Research Article

Speech Interfaces for Equitable Access to Information Technology

Madeline Plauché
International Computer Science Institute (ICSI)
1947 Center St. Suite 600
Berkeley, CA 94704
512-965-0218
plauche@icsi.berkeley.edu

Udhyakumar Nallasamy
Carnegie Mellon University
Pittsburgh, PA

Abstract

Speech recognition has often been suggested as a key to universal information access, as the speech modality is a "natural" way to interact, does not require literacy, and relies on existing telephony infrastructure. However, success stories of speech interfaces in developing regions are few and far between. The challenges of literacy, dialectal variation, and the prohibitive expense of creating the necessary linguistic resources are intractable using traditional techniques. We present our findings evaluating a low-cost, scalable speech-driven application designed and deployed in a community center in rural Tamil Nadu, India, to disseminate agricultural information to village farmers.

Introduction

Speech interfaces, or spoken dialog systems (SDS), allow users to control computer output (graphics, texts, or audio) by uttering key words or phrases that are interpreted using automatic speech recognition (ASR). Successful speech interfaces can accommodate the sight- or mobility-impaired by replacing or enhancing access to the computer output (screen) and input (keyboard, mouse) (Raman, 1997). Automated telephony systems are commercially available and are commonly used by businesses to reduce call center costs, as they are relatively affordable to run, once built. The main hurdle in replicating this success for the access of computing technologies in the developing world so far has been the prohibitive cost of computing devices, IT infrastructure, and designs for software and hardware that assume literacy and computer savvy (Brewer et al., 2006). Cell phones are affordable, the infrastructure they require is more readily available, and they are used extensively throughout the developing world. Often cell phones are shared by multiple users of varying degrees of literacy (Donner, 2004). Automated telephony services and other speech interfaces are attractive channels for information access especially among the oral communities in developing regions. For example, many applications for rural Indian IT that provide information on health, weather, employment, news, and agricultural could enhance the social and economic development of the rural poor (Sustainable Access in Rural India [SARI], 2005) and could be constructed around a limited-vocabulary speech interface. Others who have studied the positive impacts of speech-driven interfaces in developed regions, however, predict that such applications will likely not enhance the quality of life of those that rely on them (Jelinek, 1996).

To design a successful speech interface for use in rural India is to face considerable challenges. The ASR must perform in noisy environments where multilingualism and substantial dialectal variation are prevalent. The ASR must accurately recognize speech from languages for which neither

annotated corpora nor other costly linguistic resources exist. The front-end dialog interface should be interactive, easily adoptable, and accommodate individuals who lack formal education and computer literacy. The user interface (UI) will be effective only if it stems from a deep understanding of a community culture and value systems. However, design techniques developed for accessing sociocultural models in relatively wealthy European and North American communities are ineffective in poor communities, where leisure and formal education are spare. Finally, a successful speech interface is one that supports an application based on local content created by local providers, as the information needs of rural communities include news, events, and innovations happening within a radius of only a few kilometers.

A speech driven application for developing communities must address all of these issues in order to successfully extend IT access to the developing world. This article offers design requirements and solutions for the local content creation, ASR, and UI for speech interfaces in developing regions. We evaluate our solutions through a participatory design and deployment effort in which we collaborated with community partners to provide interactive agricultural information to rural villagers in southern India.

SDS Content

Relevant, local content is a large concern for developing regions (Chisenga, 1999). Attempts by local or national government or nongovernmental organizations to provide free, locally available health, job training, and education services that meet the basic needs of the public often do not reach unschooled populations in an accessible, reliable form. Radio and TV are affordable forms of mass media that can be effective at creating initial public awareness. However, they are much less effective in influencing people to improve their practices in health, agriculture, or education than traditional, oral methods of information dissemination that stem from within a community (Soola, 1988).

Historically, society has seldom given poor people ownership over the tools of production (Castells, 1997; Gordo, 2003). However, researchers in the field of IT for developing regions agree that involving community members in design and creation ensures that proposed solutions meet the needs of the community and provide the best chance for the sustainability of technology in a community (Braund & Schwittay, 2006). IT offers the opportunity and infrastructure for publishing and distributing all types of information in the shortest possible time and at the lowest cost. In particular, IT can be used by community partners to provide accurate, locally created information to unschooled adults in developing regions.

Information Needs

In rural Tamil Nadu, 37.47% of full-time workers and 71.64% of marginal workers work in the agricultural sector—a majority of them small cultivators or seasonal laborers (Figure 1). Across all developing regions of the world, farmers and other agricultural workers constitute over 40% of the labor force. The information needs of farmers in developing regions are likely to be vast and varied. Although the ability and inclination to base sale decisions on price information is open to question (Hornik, 1988; Blattman, Roman, & Jensen, 2003), studies have suggested that under the right circumstances, price and market information can improve farmer welfare (Eggleston, Jensen, & Zeckhauser, 2002; Prahalad & Hammond, 2002). Information on recommended techniques (pest and disease prevention, new seeds) improve production (Blattman, Roman & Jensen, 2003) and IT-based information networks can help raise the price of goods sold for small farmers (Kumar, 2004).

MS Swaminathan Research Foundation

MS Swaminathan Research Foundation (MSSRF) is an established nongovernmental organization (NGO) in Tamil Nadu dedicated to a pro-nature, pro-poor,

Figure 1. Rural farmers in the Povalamkottai village plaza, Tamil Nadu.

VRC Sempatti VKC Panzampatti

Figure 2. MS Swaminathan Research Foundation village centers.

and pro-women approach to fostering economic growth in rural, agricultural areas. Gathering and distributing accurate information for unschooled agricultural workers is central to MSSRF's efforts.

Trained community members in villages across southern India operate MSSRF village knowledge centers (VKCs), where they provide the rural poor with training and education on agricultural practices, health, social issues, and entrepreneurial development. The trained volunteers from each community, known as "knowledge workers," regularly communicate the needs of their neighbors to village resource centers (VRCs) through weekly meetings, a user registry, and door-to-door surveys. The VRCs, in turn, communicate needs to MSSRF headquarters (Chennai, Tamil Nadu), where information content (text, audio, video), additional training, video conferencing, and workshops are provided to address the needs of communities across the state. Our project was conducted in collaboration with Sempatti VRC, which is responsible for 9 VKCs in the region (Figure 2), each one serving between 2,000 and 11,000 people.

In addition to bridging knowledge between communities and MSSRF headquarters, the agricultural experts at VRCs meticulously document the crops grown in the region, including varieties, planting techniques, soil properties and fertilizers. VRCs function as regional libraries for the rural illiterate person, as they contain a wealth of short videos prepared locally by universities and community organizations on recommended agricultural and health practices. In addition, video conferences conducted between VRCs allow farmer groups or self-help groups to communicate experiences and local innovations across districts of Tamil Nadu using traditional, oral techniques.

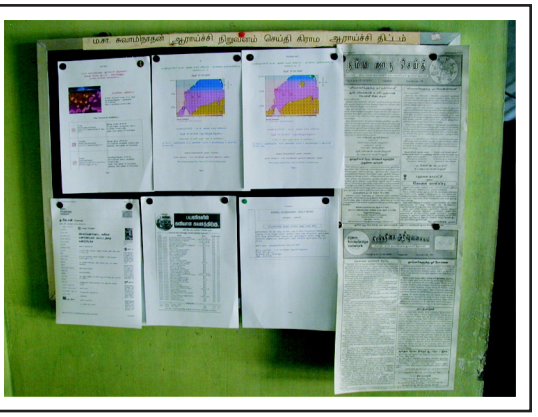

Figure 3. Daily postings of weather, news, and market prices at the Sempatti VRC.

Information from VKCs, VRCs, and headquarters is consolidated, digitized, and disseminated throughout MSSRF centers by phone, dial-up voice transfer, or wi-fi. Villagers may access accurate, up-to-date information at a nearby VKC by reading posted bulletins (Figure 3), listening to loudspeaker broadcasts, or working one-on-one with a knowledge worker (Balaji, et al., 2004). Knowledge workers often refer to a static, text-based HTML page called *Valam* (a

Figure 4. Valam ("Resources") Website available in all MSSRF centers.

Tamil word meaning *resources*) to retrieve locally relevant information on health, education, jobs, microcredit, self-help groups, and agriculture (Figure 4).

MSSRF staff in all three districts we visited reported a need to disseminate information in a form accessible to their predominantly oral population. In Sempatti, knowledge workers reported much success in educating illiterate people using a touch-screen kiosk, called the Livestock Guru (Heffernan, 2006). A major limitation to the tool, however, was its inability to be modified and updated. Throughout the design and evaluation of our technology, we relied on the experience of MSSRF staff in spreading relevant information across rural communities and adopted their belief that each village center is both a destination and a source for rich knowledge and information. Finally, the success of our project owes much to MSSRF's trusted role in the rural communities in which we designed and evaluated our technology.

ASR

ASR is the process of algorithmically converting a speech signal (audio) into a sequence of words (text). Although vendors of commercial systems and speech researchers often report the ability to correctly identify words from speech around 95% of the time, these numbers correspond to performance under optimal conditions (quiet, controlled environment, limited domain, single speaker). However, ASR is a nontrivial task because of the infinite variations of speech and will fail miserably in more challenging conditions (cocktail party, overlapping speech, etc.). State-of-the-art speech recognizers perform at only 80% on the Switchboard corpus, for example, a collection of near-natural, continuous speech recorded from multiple speakers during human-to-human telephone conversations. ASR decodes words from an audio signal by training hidden Markov models (HMMs) for phones, diphones, or triphones based on training data, generally a large corpus of speech that is hand-labeled at the phoneme or word level (Figure 5). ASR success depends on the collection and annotation of this training data, as well as the creation of a dictionary of all possible words in the language with all possible pronunciations for each word. A large vocabulary ASR also relies on language-level constraints captured by a language model either trained on a large text corpus or grammar rules meticulously created by a linguist. The creation of these linguistic resources (training data—speech, text and pronunciation dictionary) is arguably the most costly process of ASR development.

The availability of linguistic resources is taken for granted by developers who work in English, French, and Japanese, for example. The majority of the world's 6,000 languages, which are spoken in developing regions, currently have no associated linguistic resources. In India, there are two official languages (Hindi and English), 22 scheduled languages (including Tamil), and an estimated 400 more (Ethnologue 2006). The prohibitive cost of creating linguistic resources hinders the development of speech technologies for languages like Tamil, which is spoken by more than 60 million people in Tamil Nadu, Sri Lanka, and elsewhere (Ethnologue 2006). Equitable access to IT requires support for all languages, re-

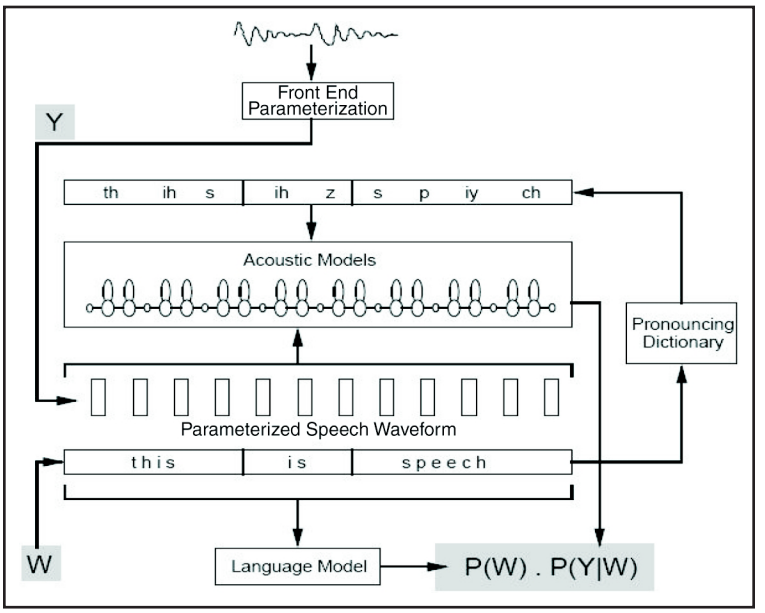

Figure 5. This diagram from Young (1996) shows the computation of the probability P(W\Y) of word sequence W given the parameterized acoustic signal Y. The prior probability P(W) is determined directly from a language model. The likelihood of the acoustic data P(Y\W) is computed using a composite hidden Markov model representing W constructed from simple HMM phone models joined in sequence according to word pronunciations stored in a dictionary.

gardless of their political dominance or number of speakers.

Researchers and developers of ASR strive for performance that mimics a human's capacity to understand speech (e.g., speaker-independent, large-vocabulary, continuous speech recognition). As a result, of the three basic principles to speech recognition performance (Table 1), the first principle, "The more data, the better," has dominated the direction of ASR research in languages with plentiful linguistic resources. One estimate, however, puts the amount of training data needed for current systems to achieve human levels of recognition between 3 and 10 million hours of acoustic training data, that is between 4 and 70 human lifetimes of exposure to speech (Moore 2003). The singular drive for more training data requires substantial cost, time, and expertise to collect and is ill suited to ASR in developing regions. In this article, we show how simplifying the recognition task and adopting adaptation techniques that tune the recognizer's models to match input data can achieve equivalent performance to better accommodate the economic and linguistic conditions of developing regions.

Field Study 1: Speech Collection for ASR

Speech recordings of rural villagers in three districts of Tamil Nadu were conducted in 2004 and 2005 to create adequate training data for machine recognition of a small vocabulary (less than 100 words) of isolated Tamil words. In the following sections, we evaluate the performance of a small Tamil word recognizer in the face of dialectal variation and limited training data. We focus on a recognition task that is relatively simple, yet adequate to power an SDS. In addition, we discuss the challenges of collecting speech data from unschooled adults and discuss alternative avenues for the creation of effective ASR models in regions such as Tamil Nadu.

Speech Recordings

During two field visits, we recorded the speech of 77 volunteers in three separate districts in Tamil Nadu, India (Figure 6). All volunteers were native speakers of Tamil over the age of 18. The researchers sought a balance of gender, education level, and age among participants, but the demographics of this study varied greatly by site (Table 2) and by recruiting method.

Coimbatore volunteers were either undergraduate students at Amrita University or laborers recruited by word of mouth; this method proved to be unsuccessful for the latter group. In Pondicherry, literate farmers and their wives were recruited as volunteers by MSSRF. In Madurai district, volunteers were recruited through MSSRF and Aravind eye

Table 1. Basic Principles of Speech Recognition Performance

The more data, the better.
The more input matches training data, the better.
The simpler the task, the better.

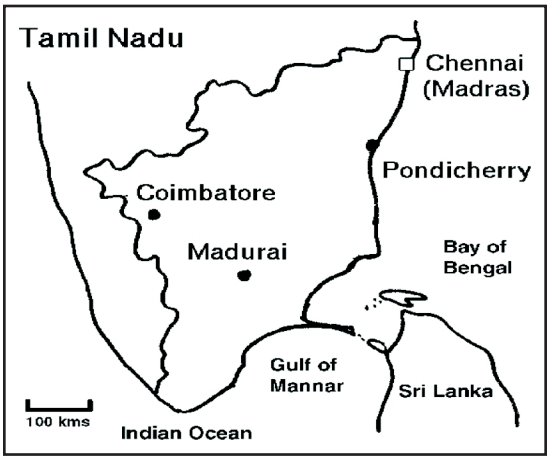

Figure 6. Map of Tamil Nadu, India.

Figure 7. Recording a literate woman in Ettimadai, Coimbatore district.

camps, where free eye care is provided in rural villages to between 200 and 500 people a day. We found that working alongside trusted organizations that serve the rural poor was the most efficient method for recruiting and recording villagers, especially those with little or no formal education.

Traditional data collection for small vocabulary speech databases relies on read speech in a quiet, controlled setting. Recordings for this study were conducted in the quietest available space, which, in many cases, was outdoors or in other public areas. Equipment and elicitation techniques varied by site; researchers had to be flexible to environmental conditions. Initially, literate Tamil speakers in Pondicherry (2004) were recorded saying 30 Tamil command words (e.g., "repeat," "send," "next") in relatively quiet offices with a lapel microphone or desktop microphone and a laptop running mouse-driven software that displayed a written word and recorded the speaker saying the target word.

Data collection in rural Tamil Nadu in 2005 relied instead on flashcards and a telephone handset with an embedded microphone connected to a Sony MD Walkman (MZ-NH900) (Figure 7). This allowed the speaker to comfortably hold the microphone close to the mouth but slightly to the side of the lips, to avoid "p-pops," bursts of high airflow during speech. In addition to capturing quality speech recordings in a variety of environmental conditions (average signal-to-noise ratio was 29), the modified telephone handset did not require us to clip or fasten equipment to the speaker's clothing and did not require the use of a table.

Bilingual flashcards with digits 0–10 written in both numerical and orthographic form were randomized and shown to speakers one at a time. Speakers were recorded reading the numbers aloud. The protocol was repeated five times per speaker. If a participant could not read the flashcards, a researcher or interpreter would translate the flashcards

Table 2. Number, Age, Gender, and Literacy of Speakers by Site.

		Average Age of		
Site (Year)	**Number of Speakers**	**Females**	**Males**	**% Nonliterate**
Coimbatore (2005)	15	20.7	31.4	13
Madurai (2005)	33	49.3	55.7	39
Pondicherry (2005)	7	37.5	47.3	0
Pondicherry (2004)	22	n/a	n/a	0
All Data	77	35.8	44.8	19.50

Here, "nonliterate" refers to speakers who could not read the flashcards and reported an inability to read or write their name. "n/a" signifies that speaker information is not available.

Figure 8. Recording an illiterate woman in Ettimadai, Coimbatore district.

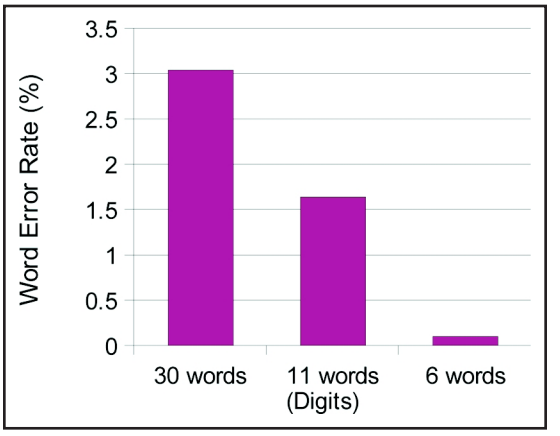

Figure 9. Word error rate by task complexity.

into a display of fingers (Figure 8). (A fist represented zero.) The flexible protocol provided a direct method for evaluating competency at literacy and numerical literacy. Both the flashcards and finger-counting methods were designed to elicit single word utterances free from external influences in word choice or pronunciation. All participants also answered a questionnaire to determine linguistic and educational background with the aid of a local interpreter.

Recording illiterate speakers saying the words for digits 0–10 took approximately six times as long as recording the same data from literate speakers. This discrepancy was due to difficulties explaining the task, limitations in the protocol (no reading aloud), inflexible and demanding occupations of participants, and apprehension involving participation, resulting in many missed appointments. In addition, illiterate speakers in this study had limited access to infrastructure appropriate for recording (no housing, no power, public buildings) and longer social protocols for foreign visitors.

Tamil Word Recognizer

All speech collected in Tamil Nadu was recorded at 44 kHz, stereo, then downsampled to 16 kHz, mono. Each of the approximately 10,000 speech samples were extracted with a collar of 100 ms of silence and labeled by hand in the laboratory. A whole-word recognizer was trained on the speech from 22 speakers in Pondicherry (2004) using the Hidden Markov Toolkit (HTK). The models used 18 states with 6 diagonal gaussians per state. Whole word models were used to avoid the need for a pronunciation dictionary. The following section describes various experiments to show how ASR accuracy depends on task complexity or vocabulary, dialectal variations, and amount of training data.

Experiment 1: Task Complexity

The speech collected from our field recordings was input for recognition in three trials of varying complexity: all words, digits only, and six command words. As expected, word error rates dropped for tasks with fewer options for correct word identity (Figure 9). An SDS with a small vocabulary of command words, or one that limits word options at each node of the dialog turn would require very little training data (less than 3 hours) to achieve accurate recognition.

Experiment 2: Dialectal Variation

Multilingualism, dialectal, and accent variations are prevalent in developing regions. India has 22 "scheduled" (official) languages but estimates range from 450 (SIL, 2005) to 850 languages (Noronha, 2002) overall. India is the ninth most linguistically diverse country, with a 93% probability that any two people of the country selected at random will have different mother tongues (Ethnologue, 2006). The Tamil language, like most languages, varies by geography (six main dialects), social factors (caste, education), and register (spoken vs. written). Recording the words for digits in three different districts of Tamil Nadu revealed that the pronunciation of the consonant in the Tamil word for "seven" and the choice of word for "zero" varied significantly ($p < 0.01$, $N = 385$; $p < 0.01$, $N = 379$) by geography.

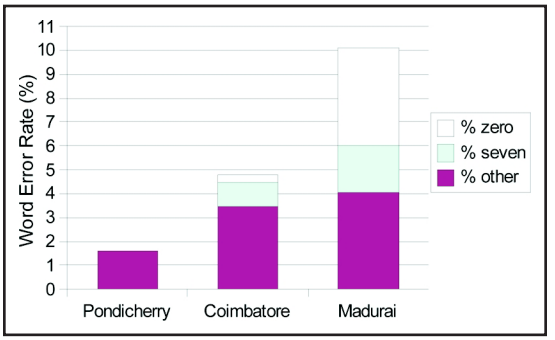

Figure 10. Word error rate for digits by site. Errors for the words "zero" and "seven" are explicitly indicated. The recognizer was trained on data from Pondicherry (2004) speakers.

Age, gender, and education level were not predictive factors in the phonetic and lexical variations.

To evaluate the influence of phonetic and lexical variations on a small vocabulary recognizer, we again trained the recognizer on the speech of the 22 Pondicherry (2004) speakers. Then, we tested the recognizer's performance on speakers from all three districts in our study (Figure 10). Digits spoken by Pondicherry (2005) speakers were recognized at less than 2% error. Coimbatore and Madurai speakers caused significantly higher word error rates ($p < 0.01$, $N = 3{,}168$). These results clearly show that a spoken dialog system should be trained on speech collected from people who are potential users of the fielded system to ensure there are no huge variations in dialect and choice of vocabulary between training speech and the field data.

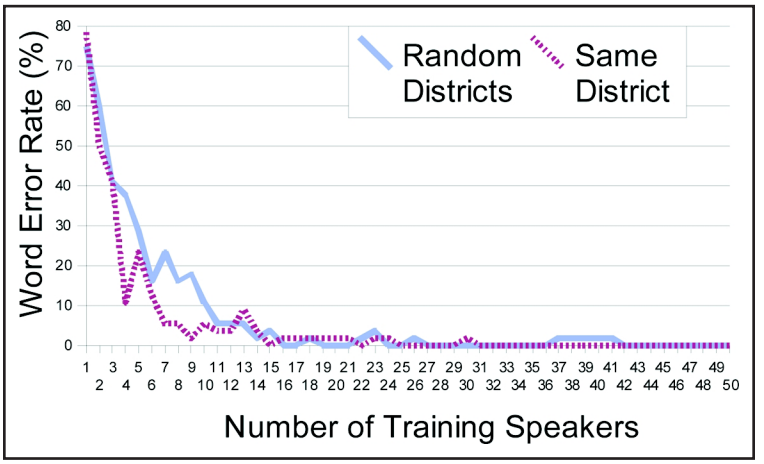

Figure 11. Word error rate by amount of training data.

Experiment 3: Available Training Data

Recognition performance depends largely on the quantity of available training data. Given that linguistic resources are limited in developing regions and that data collection is challenging and labor intensive, we ran simulations to determine the least amount of data needed to achieve acceptable error rates for the operation of an SDS.

For each simulation, one speaker was randomly selected. His speech was set aside as the input speech. First, the test speech was decoded by a recognizer trained only on the speech of a second speaker. The resulting word error rate was approximately 80% (Figure 11). Next the recognizer was trained on the speech of two speakers, three speakers, and so on. Word error rates dropped with the addition of more training data. We replicated the experiment under two conditions: first, speakers were added randomly from all districts; second, speakers from the test speaker's district were added first.

The results show that training a simple whole-word recognizer on the speech of approximately 15 speakers results in 2% error rates or less, which were found to be acceptable rates of error in parallel SDS user studies (Plauché & Prabaker, 2006). When fewer than 15 speakers are available for training, recognition for a given speaker is slightly better if trained on speakers from the same district.

ASR Design Considerations

The results from these ASR experiments confirm the basic principles of ASR (Table 1). Errors decrease with simple tasks, with matching input and training data, and with more training data. The specific trials shown here can inform our design for an SDS for developing regions. First, to achieve optimal usability with limited-resource ASR, our SDS design should limit the complexity of the ASR task to approximately 10 words or fewer at any given dialog node.

Our main finding in recording the speech of rural villagers of Tamil Nadu is that traditional data collection techniques favor literate speakers and that villagers only 300 kilometers apart use dif-

ferent words and different pronunciations for everyday concepts that are difficult to predict in advance. A speech recognizer designed to recognize the word choice and pronunciation of users in one village will likely fail for villagers a few hundred kilometers away. Likewise, the relevance of information varies from region to region (e.g., tide and sea conditions on the coast, rainwater collection in dry regions), necessitating a different set of command words in each case.

Our second proposal is the integration of speech collection into the SDS design. By recording the speech of people while they use the SDS, we can ensure that input and training data match not only in dialect, but also in other nonlinguistic factors, such as channel and room acoustics. In addition, the time-consuming, artificial step of collecting training data by recording disjointed instances of speech, which is particularly ill suited to an illiterate population, would be unnecessary. By integrating data collection into the SDS, the needs of the user (gaining access to relevant information) and the needs of the system (gathering instances of speech to enable recognition) are simultaneously met.

Finally, the ASR for each village could achieve adequate accuracy for SDS usability by cheaply and quickly initializing the models with the speech of only 15 speakers. In a later section, we will discuss language adaptation techniques that can further improve ASR performance when limited training data are available by (semi-) automatically incorporating user speech into acoustic models.

UI Design

The UI component is as important to the success of an SDS as ASR accuracy and application content. The human-computer interaction community offers these two guidelines for UI design of interactive systems (Del Galdo & Neilsen, 1996): 1. Let users feel in charge, not at the mercy of the system. 2. Spare users as much effort as possible.

An appropriate and effective user interface is one that suits the task to be accomplished. According to Lansdale and Ormerod (1994), question and answer interfaces work well when the user need only provide a small amount of information (cash machine). Repetitive tasks and tasks in which the user must provide a large amount of information before a system action can take place, are best served by form-filling tasks (calendars, travel). One strength of form-filling tasks is that they are compatible with paper-based forms (health surveys, land deed requests). Menus allow the user to choose from a set of options that need not be known in advance (information retrieval). While it is often assumed that certain dialog styles are more or less suited to novice users, it is the nature of the task that dictates appropriateness of dialog style rather than the level of expertise of the user. A spoken dialog system (SDS), which would allow users to access information by voice, either over the phone lines or at a kiosk, could play a role in overcoming current barriers of cost and literacy faced by traditional UI devices in developing regions. Speech-driven UIs are less expensive than display-based UI solutions and more accessible than text-based UI solutions. Most people in developing regions have never used a computer and generally feel uncomfortable using it for the first time for fear of breaking an expensive machine. Previous user studies (Parikh et al., 2003; Medhi, Sagar, & Toyama, 2006) in southern India found that voice feedback in the local language greatly helps with user interest and comfort.

Audio output that enhances a graphical interface or powers a telephony system can be created from speech synthesis or pre-recorded audio files. Synthesized speech may have poor pronunciation, but it requires a lesser amount of memory and offers an unlimited vocabulary. Speech synthesis is readily available for a handful of languages, using the open source, Festival system (Black, Taylor & Caley, 1999), but each new language voice requires months to develop (Dutoit et al., 1996). In our project, we relied on prerecorded speech from a native speaker for all audio output. The following sections describe some of the factors that influence the design of speech based UI in developing regions.

Literacy

Literacy is usually used to describe an individual's competency at the tasks of reading and writing, or her exposure to formal schooling. It is important to note that definitions of literacy only apply to people who live within a literate society. In traditional, oral societies, men and women of considerable learning, wisdom, and understanding, such as priests and traditional healers, transmit cultural and societal history through oral methods, though these individuals would be considered non-literate by most definitions of literacy.

In oral communities, information is primarily disseminated by word of mouth. Literacy increases access to information over wider distances in space and time. Of the estimated 880 million adults who are not literate, two-thirds are women and two-thirds live in India (Lievesley & Motivans, 2000), where health, nutrition, and earning potential positively correlate with literacy (Psacharopoulus, 1994; Census of Tamil Nadu, 2001; Borooah, 2004). In rural Tamil Nadu, illiteracy rates can be as high as 50% for men and 80% for women (Census of India, 2001).

Unschooled adults primarily rely on empirical, situational reasoning rather than abstract, categorical reasoning (Scribner, 1977). Design features considered standard or intuitive in traditional user interface literature, such as hierarchical browsing, and icons that represent concepts are found to present a challenge to individuals with no formal schooling. Researchers agree on the following requirements for user interface designs that accommodate unschooled individuals (Deo, et al. 2004; Parikh, Kaushik, & Chavan, 2003; Medhi, Sagar, & Toyama, 2006; Plauché & Prabaker, 2006):

- Ease of learning, ease of remembrance
- No textual requirements
- Graphics (and possibly speech in local language)
- Support for internationalization
- Accommodates localization
- Simple, easy to use, tolerant of errors
- Accurate content
- Robust in (potentially distracting) public spaces

Localization

For each new language and culture, the following UI design elements are subject to change: fonts, color, currency, abbreviations, dates, register, icons, concepts of time and space, value systems, behavioral systems. Traditional approaches to accessing models of culture include questionnaires, storyboards, and walkthroughs with a large sample of potential users at each stage of UI development (Schneiderman, 1992; Delgado & Araki, 2005). These user study techniques, however, present difficulties for unschooled, marginalized populations because of their daily requirements and ambient infrastructure (Huenerfauth, 2002; Brewer et al., 2006; Plauché et al., 2006). This might account for the relatively few publications reporting user studies in developing regions, despite the growing interest in researchers and developers of technology for rural populations. We predict that successful UI design for predominantly oral communities will build on existing means of information transfer and existing linguistic and cultural expertise by enabling community authorship of content.

Field Study 2: OpenSesame SDS

We developed OpenSesame, an SDS template for creating multi-modal spoken dialog systems for developing regions. We worked collaboratively with agricultural and community experts of MSSRF staff to port one unit (Banana Crop) of the text-based Valam website to the interactive OpenSesame application. User studies for the Banana Crop SDS were conducted using live speech recognition in Dindigul district, Tamil Nadu. The audio input recorded during user interactions with OpenSesame SDS served to simulate integrated data collection and ASR adaptation techniques, as discussed in the following sections.

Multimodal Prototype

The OpenSesame SDS runs on a multimodal prototype that allows both speech and touch input. The output includes graphics, a small amount of text, and prerecorded audio files. We constructed a modifiable flex button system by soldering the "reset" buttons from used computers to the function keys of a dedicated keyboard (Figure 12). The low-

Figure 12. Multimodal prototype accepts both touch and speech input.

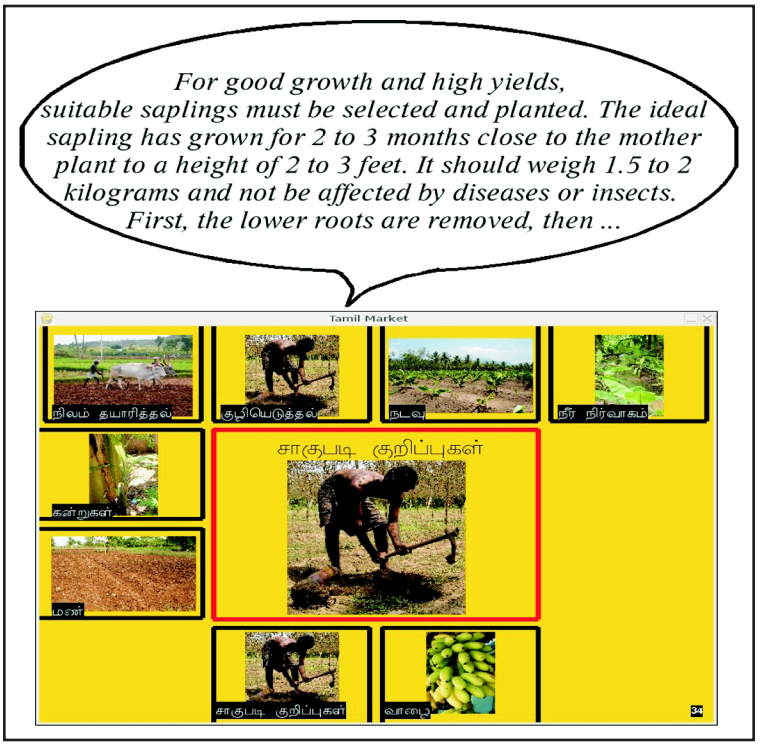

Figure 13. Screen shot of Banana Crop Application. The center square correlates to audio output, the smaller squares indicate available command options, also accessible via button panels.

cost equivalent of a touch screen (similar to ATM cash distribution machines), allows the SDS to function well in noisy environments where speech recognition fails by incorporating an additional input modality (touch). Construction of the prototype is transparent, cheap, and easy to construct from locally available materials.

Banana Crop SDS

Researchers and MSSRF staff collaborated to create an interactive version of one unit (Banana Crop) of the Valam website, using the OpenSesame SDS template. Banana Crop SDS adhered to the design guidelines for UIs previously described and was completed in less than 3 weeks. Our rapid, collaborative process involved identifying appropriate content, verifying the accuracy of the text version, gathering digital pictures, recording the speech output, and synchronizing all elements. MSSRF staff used their expertise and connections with local agricultural experts, universities, farmers, and merchants to locate relevant sites and stage demonstrations of recommended techniques. Their expertise in recommended agricultural practices informed the portrayal of content. For example, farmers identify banana varieties primarily by their fruit, not by the tree. Photos were prepared accordingly. The researchers provided design and scheduling recommendations based on the limitations and strengths of the technology. Synchronizing the images and the audio output was the most time consuming part of development, which led us to later develop a graphical user interface (GUI) editor for easy modification of OpenSesame applications.

Twenty-eight acoustically dissimilar and locally appropriate vocabulary words were selected to correspond to the Valam website subheadings (*Soil Preparation, Varieties,* etc.). The menu system was only three levels deep and presented no more than eight options at time. The system was highly redundant, explicitly listing options at every screen and disseminating information in the form of an audio slide show in Tamil when no input was provided. The result is an interactive dialog system that educates the user through digital photographs and narrative in Tamil on the recommended practices for growing, harvesting, and propagating a banana crop according to local conditions and needs (Figure 13).

Banana Crop ASR

The recognizer for the SDS must recognize multiple speakers and be robust to noisy conditions under conditions of limited linguistic data. The recognizer that powers Banana Crop SDS was trained on the transcribed Tamil speech recordings described in the previous section (Field Study 1). The speech recognizer is built using the hidden Markov model toolkit (HTK) (Young, 1996). A pronunciation dictionary was prepared by listing each vocabulary word along with its phonemic representation. Training models at the sub word level (e.g., phones and

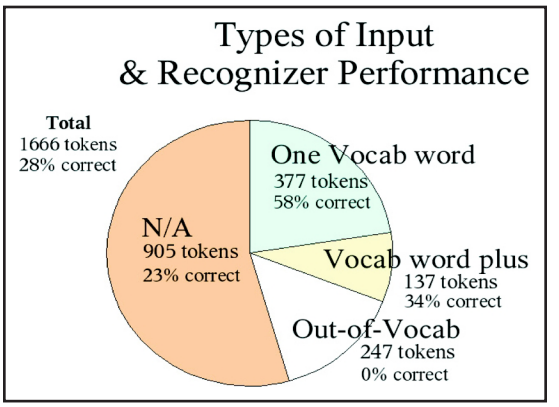

Figure 14. Categories of input from all six sessions. The percentages shown do not total 100, as they refer to the recognition results within each category, not the portion of all data the category represents. One Vocab word is referred to as the SDS Tamil 2006 database in later sections.

triphones) allow a recognizer to accommodate new words and phonetic contextual variations.

A test database was prepared by recording five MSSRF staff members saying each vocabulary word three times each to evaluate the ASR. When we trained our recognizer on monophone models, recognition on the test database yielded 73% accuracy. Triphone models (single Gaussian) performed at 97% accuracy. For subsequent user studies, our recognizer used triphone models and state-based parameter tying for robust estimation.

Field Evaluations

The Banana Crop SDS was evaluated by rural villagers in three different conditions across six different sites (Table 3). Approximately 50 people (roughly equal women and men) actively navigated the system using either touch or speech input. An estimated additional 200 people were onlookers who offered feedback based on that role. The participant's audio commands to the system were recorded during use. Sessions with each person were generally short, involved very little training, and invited informal feedback. In particular, people were asked to comment on the content, how easy the touch or voice input was to learn, and any preferences between the two modalities. We did not attempt a formal user study of the SDS. Our goal was to use the SDS to record speech during user interactions in order to design ASR adaptation techniques that would optimize performance by gradually integrating user speech into existing models.

ASR Results

The overall categories of input recorded across all sessions are shown in Figure 14. The majority of input was hand-labeled as "N/A," or "Not Applicable." This category includes sound files which are either empty, contain no speech, or contain irrelevant background speech. The recognizer correctly identified 23% of these tokens as "silence." Approximately 15% of all input were utterances directed at the application but not included in the recognizer's restricted vocabulary (out-of-vocabulary). The recognizer did not include a model for out-of-vocabulary input, so recognition performance on this set was 0%. Input that contained a vocabulary word, either alone (one vocab word) or with other input (vocab word plus), represented less than a third of all input data, and was recognized at rates of 58% and 34%, respectively.

Recognition performance on isolated vocabulary words was much worse for speech recorded during SDS interactions (58% accuracy) than for the speech recorded from MSSRF staff as they read words aloud in a quiet office (97% accuracy). Although ASR is known to degrade in noisy environments, the speech from MSSRF staff did not vary significantly from SDS sessions in signal to noise ratio, which was overall remarkably good (~20dB). The degradation is more likely due to the dissimilarity in speaking style between reading aloud and issuing commands to a machine.

Further investigation into recognition performance by site was conducted only on input comprised of a vocabulary word either alone (one vocab word) or with other speech (vocab word plus) (Figure 15). Performance does appear to be subject to social and environmental factors, as the highest rate of performance is found in the Sempatti session, a controlled user study with all literate subjects. The lowest performance occurred in S. Kanur, a farmer focus group in a much more distracting setting: a schoolroom with approximately 100 people and 2 onlookers for every participant.

Although overall recognition was poor, participants reported that the interface is easy to use. The most educated participants often commented that the system would be "good for people who cannot read." We noted that the least educated partici-

Table 3. Recording Conditions

Conditions	Users	Site Description
Controlled user study	3 men (literate)	Sempatti VRC: • One user at a time • Group feedback • 30 min. sessions • Speech only
	8 women 5 men (literacy varied)	Panzampatti VKC: • One user at a time • Individual feedback • 10–20 min. sessions • Speech and touch
Farmer focus group	15 women 20 men (literacy varied)	S.Kanur: • Group use • Group feedback • 5 min. sessions • Speech and touch
	10 women 20 men (literacy varied)	Gandhigram: • Group use • Group feedback • 5 min. sessions • Speech and touch
Village outreach	5 men (literacy varied)	Athoor: • One user at a time • Group feedback • 10 min. sessions • Speech only
	8 men 4 women (literacy varied)	P.Kottai: • One user at a time • Group feedback • 10-min. sessions • Speech only

pants preferred to listen to the system for several minutes before speaking to it. When prompted explicitly, some subjects reported preferring the touch screen as a means of input, others preferred speech. Many corrections and suggestions were offered for Banana Crop SDS, especially the addition of more crops to the system.

The three recording conditions (Table 3) were adopted out of flexibility to the available infrastructure, which ranged from a dedicated room in a village center (controlled user study) to a mat outside a home (village outreach). We tried to balance controlled user studies with existing methods of community information flow used by MSSRF (farmer focus groups). When MSSRF staff played a large role in the evaluative sessions, we observed lively and informed debates about recommended agricultural practices. MSSRF staff heard feedback from farmers who shared their successes and failures with current practices and explained what services and materials were provided at the nearby community centers. The farmer focus groups enabled us to observe the advantages of audio enabled software for multiple user settings, which were not apparent in our controlled user study conditions.

ASR Adaptation

So far, we have presented design and technology considerations for speech interfaces that meet criteria for equitable access, in particular for users in developing regions. Our belief is that speech interface solutions, especially those that can be easily modified by local experts, can allow oral populations access to digital, local resources. The technology that powers such a speech interface must also be easily customizable to new languages and dialects. Here, we introduce ASR adaptation, a technique for automatically or semiautomatically optimizing a recognizer by gradually integrating new, untranscribed data into the models for speech. Small vocabulary speech recognizers that are initialized with available data then tuned to user speech input with adaptation techniques can scale to new domains and new dialects more quickly and more affordably than large vocabulary, continuous speech systems.

Cross-Language Adaptation

Only a handful of speech technology efforts (Nayeemulla Khan & Yegnarayana, 2001; Saraswathi & Geetha 2004; Udhyakumar, Swaminathan, & Romakrishnan, 2004) have been dedicated to Tamil, which is spoken by more than 60 million people in

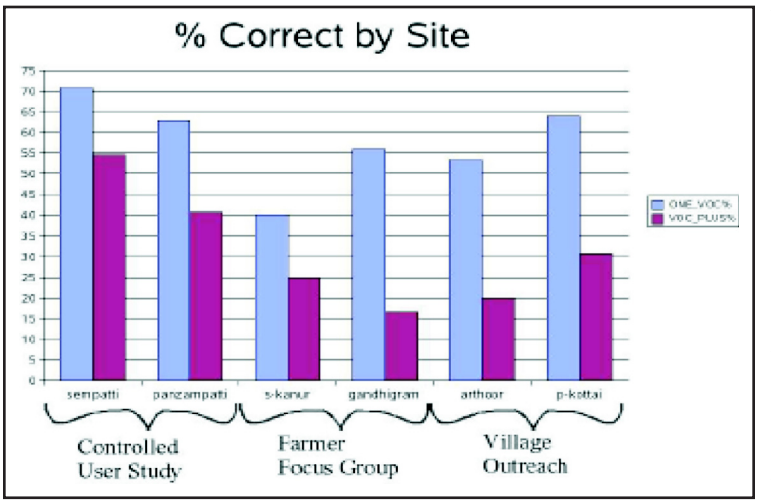

Figure 15. Recognition performance by site.

Tamil Nadu, Sri Lanka, and throughout the world (Comrie 1987). When annotated corpora for a given language are limited or unavailable, as is the case for most languages spoken in developing regions, a recognizer trained on transcribed data from one or more (source) languages, can be used to recognize speech in the new (target) language. This technique, called *cross-language transfer,* yields the best results when the source and target languages are linguistically similar and when the recognizer is trained on multiple languages.

Language adaptation is a technique in which the recognizer is trained on a large source language corpus and then the acoustic models are adapted to a very limited amount of target language data. Performance correlates to the amount of data available in the target language and the number of different speakers used for training. During *bootstrapping,* acoustic models are initialized from a small amount of transcribed source data. The speech recognition system is then iteratively rebuilt, using increasing amounts of training data and adaptation (Schultz & Waibel, 1998; Kumar & Wei, 2003; Udhyakumar et al., 2004).

Table 4. Language Adaptation Techniques and Data Conditions[1]

Adaptation Technique	Availability of Data
Cross-language Transfer	No data
Language Adaptation	Very limited data
Bootstrapping	Large amounts of data

[1] *See Waibel et al., 2000.*

Experiment 4: English-to-Tamil Adaptation

Based on the speech we collected in the field with Banana Crop SDS, we ran a series of recognizer experiments to determine how to optimize the small vocabulary recognizer to the speech of a particular community given no or limited Tamil training data. We simulated ASR performance under conditions of no available training data (cross-language transfer) and very limited training data (language adaptation) using the following databases: SDS Tamil 2006, Tamil 2006, Tamil 2005, and English TIMIT (Table 5).

In the field, the recognizer trained only on Tamil 2005 data recognized commands for Banana Crop SDS with 58.1% accuracy. We noted a substantial improvement (68.7%) with the addition of cepstral mean subtraction, an increase in model size from single Gaussian to 16 Gaussians, and the collapse of certain contrastive phonetic categories (long vs. short vowels) in the pronunciation dictionary (Figure 16). Simple noise robustness methods such as cepstral mean subtraction factor out environmental noise and generalize across tasks and speakers.

When an annotated corpus for a given language is unavailable, the options are to build one by collecting and transcribing speech, as we did in 2005, or to train a recognizer on an available corpus in another language. We first mapped the Tamil phonemes to English phonemes as closely as possible. Then, training and decoding were performed using HTK (Young, 1997). The acoustic models are trained first with a default flat initialization. Then triphone models are developed based on monophone HMMs and the recognizer decodes in a single pass using a simple, finite state grammar. Test results for the recognizer on speech input from the field (SDS Tamil 2006) show that the accuracy was significantly better when trained on a small amount of same language data than when trained on a greater amount of mismatched data. A Tamil SDS powered by a recognizer trained on English speech would only predict the correct word 30% of the time.

Table 5. English and Tamil Data Sets

Data Set	Size	Dictionary Size	Description
SDS Tamil (2006)	Very small (377 words)	Very small (28 words)	Agricultural words spoken by villagers retrieving information from Banana Crop SDS indoors and out in Dindigul district
Tamil (2006)	Very small (170 words)	Very small (28 words)	Same agricultural words read out loud by MSSRF staff in a fairly quiet office in Dindigul district
Tamil (2005)	Small (10K words)	Very small (50 words)	Digits and verbs read or guessed out loud by speakers of all literacy levels indoors and out in three districts
English (TIMIT)	Medium (50K words)	Medium (6K words)	Phonetically balanced sentences read out loud in a quiet laboratory setting

Finally, we initialized the recognizers on either English or Tamil, as described above, and then adapted the recognizer to the very small database of Tamil speech collected from five volunteers from the MSSRF staff (Tamil, 2006) using maximum likelihood linear regression (Young, 1997). The Tamil 2006 database is an available, yet very limited, language corpus that was rapidly collected and annotated (approximately 1 hour of nonexpert time). Adaptation to Tamil 2006 improves performance for both the recognizer trained on English and the recognizer trained on Tamil. It is interesting that the results are comparable (82.2% and 80.4%, respectively). There is very little to be gained by collecting and annotating a corpus like Tamil 2005, which took an estimated 100 hours of expert time, when adapting an English-trained system to a very small, easily prepared data set like Tamil 2006 yields similar results. This technique is supervised, as the adaptation data is manually transcribed before adaptation.

ASR adaptation can overcome the high costs of recording and annotating a large training corpus. Adaptation can also be *unsupervised* when a recognizer is automatically improved by gradually integrating new, untranscribed speech into intialized acoustic models (Kemp & Waibel, 1999; Lakshmi & Murthy, 2006). A confidence measure is used to rank the data; those with the highest scores are selected for integration. Similar adaptation efforts have sought to include, yet minimize, human participation for training acoustic models. In *supervised adaptation,* the utterances with the lowest confidence scores are deemed to be the most informative. They are automatically selected for hand transcription and integration into the training data, resulting in a reduction of the amount of labeled data needed by as much as 75% (Lamel, Gauvain, & Adda, 2000; Riccardi & Hakkani-Tür, 2003). Further exploration of (semi-) automatic adaptation techniques will surely result in robust, rapid development ASR for limited-resource environments.

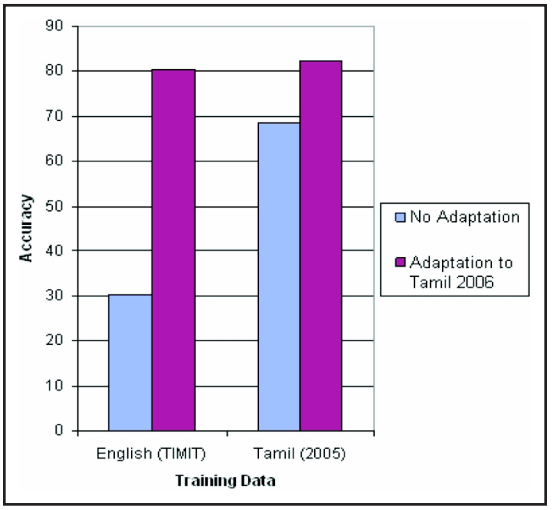

Figure 16. Accuracy of ASR on SDS Tamil 2006.

Future Plans and Conclusions

This article reviews literature on the language, access, and information requirements likely to be found in primarily oral, limited-resource environments. Speech technologies and techniques that are

small, scalable, and easy to modify and update by local stakeholders in community development can be constructed to deliver accurate, locally relevant information to individuals regardless of their literacy level. Integrated data collection and language adaptation are found to be useful techniques for collecting linguistic resources according to both user and system needs.

In future studies, we plan to determine the smallest amount of adaptation data required to reach adequate levels of ASR accuracy. We would also like to explore how speech/no speech detectors and out-of-vocabulary models could play a role in a robust, adaptive SDS/ASR system. Recall that 75% of SDS input consisted of unusable data. We envision an SDS that is initialized with a large amount of available data perhaps from a different language, then as it is used in a village or community, participants' speech is recorded, prefiltered, and gradually integrated (automatically or semiautomatically) to adapt to the dialect and speaking style. We hope to see further work in simple, affordable designs for speech synthesis and UI, especially for text-free browsing and searching across libraries of audio and digital media. ■

Acknowledgments

The authors thank their volunteers, families, hosts and partner organizations in Tamil Nadu, as well as colleagues and reviewers who contributed their encouragement and insights. This material is based upon work supported by the National Science Foundation under grant number 0326582. Any opinions, findings and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Black, A., Taylor, P., & Caley, R. (1999). The Festival speech synthesis system. Available at http://festvox.org/festival

Blattman, C., Roman, R., & Jensen, R. (2003). Assessing the need and potential of community networking for development in rural India. *The Information Society 19*(5): 349–364.

Borooah, V. K. (2004). Gender Bias among Children in India in Their Diet and Immunization against Disease. *Social Science and Medicine 58:* 1719–1731.

Braund, P., & Schwittay, A. (2006). The missing piece: Human-driven design and research in ICT and development. In *Proceedings of ICTD.* Berkeley, CA: ICTD.

Brewer, E., Demmer, M., Ho, M., et al. (2006). The Challenges of Technology Research for Developing Regions. *IEEE Pervasive, 5*(2): 15–23.

Castells, M. (1997). *The power of identity: The information age—Economy, society and culture*. Cambridge, MA: Blackwell Publishers.

Census of India (2001). *Census of India 2001,* Paper 1 of 2001. Office of the Registrar General, India.

Census of Tamil Nadu (2001). *Primary Census of Tamil Nadu.* Directorate of Census Operations, Tamil Nadu.

Chisenga, J. (1999). Global Information infrastructure and the question of African content. In *Proceedings of the 65th IFLA Council and General Conference.* Bangkok, Thailand.

Comrie, B. (1987). *The World's Major Languages.* New York: Oxford University Press.

Delgado, R. L.-C., & Araki, M. (2005). *Spoken, Multilingual and Multimodal Dialogue Systems; Development and Assessment.* Northwest Sussex, England: John Wiley & Sons.

Del Galdo, E. M., & Nielsen, J. (1996). *International User Interfaces.* New York: John Wiley & Sons.

Deo, S., Nichols, D.M., Cunningham, S.J., et al. (2004). Digital Library Access for Illiterate Users. In *Proceedings of the 2004 International Research Conference on Innovations in Information Technology,* Dubai.

Donner, J. (2004). Microentrepreneurs and Mobiles: An Exploration of the Uses of Mobile Phones by Small Business Owners in Rwanda. *ITID, 2*(1): 1–21.

Dutoit, T., Pagel, V., Pierret, N., et al. (1996). In Proceedings of ICSLP'96, 3: 1393–1396. Philadelphia.

Eggleston, K., Jensen, R., & Zeckhauser, R. (2002). Information and communication technologies, markets and economic development. In G. Kirkman and J. Sachs (Eds.). *Global Information Technology Report 2001–2002: Readiness for the*

Networked World, Oxford: Oxford University Press. 62–75.

Ethnologue (2006). Available online at http://www.ethnologue.com

Gordo, B. (2003). Overcoming digital deprivation. *IT & Society,* 1(5): 166–180.

Heffernan, C. (2006). Fighting poverty with knowledge: The livestock guru. *Report for DFID's Livestock Production Program.* DFID, London.

Hornik, R. (1988). *Development Communication: Information, Agriculture, and Nutrition in the Third World.* New York: Longman.

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

Jelinek, F. (1996). Five speculations (and a divertimento) on the themes of H. Bourlard, H. Hermansky, and N. Morgan. *Speech Communication,* 18:242–246.

Kemp, T., & Waibel, A. (1999). Unsupervised Training of a Speech Recognizer Using TV Broadcasts: Recent Experiments. In *Proceedings of EUROSPEECH'99,* Budapest, Hungary: 2725–2728.

Kumar, C. S., & Wei, F. S. (2003). A bilingual speech recognition system for English and Tamil. In *Proceedings of the Fourth Information, Communications and Signal Processing and Fourth Pacific Rim Conference on Multimedia,* Singapore, December.

Kumar, R. (2004). eChoupals: A study on the financial sustainability of village Internet centers in rural Madhya Pradesh. *Information Technology and International Development, 2*(1):45–73.

Lakshmi, A., & Murthy, H. A. (2006). A syllable-based speech recognizer for Tamil. In *Proceedings of INTERSPEECH 2006—ICSLP,* Pittsburgh, 1878–1881.

Lamel, L., Gauvain, J.-L., & Adda, G. (2000). Lightly supervised acoustic model training. In *Proceedings of ISCA ITRW ASR2000,* Paris, September, 150–154.

Lansdale, M. W., & Ormerod, T. C. (1994). *Understanding interfaces: A handbook of human-computer dialogue.* San Diego, CA: Academic Press Professional.

Lievesley, D., & Motivans, A. (2000). *Examining the Notion of Literacy in a Rapidly Changing World.* Paris: UNESCO Institute for Statistics.

Medhi, I., Sagar, A., & Toyama, K. (2006). Text-free user interfaces for illiterate and semi-literate users. In *Proceedings of ICTD.* Berkeley, CA: ICTD.

Moore, R. K. (2003). A comparison of the data requirements of automatic speech recognition and human listeners. In *Proceedings of EUROSPEECH'03,* Geneva, Switzerland: 2582–2584.

Nayeemulla Khan, A., & Yegnanarayana, B. (2001). Development of a speech recognition system for Tamil for restricted small tasks. *Proceedings of National Conference on Communication.* Kanpur, India.

Noronha, F. (2002). Indian language solutions for GNU/Linux. *Linux Journal, 2002.* (103): 4.

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

Plauché, M., & Prabaker, M. (2006). Tamil market: A spoken dialog system for rural India. *Working Papers in Computer-Human Interfaces (CHI),* April.

Plauché, M., Wooters, C., Ramachandran, D., et al. (2006). Speech recognition for illiterate access to information and technology. In *Proceedings of ICTD.* Berkeley, CA: ICTD.

Prahalad, C. K., & Hammond, A. (2002). Serving the poor profitably. *Harvard Business Review, 80:* 48–57.

Psacharopoulus, G. (1994). Returns to investment in education: A global update. *World Development, 22* (9):1325–1343.

Raman, T. V. (1997). *Auditory user interfaces: Towards the speaking computer.* Norwell, MA: Kluwer Academic Publishers.

Riccardi, G., & Hakkani-Tür, D. (2003). Active and unsupervised learning for automatic speech recognition. In *Proceedings of EUROSPEECH'03,* Geneva, Switzerland.

Saraswathi, S., & Geetha, T. V. (2004). Building language models for Tamil speech recognition system. *Proceedings of AACC.* Kathmandu, Nepal.

Schneiderman, B. (1992). *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* Boston: Addison-Wesley.

Schultz, T., & Waibel, A. (1998). Multilingual and crosslingual speech recognition. In *Proceedings DARPA Workshop on Broadcast News Transcription and Understanding,* 259–262.

Scribner, S. (1977). Modes of thinking and ways of speaking: Culture and logic reconsidered. In P. N. Johnson-Laird & P. C. Wason (Eds.). *Thinking: Readings in Cognitive Science.* Cambridge: Cambridge University Press.

Soola, E. O. (1988). Agricultural communication and the African non-literate farmer: The Nigerian experience. *Africa Media Review, 2*(3):75–91.

Summer Institute of Linguistics (SIL) (2005). Available online at *http://www.sil.org/literacy/LitFacts.htm*

Sustainable Access in Rural India (SARI) (2005). Available online at *http://www.tenet.res.in/sari*

Udhyakumar, N., Swaminathan, R., & Ramakrishnan, S. K. (2004). Multilingual speech recognition for information retrieval in Indian context. In *Proceedings from the Student Research Workshop, HLT/NAACL,* Boston, MA.

Waibel, A., Guetner, P., Mayfield Tomokiyo, L. et al. (2000). Multilinguality in speech and spoken language systems. In *Proceedings of IEEE, 88*(8):1297–1313.

Young, S. (1996). A review of large-vocabulary continuous speech recognition. *IEEE Signal Processing Magazine,* 45–57.

Young, S. (1997). *The HTK BOOK. Version 3.2.* Cambridge University, U.K.