literate participants in a user study is non-trivial. Verbal instruction, video-based instruction, and text-based instruction are not viable options to teach a low-literate user study participant how to use an interface. Human-guided tutorials are one possible solution to this problem.

The national language is often not the optimal choice for a speech interface for low-literate users. Such users are less likely to have had contact with formal schooling. It is advisable to build and test systems in the local language and specific dialect of the given users.

Text content designed for literate consumption cannot be used as-is for low-literate users. Significant adaptation by a content expert is required to make the content comprehendible in an oral medium. Long segments of text or audio must be avoided, and shorter segments with interactivity are required. Further, the structure of information in text documents may prove difficult for low-literate users to follow.

Significant training is required for low-literate participants to be able to use any interface. Even participants with 8 years of education had difficulty skimming a page to find its key points. Low-literate users were not able to successfully use either the touch-tone or speech interface without significant hands-on training.

6.2. Contributions

As a counterpoint to existing research, the results presented in this thesis show that with adequate user training, well-designed speech interfaces yield a significantly higher task success than equivalent touch-tone interfaces. Further, this thesis empirically proves that literacy is a statistically significant determinant of task success for both interface types.

This thesis also presented technical results from a novel speech recognition method for resource-scarce languages, showing that robust speech recognition capability in such languages may be had using existing technology today, without the need for linguists well-versed in each language. This has wide implications for the research and development of interfaces for resource-scarce languages.

Schultz & Kirchhoff (2006) state that the creation of a speech interface in any language requires the existence of a number of resources, which in most cases do not exist for resource-scarce languages in the developing world. These resources are:

1) Basic linguistic and cultural knowledge
2) Linguistic resources
3) Software tools
4) Guidelines for the design of user interfaces

This thesis presents novel research that contributes to the above resource requirements in the context of developing regions. The user studies present strong interface guidelines, as well as empirical results proving the effectiveness of interfaces built using these guidelines. The Salaam methods provide an approach that reduces the need for linguistic resources, and enables the interface developer to leverage software tools already created for existing languages.

As a recommendation for future work, we now present a framework on Orality-grounded Interface Design and Evaluation that provides a basis with which to examine cultural knowledge from an interface perspective, and also provides concrete guidelines for the design of user interfaces for low-literate users.

6.3. Future Work

6.3.1. The Relevance of Orality Theory

In ICTD literature, users that are not familiar with the tools of literacy (reading and writing) are commonly described as illiterate, non-literate, low-literate, or semi-literate – and it is these users that most ICTD interventions focus on. We suggest that a better understanding of these users can be achieved through the concept of orality, which has been notably theorized by Walter Ong in his seminal work "Orality and Literacy" (Ong, 1982), which has subsequently been extended by many other researchers[14].

Orality describes 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. The very categories through which such people are described is a case in point: they are described in terms of what they are not, such as "illiterate", rather than what they are: oral. Terms such as "illiterate" devalue the identity and knowledge of oral cultures by implicitly suggesting that lack of literacy is equivalent to backwardness. Orality theory emphasizes that we need to understand such communities in their own terms, rather than from the perspective (and biases) of literate users.

A summary of orality is given in Section 8. Most notably, it presents a number of differences in the "psychodynamics" of thought between oral and literate people – for instance, oral cultures do not have bulleted lists, nor hierarchically presented information (much less, the concept of 'navigating' an information hierarchy), nor even the term "to look up" information (which is a text-based term). In oral cultures, information is produced, transmitted, shared, and consumed in a local context, embedded in the human lifeworld, and is concrete rather than abstract. Further details on orality are given in Section 8.

The implications of orality for HCI in general, and HCI in ICTD(HCID) specifically, are profound. In the remainder of this section, we present a new framework, Orality-grounded Interface Design and Evaluation, which is based on the implications of orality for HCI.

6.3.2. Orality-grounded Interface Design and Evaluation

Drawing upon Ong's conceptualization of orality, we suggest the framework of Orality-grounded Interface Design and Evaluation. This framework calls for a fundamental rethinking of interface design and evaluation methodology, by drawing upon a deeper understanding of oral end-users. While the corollaries of the framework echo the sentiments of many user-centric methodologies that are popular in ICTD research (such as User-Centered Design, Participatory Design, and

[14] See http://www.oraltradition.org

Participatory Action Research), there is a fundamental difference: they do not provide a model for how such users think, learn, and process information. Orality-grounded Interface Design and Evaluation provides a testable (and falsifiable) model of oral users, and thus enables researchers to engage in more grounded research and design.

Below, we present recommendations that comprise the fundamental differences between standard HCI methodology and Orality-grounded methodologies for interface design, and for interface evaluation.

Interface Design

When designing interfaces for information access by oral users, several dimensions of orality are useful to consider:

Information needs to be rooted in common experience with specific examples. Abstract descriptions, such as “Breastfeeding should be continued even when the child has Diarrhea” is not as effective as describing a specific example of a mother who is faced with this particular issue. Ideally, **new information should be described in terms of familiar cultural memes, and preferably using the culture’s own oral formulae.**Thus, instead of using a generic “mother”, the information should draw upon existing characters (perhaps a widely known maternal character in this case) in the community’s folklore. Further, drawing upon a culture’s oral formulae is best left to local members of the culture. IDE hired local storytellers to write and perform a play centered around a treadle pump (Polak, 2008), which is an excellent strategy since local storytellers understand their culture’s oral formulae better than anyone else.

Narrative stories will work better than neutrally listed bullet points. Oral cultures do not have neutral lists – “even genealogies out of such orally framed tradition are in effect commonly narrative. Instead of a recitation of names, we find a sequence of ‘begats’, of statements of what someone did: ‘Irad begat Mehajael, Mehjael begat Methusael, Methusael begat Lamech’ (Genesis 4:18’) (Ong, pg. 97). Thus, listing bullet points of information such as how to improve cow yield will not work as well as telling the story of a farmer who used a specific method and how he was able to increase his yield (Gandhi, 2007). Moreover, **dramatic descriptions will work better than neutral ones.** IDE created a full-length movie that showed the value of a treadle pump through a dramatic storyline (Polak, 2008), which is much more effective than a neutral description of how such a pump could improve a farmer’s bottom line.

Rhythm aids recall. IDE hired local troubadours to compose a song about a treadle pump and to perform it at farmer’s markets and fairs in Bangladesh (Ibid). Even without the creation of an entire song, content with rhyme and alliteration is likely to be both understood and remembered by oral users more effectively than prose.

Linguistic style should be additively structured, not hierarchically. While it is common in literate material, using subordinative conjunctions such as “while”, “then”, “since”, “although” is not common in oral users’ communication. Complicated sentence structures impose a cognitive load on the user. Instead, it is preferable to use coordinating conjunctions, such as “and”, “or”, “so”, which do not create hierarchy.

Redundancy needs to be embedded in the content. Redundancy is an important part of oral communication, mainly because of the ephemeral nature of speech. The user should also be given ample opportunity to request repeated presentations of content that has been given before. However, it should be noted that explicitly requesting repetition may be less natural for the oral user than having the correct amount of redundancy already embedded within the content.

Each and every word needs to be understood. In an oral community that speaks one common language, there are no unfamiliar words, and oral users never face unknown words in daily life. Thus, even one unfamiliar word can confuse the user completely, and care should be taken to ensure that no such words exist in the system’s content. This is a common when content comes from experts, who are literate, and have a much more diverse linguistic background than an oral person.

Abstract categories should be avoided. Oral users do not categorize the same way as literate users, and do not think at all in terms of abstract categories (Luria, 1976). Thus, the use of categories should be minimized, and if their use is essential, it should be kept in mind that the designer’s choice of categories will most likely not match the expectations of the user. Also, **hierarchies in information architecture should be avoided.** Browsing multiple depths of information (e.g., as in a web-page, or when using navigation metaphors of “up a level”) is difficult for oral users (Deo, 2004).

Requiring adherence to words or phrases is less likely to succeed. Oral users perceive speech as a continuous stream, rather than as discrete words. In our work with speech recognition interfaces, the most common complaint of oral users is that they are afraid that they will “say the wrong thing”; they understand they have to speak in individual words, which is very unnatural for them, and their comment reflects their discomfort with this requirement. While it is may be technically difficult, it may be necessary to widen the speech recognizer’s grammar to allow for more natural utterances, rather than forcing the user to speak in single word commands.

Oral people do not internalize new information the same way as literate people. Since internalizing new information is comparatively expensive for oral people (they cannot offload their memory requirements onto the technology of writing), it appears that they are more selective when choosing whether to internalize any new information. In our research, we have seen many times that even when users understand each word in the content perfectly, when asked a question on that content, users will respond based on prior knowledge. For instance, even when literate health workers read a specific paragraph in their manual fluently, their response to a question on that content was the opposite of what it said. Similar effects have been found in other research on oral people (Luria, 1976), and poor readers (Doak, 1995), and has also been anecdotally reported by other HCID researchers (Etienne Barnard, personal communication). This might happen because the (formal) content is not seen as relevant, or that it isn’t organized in the same linguistic style that users expect (e.g. narrative, concrete examples, instead of neutral and abstract information), or for other reasons – however, it is important in the design of an interface to know that this is a problem that needs to be engaged with.

Oral people give more importance to the source of information than literate people. Writing establishes "context-free language" (Ong, 1982), where information is not linked to any particular source. For oral people, however, all information is social, and traceable to a person. When farmers were shown videos of other farmers demonstrating agricultural best practices, the most common question they asked was what the demonstrating farmer's name was, and which village he was from (Gandhi, 2007). This may also explain why source-neutral information (of the form presented in a book, or in a persona-less system) is not internalized, because it is missing the essential feature of social context – not only of the information's relevance to the real world, but also its traceability to a trustworthy human being. Thus, systems (such as Digital Green) that provide information through people (perceived to be trustworthy) are likely to be more successful than systems that transmit information without this feature.

Given all of the recommendations listed above, it becomes apparent that **information designed for literate users (virtually all written material) is not appropriate for oral delivery.** Oral and literate users require content with very difference organization, presentation and context.

Furthermore, **end users may be the best resource for content creation and content adaptation.** Literate people produce both written and spoken content optimized for literate people (Ong, pg 56). Even "human access points" (Marsden, 2008) – literate technologists from the local community – may not understand how to effectively alter content for optimal consumption by oral users, although they would do a better job than a naïve, non-local designer. However, it may be the case that involving users is essential for content creation and content adaptation. The success of the Digital Green project is a case-in-point – it doesn't just involve users in the creation of the content, but features them as the star performers in the "Farmer Idol" videos that form the backbone of the system.

Interface Evaluation

Overall, standard user studies do not work for oral people. Even though literate participants may find user studies a novel concept, they are likely to have experienced the individual components of a user study in school, and would likely consider them somewhat familiar. For oral people, a user study is an alien experience. The various steps involved in a user study – the facilitator reading out instructions to the participant, teaching the steps in the task and then asking the participant to perform it themselves to answer examination-style questions – is a clinical abstraction that is completely alien to the lifeworld of a typical oral person. Oral people do not think and learn in the way that user studies expect them to, and they are not used to being asked abstract, context-free questions. Thus, it is arguable whether results from such studies are of much analytical value in oral contexts.

Further work is needed to explore alternative methods for user studies adapted to the HCID context.

The context and motivation for the system should not be presented in the abstract. Given oral users' preference for situationally-grounded examples relevant to their lifeworld, it is imperative to present the system as a solution to a problem that users deem important, and for which users perceive current solutions as inadequate. When possible, **the system should be introduced as a solution to a widely remembered, specific instance of a local problem.** For example, in healthcare, there may be a recent story of a person affected by an illness that no one knew how to cure, which everyone remembers vividly, perhaps because of its tragic nature. By asking locals about such examples beforehand, and then portraying the health information access system as a potential solution as part of a narrative tying in the local example, potential users may see the social context of the system more clearly than they would if it was to be given as a solution to the problem *in general* (e.g., "imagine if a person in your village became ill…"), which is still better than a generic solution *in the abstract* (without a narrative, e.g., "the system can be used to give health information when the need arises"). A motivation presented with a narrative tied to the specific local folklore of the community could make the intervention seem more relevant to the needs of the community.

This suggests why the full-context video method (Medhi, 2007) works so well – because it creates a dramatized and visual example of how the interface could be useful, instead of introducing it in the abstract. It is unclear whether a video (which is difficult to customize to the needs of a specific local community) is more effective at providing context than face-to-face interaction, with a narrative that is customized to the specific experiences and memory of the user or community (which is less vivid than a video). Given the high success rate of the full-context video method (100% of the users exposed to the video were able to successfully complete the task), there may be no need for improvement.

Oral participants do not remember neutral user study tasks, and do not actively engage with them. It is a known issue that artificial scenarios and fictitious needs are unlikely to be internalized by participants. For oral users, tasks cannot be presented textually (and so cannot be reviewed at will), and further, if the task is abstract and neutral (standard HCI methodology), it is very unlikely that it will even be remembered, much less internalized. Engaging with a task that one does not remember or believe in is also unlikely. The Bollywood method (Chavan in Shaffer, 2004) is an excellent solution to this problem, since it couches the task in a narrative, and moreover, a dramatized narrative. **Oral users are much more likely to remember and engage with a dramatic narrative** than with an abstract and neutral request for information. It is unclear how much of the difference is attributable to a) participants' improved memory of the dramatized tasks, b) their appreciation for these tasks' wider social context and relevance to the human lifeworld, c) their perceived freedom to critique when in role play (what Chavan attributes its success to), or other factors. Further work is needed to explore what aspects of this method work well, and why.

Giving user study participants descriptive instructions on how to use the system is not sufficient. Oral people do not learn by verbalized instruction – rather, they learn through apprenticeship and practice. This is neatly analogous to the difference between declarative and procedural knowledge in the field of Artificial Intelligence. In our own work, we have found that a mentor-apprentice model, where a facilitator does a guided walkthrough of a sample task with a participant, works very well. Also, participants' understanding of the system improves further with an "incrementally-removed training wheels" approach with multiple practice tasks, where explicit help is provided on each step for the first sample task, and is gradually reduced for each subsequent practice task, until no help is provided on the final sample task unless absolutely required. The approach used in the full-context video method is different, and we hypothesize that oral users would learn better with a personal, mentored learning experience than by watching a video demonstrating another person's use of a system. Further work is needed to ascertain the best strategy for teaching the interface to the user – both for the design of the system, and also for evaluating the system in a user study.

Likert Scales are not an appropriate measurement tool for oral people. A Likert Scale question consists of an abstract statement, such as "I found this system easy to use", and a list of numbered options (usually from 1 through 7), with 1 representing "strongly disagree" and the highest number (e.g. 7) representing "strongly agree". Participants are asked to select the option that most closely describes their level of agreement with the given statement. When used with oral people, this tool is problematic for a number of reasons.

- The Likert Scales method requires that the options to be written down and presented textually – however, oral people cannot read, and verbal presentation of Likert Scales has not been proven as a valid methodology.
- The concept of "context-free", abstract statements is a literate construct. Oral people do not conceive of statements in this form – there has to be *someone* who said any given statement. Moreover, who this *someone* is makes a difference to how likely they are to agree with the statement: issues of credibility, trust, authority, politeness, political considerations, etc. are involved.
- Categorizing agreement into discrete chunks is also a literate construct. Oral people do not think in terms of such categories, and we have seen countless examples of oral user study participants confused at such categorization.
- Praise and criticism is often given indirectly in oral cultures, and thus asking participants to directly praise or critique an artifact with their subjective evaluation might not be in line with participants' expectations.
- Self-analysis is usually difficult for oral people (Luria, 1976). Thus, participants may not be able to verbalize what they really feel about a system, much less quantify it.

The "auto-rick radio" method (Chavan, 2007) provides an interesting alternative to Likert Scales. Instead of representing a subjective evaluation on a linear scale, participants are asked to twist a knob (shaped like a radio's volume knob) to represent their feelings on a particular dimension. Chavan states that "for these users, the concept of a difference in degree (moving from negative to positive) being represented by a horizontal straight line seemed very conflicting. The feeling was that if the different points in the scale represented different degrees of an attribute, then they could not appear to be on the same level, as they did with the straight horizontal line… Hence a knob control was devised which resembled the volume control knob of the radio that all users were very familiar with." (Ibid) However, this explanation confounds various issues: the difficulty in quantifying emotion for oral users, the difficulty in relating emotion to spatial terms, and the difficulty of expressing their choice using a novel interface. By providing a familiar interface (the radio volume knob), participants may have found it easier to express themselves, but whether their expression correctly represented their subjective emotions remains to be seen. Rigorous methodological research is required to explore these issues in detail.

One potential alternative to Likert Scales in general is to instead solicit open-ended answers and have the experimenter decide on the appropriate category – after all, what is the point of asking a participant to do something (quantify emotion) that they really cannot do? Chavan suggests "using storytelling to find dissonances" as a means of asking users about their experience with a particular product (Ibid) – however, this method could be extended to subjective measurement, where participants could be asked, for instance, to narrate instances where they had trouble while performing the user study task (as an alternative to asking them to subjectively rate its difficulty). The experimenter could use the quantity and quality of these instances to determine a quantification of the user's experience, if desired. However, if the participant and experimenter are both present in the same room during the study, the participant would likely wonder why the experimenter is asking such questions, since surely s/he saw the participant's performance and thus should obviously know the answers. Thus, care would need to be taken to set up the scenario appropriately (perhaps by having a different experimenter conduct the subjective evaluation) to ensure that the participant views these questions as valid and worthy of a real response.

Thus, it can be seen that standard user studies have various layers of problems, and while the above discussion point out some suggestions for how the methodologies can be adapted, it may be the case that a fundamentally different method of evaluation needs to be explored for the HCID context. Alternatively, testing deployed interfaces "in the wild" should yield much more representative results than user studies would (much more so than in literate contexts). There are a significant number of research projects designed this way, such as Digital Green (Gandhi, 2007), eSagu (Reddy, 2007), and Warana Unwired (Veeraravaghan, 2007), and this may be a better (though in most cases, more difficult) method of evaluation.

In Summary

While the above framework may be thought provoking, further research is required to fully understand its implications for interface design and evaluation. Clearly, there is much work to be done in integrating orality theory with HCI methodology. In this section, we have aimed to begin this process, by outlining various recommendations and corollaries of Orality-grounded HCID. However, much work is need to substantiate, qualify, verify, refute, and build upon these recommendations.

Specifically, in terms of speech interfaces, we can see that abstract interaction in the form of single-word, hierarchical menus can potentially be improved by applying the principles presented above. Further work is needed to apply and test these ideas to discover new interaction paradigms targeted specifically towards oral users. Given the diversity of users in any given region, it is highly likely that the findings from such research would be widely applicable to users in the so-called "developed world" as well.

6.4. Other Directions

In this thesis, a number of prototypes were designed, developed, and tested with community health workers of varying literacy levels. An alternative approach to this would be to instead create a Wizard-of-Oz system open to community health workers, where they would be able to ask questions, at any time, with responses by a trained agent with access to a large information database. These interactions could then be analyzed in terms of various metrics, for the purpose of both understanding what information is important to such health workers and the dialog structure of the conversation between the health worker and the agent – which could then be used to create an automated system. The system could initially allow free-form conversations, and over time, this could be constrained to a more automated approach. This would also test the demand for such a system.

While this approach would be informative, and would most likely yield substantial findings, there are a number of concerns that should be kept in mind. First, it is unclear how low-literate health workers would react to a system (that they have learnt and gotten used to) that changes – in fact, performs worse – over time. Moreover, the natural quality of free-form conversation is significantly different than constrained interaction, and so the switch from open to constrained conversation may cause substantial confusion on the part of the users. Next, the social cost of speaking to a human being may result in health workers not wanting to ask the agent to repeat information again and again until they understand it – while this may be perfectly acceptable behavior with an automated system. Finally, the opposite may also be true – speaking with a human operator may provide enough excitement and incentive that would not exist with an automated system.

7. Appendix A: Technical Solutions

7.1. Urdu language text tools

While Unicode-based Urdu language support has recently been standardized and made available, a proprietary application called InPage has been the *de facto* standard for Urdu word processing for the past decade, with its own encoding format. For the HANDS pamphlet, the text was already in InPage format, while the AKHSP pamphlets needed to be manually typed up by a typesetter – who was familiar only with InPage. Thus, these InPage files needed to be converted to Unicode, although none of the available online tools for this purpose were completely accurate. Some of the discrepancies occurred as a result of fundamentally different methods of representing the same letter with its diacritics between InPage and Unicode. Again our solution was to write such a converter ourselves.

7.2. Large corpora Urdu language audio recording

Recording the reading out of multiple pages of Urdu text required another component to present each sentence to the voice talent and record them speaking it aloud. Such software is commonly found in speech synthesis packages (as well as in MSS) – however, none of the available options were ideal for our use. First, these applications displayed each sentence in isolation – which is optimized for recording meta-prompts that have limited relation to one another. However, in our case, the voice talent needed to record not just meta-prompts (e.g., “What would you like to hear next?”) but also the content itself. The content was composed of paragraphs and paragraphs of text, with each sentence needing to seamlessly lead to the next. Thus, the voice talent needed to be able to see the rest of the text, so that she could have an idea of how to place emphasis based on context (paragraph end, bulleted list, etc.). Audio for each sentence was recorded separately, since a single error in a paragraph would require the entire paragraph to be re-recorded. Next, the text had to be in a readable Urdu font – other recording applications either did not support Unicode, or if they did, did not allow font changing. Our solution again was to write our own component. All audio was recorded in 44.1kHz, 16-bit, and downsampled to 8kHz for use in a telephony environment. Recordings were done using an M-Audio Fast Track Pro USB audio device, with a Shure SM-58 microphone, and a pop filter.

7.3. Mobile User Study Architecture

For telephony access to the system, we used two desktop computers: one for speech recognition, and one for telephony I/O, connected to a telephone line. These desktops were housed in Karachi, the largest city in Pakistan, where researchers were based. This approach meant that when field testing the speech interface in rural areas, it was not possible to make modifications to the system, even though the most crucial lessons were learnt in the field. It is striking that other researchers investigating telephony interfaces for low-literate users have faced the same problem (Nasfors, 2007).

In some cases, we were able to use a commercial remote desktop application on a mobile device to access the system and make changes when in the field – however, this approach was useful mainly for minor bug fixes, not for significant interface changes.

To solve this problem, we developed a method to shrink the footprint of our servers to a single laptop, along with a Voice-over-IP device (roughly the size of two decks of cards) and a regular desk telephone set. This enabled us to have the server completely self-contained in a mobile form factor, and meant that all aspects of the system could be modified in the field. In the final user study, this allowed us to make significant changes based on lessons learnt during a pilot study before beginning the larger user study.

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:

- A Windows Server running Microsoft Speech Server, containing all the logic for the information access interfaces, also running a Voice-over-IP gateway
- A Linux server running Asterisk/Trixbox for Voice-over-IP support, with a hardware telephony interface for analog telephone lines
- 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. In one instance, this meant postponing a user study even though all arrangements had been made.
- 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 (the size of a 3.5" hard drive) 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.

8. Appendix B: An Overview of Orality

8.1. Orality vs. Literacy

We argue that HCID research can be substantially improved by grounding it in an understanding of orality. To design for oral users, it is imperative to explore how they a) organize and transmit information, b) learn information, and c) remember information. Additionally, we need to grasp what Ong refers to as the "psychodynamics of oral thought" to understand the fundamental differences between oral and literate users. Below, we expand on each of these aspects.

8.1.1. How information is organized and transmitted

A key difference between oral and literate users is that knowledge production and transmission is a communal process instead of an individual one. The sum of an oral community's knowledge – its history, its identity, its culture, and its whole way of life – is represented and remembered in the oral tradition, often referred to as folklore. Amongst other dimensions, this tradition comprises stories, myths, proverbs, riddles, poems and songs. These forms provide a pool of wisdom from which individuals draw upon as a means of daily survival, as well as in dealing with new situations.

Notably, this knowledge is not statically stored – rather is it kept alive through constant verbal, extempore performance. Storytellers dynamically render different parts of this knowledge at different occasions, altering the content based on the current political scenario, social sentiment and immediate audience. Since it is never written down, it is neither fixed in form or content, nor repeated word-for-word.Hence, knowledge is dynamic and evolving, fluid and creative, and is composed iteratively over generations.

One of the striking characteristics of oral knowledge is that it "knows no lists or charts or figures." (Ibid, pg. 97) The emphasis is on narrating events in time, with a correspondence between these narratives and human experience.

8.1.2. How information is learnt

According to Ong, "human beings in primary oral cultures… learn by apprenticeship… by discipleship, by listening, by repeating what they hear, by mastering proverbs and ways of combining and recombining them, by assimilating other formulary materials, by participation in a kind of corporate introspection—not by study in the strict sense." (Ibid) What is especially relevant to HCID practitioners is what is *not* mentioned in the above quotation: oral people do not learn from a neutral, stand-alone object (such as a book, or automated system) which contains a set of abstract instructions to be applied across situations. Rather, they learn *in situ*, embedded in concrete situations and practical experience.

8.1.3. How information is remembered

Since information in oral cultures is transmitted through oral delivery – and hence it is *recited* and *heard* – it needs to be presented in a form that is conducive for aural-oral reception and retention. As Ong elaborates:

"In a primarily oral culture, to solve effectively the problem of retaining and retrieving carefully articulated thought, you have to do your thinking in mnemonic patterns, shaped for ready oral recurrence. Your thought must come into being in heavily rhythmic, balanced patterns, in repetitions or antitheses, in alliterations or assonances, in epitethic and other formulary expressions, in standard thematic settings… in proverbs which are constantly heard by everyone so that they come to mind readily and which themselves are patterned for retention and ready recall, or in other mnemonic form. Serious thought is intertwined with memory systems. Mnemonic needs determine even syntax."

In short, rhythm, repetition, re-use of locally-known idioms, and dramatized settings are key to keeping orally constructed knowledge alive and remembered.

8.1.4. Psychodynamics of Oral Thought (contrasted with Literate Thought)

Additive, not Subordinative

While the preferred grammatical form of literate writing is subordinative, oral thought prefers additive forms. Ong's example from an older translation of the Bible (one considered to have a large amount of oral residue) contrasts neatly with a modern translation. The older translation reads:

In the beginning, God created heaven and earth. **And** the earth was void and empty, **and** darkness was upon the face of the deep; **and** the spirit of God moved over the waters. **And** God said: Be light made**. And** light was made. **And** God saw the light that it was good;**and….**

The word "and" is used to join phrases and sentences liberally, which is not the case with a modern translation:

In the beginning, **when** God created the heavens and the earth, the earth was a formless wasteland, and darkness covered the abyss, **while** a mighty wind swept over the waters. **Then** God said, "Let there be light", and there was light. God **then** separated the light from the darkness…

Here, the use of conjunctions to join phrases is the common case, with a much more hierarchically and subordinatively organization than the additive pattern shown in the previous passage. Ong argues that the former example is as natural to the oral mind as the latter example is to the literate mind.

Aggregative, not analytic

Oral tradition favors the use of reusable formulas. Thus, it is not just "the soldier", but "the brave soldier"; not just "the princess", but "the beautiful princess"; and not just "the oak", but "the sturdy oak". These clusters aid memory by creating archetypal concepts that are repeatedly used in different narratives. Further, "traditional expressions in oral cultures must not be dismantled: it has been hard work getting them together over the generations, and there is nowhere outside the mind to store them… Once a formulary expression has been crystallized, it had best be kept intact. Without a writing system, breaking up thought – that is, analysis – is a high risk procedure" (Ibid, pg. 39). These reusable formulae are in some ways the fundamental building blocks of oral tradition.

Redundancy

All spoken information is ephemeral, and if information were never repeated, even the slightest loss of concentration would mean a complete misunderstanding on the part of the listener. Redundancy is necessary to ensure that the speaker and listener both remain on track. When reading, however, it is not necessary to repeat any information, as the reader can always backtrack to re-read earlier material – to the point where this becomes second nature. The differing economies of hearing and reading result in a significant difference between the outputs of oral and literary content in terms of the redundancy inherent in the content. Even though redundancy may be a natural part of the thought process, "with writing, the mind is forced into a slowed-down pattern that affords it the opportunity to interfere with and reorganize its more normal, redundant processes" (Ibid, page 40).

Conservative / Traditionalist

A corollary of redundancy is conservatism. Since oral knowledge vanishes unless repeated again and again, oral cultures place a premium on repeating previously-held knowledge, rather than experimenting and discovering new knowledge. To the technologically-minded, this can best be explained with an analogy to hard drive capacities: when hard drive space was expensive, one had to keep only a limited amount of data, which was most likely static – however, once capacity became cheap, it was easier to keep storing newer and newer data, as there was no need to choose between the old and the new. Oral cultures function with the former mindset.

This is not to say that oral cultures never update their knowledge store or that they lack originality. "Narrative originality lodges not in making up new stories but in managing a particular interaction with this audience at this time—at every telling the story has to be introduced uniquely into a unique situation, for in oral cultures an audience must be brought to respond, often vigorously. But narrators also introduce new elements into old stories" (Ibid, pg 41).

Close to the Human Lifeworld

"In the absence of elaborate analytic categories that depend on writing to structure knowledge at a distance from lived experience, oral cultures must conceptualize and verbalize all their knowledge with more or less close reference to the human lifeworld, assimilating the alien, objective world to the more immediate, familiar interaction of human beings… An oral culture has no vehicle as neutral as a list… [it] likewise has nothing corresponding to how-to manuals for the trades." (Ibid, pg. 42) Trades are learnt primarily through apprenticeship, involving observation and practice, with minimal verbalized knowledge if any. Further, oral cultures do not preserve knowledge of skills in abstract, self-contained corpora.

Finally, oral people do not treat all human-lifeworld-based information as equally important – it is *their* specific lifeworld that is relevant, and hence, important. For instance, they give specialized names to details that matter in their specific lives, but generalize to others that do not (e.g., “that is merely a flying animal”).

Agonistically Toned

To literate individuals, oral people might appear to be extraordinarily agonistic or argumentative. This is because stories and proverbs – the basis of oral tradition – are not just meant to be a store of knowledge, but also as means to engage a dialectic dialog. It is this argumentation that Ong considers to be the predecessor of the dialectic method of Socrates and Plato. Thus, oral thought is often instantiated as an interplay of competing ideas.

Empathetic and Participatory

With literacy, knowledge is disengaged and is supposedly “objective”. In an oral culture, “learning means achieving close, empathetic, communal identification with the known”. Oral people usually do not memorize (other than for rituals) – instead, learning involves an amalgamation of the new with the self. “Literates are usually surprised to learn that the bard planning to retell the story he has heard only once wants often to wait a day or so after he has heard the story before he himself repeats it. In memorizing a written text, postponing its recitation generally weakens recall. An oral poet is not working with texts or in a textual framework. He needs time to let the story sink into his own store of themes and formulas, time to ‘get with’ the story. In recalling and retelling the story, he has not in any sense ‘memorized’ its metrical rendition from the version of the other singer” (Ibid, pg 59).

Homeostatic

On the order of generations, oral cultures eliminate memories that are obsolete, simply by not repeating that information. As a consequence, oral cultures never need dictionaries, since all words in usage are commonly known and understood by everyone in the culture. The number of words never grows unmanageably, as obsolete words are pruned away. Literate cultures, on the other hand, can access all words from the present day up till thousands of years ago, and require dictionaries to be able to store and access these words.

Situational, not Abstract

Perhaps the most fundamental difference between oral and literate cultures is that of situational vs. abstract thinking. Because oral knowledge is rooted in the human lifeworld, oral people are most comfortable in thinking and learning situationally, instead of in the abstract. This does not mean that oral people cannot think categorically; it is just that they make categories differently than literate people. For example, when asked to categorize a set of objects, oral people categorize based on features important to the human lifeworld (e.g. usefulness) as opposed to abstract features (e.g. function). For instance, “subjects were presented with drawings of four objects, three belonging to one category and the fourth to another, and were asked to group together those that were similar and could be placed in one group or designated by one word. One series consisted of drawings of the objects *hammer, saw, log, hatchet*. Illiterate subjects consistently thought of the group not in categorical terms (three tools, the log not a tool) but in terms of practical situations—‘situational thinking’—without adverting at all to the classification ‘tool’ as applying to all but the log. If you are a workman with tools and see a log, you think of applying the tool to it, not of keeping the tool away from what it was made for—in some weird intellectual game” (Luria, 1976 in Ong, 1982). Further, any form of thought other than grounded, “operational thinking” is likely to be considered “not important, uninteresting, trivializing.” (Ong, 1982)

Operational thinking is also the mode in which oral individuals interpret reality. For example, Luria found that oral people interpreted circles as a plate, sieve, bucket, watch or moon, although school-going children readily identified them as circles (Luria, 1976).

Likewise, oral people do not think in formal, syllogistic logic. For instance, when given the query, “precious metals do not rust, gold is a precious metal – does it rust or not?”, one oral respondent said: “Do precious metals rust or not? Does gold rust or not?” Another said, “Precious metal rusts. Precious gold rusts.” Syllogisms are special riddles where the conclusions are derived from the given premises alone. “Persons not academically educated are not acquainted with this special ground rule but tend rather in their interpretation of given statements, in a syllogism as elsewhere, to go beyond the statements, as one normally does in real life or in riddles (common in all oral cultures).”

Further, requests for definition are resisted in oral cultures. When asked to explain what a tree is, one oral peasant replied, “Why should I? Everyone knows what a tree is, they don’t need me telling them.” Ong’s response to this is “why define, when a real-life setting is infinitely more satisfactory than a definition… The peasant was right, there is no way to refute the world of primary orality. All you can do is walk away from it into literacy.”

One might argue that the above questions were not asked in the correct context. However, it appears that there is no conceivably “correct” way to ask oral people such questions. In fact, in oral cultures, intelligence is something that is measured based on a person’s actions or skills – not by their response to verbal questions in a test: “written examination questions came into general use (in the West) only well after print had worked its effects on consciousness, thousands of years after the invention of writing” (Ibid, pg 55). Thus, asking abstract questions to test someone’s knowledge without any social context is an alien concept in an oral culture.

It is important to note that just because oral people think situationally or concretely does not mean they are incapable of abstract thought. Indeed, there are complex layers of meaning in oral knowledge, that are captured in idioms, folklore, mythology, riddles and historical narratives.

9. References

Abraham, R. (2006). ***Mobile Phones and Economic Development: Evidence from the Fishing Industry in India.*** Proc. International Conference on Information and Communications Technologies and Development, 2006

Ackerman, S. (2002). ***Mapping User Interface Design to Culture Dimensions***. Proc. International Workshop on Internationalisation of Products and Systems

Afsar, H. and Younus, M. (2005). ***Recommendations to Strengthen the role of Lady Health Workers in the National Program for Family Planning and Primary Health Care in Pakistan: The Health Workers Perspective***. Journal of Ayub Medical College, Jan-Mar 2005. http://www.ayubmed.edu.pk/JAMC/PAST/17-1/HabibYounus.htm Accessed 30th Aug 2006

Bansal, D., Nair, N., Singh, R. & Raj, B. (2009). ***A Joint Decoding Algorithm for Multiple-Example-Based addition of Words to a Pronunciation Lexicon***, Proc. ICASSP 2009.

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

Best, M., Etherton, J., Smyth, T. & Wornyo, E. (2009). ***Uses of Mobile Phones in Post-Conflict Liberia.*** In Proceedings of International Conference on Information and Communications Technologies and Development, Doha, Qatar.

Brand, P., Schwittay, A. (2006). ***The Missing Piece: Human-Driven Design and Research in ICT and Development*. **Proc. International Conference on Information and Communications Technologies and Development, 2006.

Brewer, E., Demmer, M., Du, B., Fall, K., Ho, M., Kam, M., et al. (2005). ***The Case for Technology for Developing Regions***. IEEE Computer. Volume 38, Number 6, pp. 25-38, June 2005.

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

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

Davel, M. & Barnard, E. (2008). ***Pronunciation predication with Default & Refine***. Computer Speech and Language, Vol 22, pp 374-393, October 2008.

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

Doak, C., Doak, L. & Root, J. (1996). ***Teaching Patients with Low Literacy Skills***. http://www.hsph.harvard.edu/healthliteracy/doak.html, accessed on 9/10/08.

DeRenzi, B., Lesh, N., Parikh, T., Sims, C., Mitchell, M., Maokola, W., et al. (2008). **e-IMCI: Improving Pediatric Health Care in Low-Income Countries**, Proceedings of CHI 2008.

Donner, J. (2004). ***Innovative approaches to public health information systems in developing countries: An example from Rwanda***. Presented at "Mobile Technology and Health: Benefits and Risks". Hosted by the Department of Economics, Society, and Geography at the University of Udine, Udine, Italy. (2004)

Gakuru, M. & Tucker, R. (2009). ***Sustainability of Research and Development: A case of Successful Technology Transfer in Spoken Language Technology***, CHI Workshop on Human-Centered Computing in International Development, Boston, 2009

Gandhi, R., Veeraraghavan, R., Toyama, K. and Ramprasad, V. ***Digital Green: Participatory Video for Agricultural Extension.*** In Proceedings of ICTD 2007, pages 21-30, 2007

Geertz, C. (1978). ***The Bazaar Economy: Information and Search in Peasant Marketing***. American Economic Review, 1978, 68(2), pp.28-32

Heffernan, C. (2006). ***The Livestock Guru: Demand-led knowledge transfer for poverty alleviation.*** Proc. International Conference on Information and Communications Technologies and Development, 2006.

Hofstede, G. (1997). ***Cultures and Organizations: Software of the Mind***, McGraw-Hill, New York, 1997.

Hone, K., and Graham, R. (2000). ***Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI)***, Natural Language Engineering, Volume 6 , Issue 3-4, Pp 287 – 303, 2000, Cambridge University Press, New York, USA

Huenerfauth, M. (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.

ITU. (2003). ***Mobile overtakes fixed: Implications for policy and regulation***. http://www.itu.int/osg/spu/ni/mobileovertakes/Resources/Mobileovertakes_Paper.pdf. Accessed October 30, 2005.

ITU. (2009). ***Measuring the Information Society: The ICT Development Index****.* http://www.itu.int/ITU-D/ict/publications/idi/2009/index.html. Accessed March 25, 2009.

Jones M., Harwood W., Buchanan G., Frohlich D., Rachovides D., Lalmas M., Frank M. (2008). ***Narrowcast yourself: Designing for Community Storytelling in a Rural Indian Context***, DIS2008, 25-27 February 2008, Cape Town, South Africa.

Kam, M., Kumar, A., Jain, S., Canny, J. (2009). ***Improving Literacy in Rural India: Cellphone Games in an After-School Program***. In Proceedings of International Conference on Information and Communications Technologies and Development, Doha, Qatar.

Kahssay, H.M., Taylor, M.E., & Berman, P.A. (1998). ***Community Health Workers: The Way Forward***. World Health Organization.

Kirkman, G. (2001). Out of the Labs and into the Developing World: Using Appropriate Technologies to Promote Truly Global Internet Diffusion, Journal of Human Development, Volume 2, Number 2, July 2001.

Kumar, R. (2004). ***E-Choupals: A Study on the Financial Sustainability of Village Internet Centers in Rural Madhya Pradesh***, Information Technologies and International Development, Vol. 1 issue 1, pp.45-73, Spring.

Langner, B., Kumar, R., Chan, A., Gu, L. & Black, A. (2006). ***Generating Time-Constrained Audio Presentations of Structured Information***, Interspeech 2006, Pittsburgh, USA.

Luk, R., Zaharia, M., Ho., M. & Aoki., P. (2009). ***ICTD for Healthcare in Ghana: Two Parallel Case Studies***. In Proceedings of International Conference on Information and Communications Technologies and Development, Doha, Qatar.

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

Mansell, R., & Wehn, U. (1998). **Knowledge Societies: Information Technology for Sustainable Development**. New York: Oxford University Press

Marcus, A. (2003). **User-interface design and China: A great leap forward**. Interactions 10(1), 21-25.

Marsden, G., Maunder, A., & Parker, M. ***People are people, but technology is not technology***. Philisophical Transactions of the Royal Society A:1-10. 2008

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

Medhi, I. &Toyama, K. (2007). ***Full-Context Videos for First-Time, Non-Literate PC Users***. IEEE/ACM International Conference on Information and Communication Technologies and Development, Bangalore, India.

Medhi, I., Nagasena, G. S. N., & Toyama, K. (2009). ***A Comparison of Mobile Money-Transfer UIs for Non-Literate and Semi-Literate Users***. Proc. ACM Conference on Computer Human Interaction, Boston, USA, (2009).

Nasfors, P. (2007). ***Efficient Voice Information Services for Developing Countries***, Master Thesis, Department of Information technology, Uppsala University, Sweden, 2007

Ong, W. (1982). ***Orality and Literacy: The Technologizing of the Word***. New Accents. Ed. Terence Hawkes.

OPML. (2002) ***Evaluation of the Prime Minister's Lady Health Worker Program***. Oxford Policy Management Institute. http://www.opml.co.uk/document.rm?id=690 Accessed March 2nd, 2009.

Pakenham-Walsh, N., Priestly, C., & Smith, R. (1997). ***Meeting the information needs of health workers in developing countries***. British Medical Journal; 314:90.

Parikh, T. (2005). ***Using Mobile Phones for Secure, Distributed Document Processing in the Developing World***. IEEE Pervasive Computing Magazine, 4(2):74–81, April 2005.

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

Plauche, M., Nallasamy, U., Pal, J., Wooters, C., & Ramachandran, D. (2006). ***Speech Recognition for Illiterate Access to Information and Technology*.** Proc.

International Conference on Information and Communications Technologies and Development, 2006.

Plauche, M. & Nallasamy, U. (2007). ***Speech Interfaces for Equitable Access to Information Technology***. In Information Technology and International Development, Vol 4., No. 1, pp 69-86.

Polak, P. (2008). ***Out of Poverty***. Berrett-Koehler Publishers, 2008

Prahalad C. (2005). ***The Fortune at the Bottom of the Pyramid: Eradicating Poverty Through Profits*****.** First printing ed. New Jersey: Wharton School Publishing; 2005:361-379.

Rajalekshmi, K.G. (2007). ***E-governance Services Through Telecenters: The Role of Human Intermediary and Issues of Trust.*** In Information Technology and International Development, Vol 4, Issue 1.

Ramamritham, K., Bahuman, A., Duttagupta, S. (2006). ***aAqua: A Database-backended Multilingual, Multimedia Community Forum***. Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data.

Reddy, R. (2004). ***PCtvt: a Multifunction Information Appliance for Illiterate People.*** ICT4B retreat. http://www.rr.cs.cmu.edu/pctvt.ppt

Reddy, P., Ramaraju, G. & Reddy, G. (2007). **eSagu™: a data warehouse enabled personalized agricultural advisory system**, ACM SIGMOD International Conference on Management of Data, 2007.

Rosenfeld, R., Olsen, D. & Rudnicky, A. (2000). ***A Universal Human-Machine Speech Interface***. Technical Report CMU-CS-00-114, School of Computer Science, Carnegie Mellon University.

Rudnicky, A. I. (1993). ***Mode preference in a simple data-retrieval task***. Proceedings of the ARPA Workshop on Human Language Technology. San Mateo: Morgan Kaufmann, 1993, 364-369.

Rudnicky, A., Thayer, E., Constantinides, P., Tchou, C., Stern, R., Lenzo, K., Xu, W., Oh, A. (1999). ***Creating natural dialogs in the Carnegie Mellon Communicator System***, in Proceedings of Eurospeech, 1999, pp 1531-1534

Schultz, T. & Waibel, A. (1998). ***Adaptation of Pronunciation Dictionaries for Recognition of Unseen Languages***. Workshop on Speech and Communication (SPECOM-1998), pp 207-210, St. Petersburg, Russia, October 1998.

Schultz, T, and Kirchhoff, K. (Ed.). (2006). ***Multilingual Speech Processing**.* Elsevier, Academic Press

Shaffer, E. ***Institutionalization of Usability***. Addison Wesley Professional, 2004.

Sherwani, J., Tomko, S. & Rosenfeld, R. (2006). ***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.

Sherwani, J. & Rosenfeld, R. (2008). ***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.

Sherwani, J., Yu, D., Paek, T., Czerwinski, M., Ju, Y.C., & Acero, A. (2007). ***VoicePedia: Towards Speech-based Access to Unstructured Information***, Interspeech 2007, Antwerp, Belgium.

Sinha, C. (2005). ***Effect of Mobile Telephony on Empowering Rural Communities in Developing Countries.*** International Research Foundation for Development (IRFD). Conference on Digital Divide, Global Development and the Information Society.

Tongia, R. & Subrahmanian, E. (2006). ***Information and Communications Technology for Development (ICT4D) – A Design Challenge?*** Proc. International Conference on Information and Communications Technologies and Development, 2006.

Tongia, R., Subrahmanian, E., and Arunachalam, V. S. (2005). ***Information and Communications Technology for Sustainable Development***, A Report based on two workshops organized by: Carnegie Mellon University, Pittsburgh and Indian Institute of Science, Bangalore; Washington, D.C., 2003, and Bangalore, 2004, Allied Publishers, Pvt. Ltd. Gandhi Nagar, Bangalore, India, c. 2005

UNESCO. (2008). ***UN Educational, Scientific and Cultural Organization's Institute for Statistics: Global Education Digest 2008***. Quebec, Canada.

Veeraraghavan, R., Yasodhar, N. and Toyama, K. ***Warana Unwired: Replacing PCs with Mobile Phones in a Rural Sugarcane Cooperative***. In Information and Communication Technologies and Development, 2007. ICTD '07. International Conference on, 2007.

Vodafone. (2006). ***Africa: The Impact of Mobile Phones***. Vodafone Policy Paper Series Number 2 2005, http://vodafone.com/assets/files/en/AIMP_09032005.pdf. Accessed 30 August 2006.

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

WHO, (2006). ***The World Health Report 2006 – Working Together for Health****.* World Health Organization. http://www.who.int/entity/whr/2006/whr06_en.pdf

Williams, J. & Witt, S. (2004). ***A Comparison of Dialog Strategies for Call Routing***. International Journal of Speech Technology, 7(1): 9-24

WRI (2004). ***Lessons from the Field: Overview of ICT Use in Development***. http://www.digitaldividend.org/pubs/pubs_02_tele.htm.

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