

How to Build a Spoken Dialog System with Limited (or no) Language Resources

Madelaine Plauché*, Özgür Çetin*, Udhaykumar Nallasamy†

*International Computer Science Institute, 1947 Center St., Suite 600, Berkeley, CA 94703

†Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213
{plauche, oetin}@icsi.berkeley.edu, udhay@cmu.edu

Abstract

This paper evaluates low cost, rapidly deployable speech technologies for new languages as a means to improve equitable, affordable access to information technology (IT). We describe our field work in Tamil Nadu, recording speech from within a multi-modal (speech and touch) dialog system. The performance of a speech recognizer built using cross-language adaptation is evaluated on our field recordings, with implications for an iterative, learning approach for spoken dialog systems in conditions of limited language resources.

1 Introduction

The goal of this paper is to provide practical guidelines for the development of low cost spoken dialog system (SDS) in new languages and in contexts of limited resources and oral communities. We also report on our experience in designing a rapidly deployable and easily modifiable SDS in partnership with a local NGO in Tamil Nadu through a participatory design process.

The paper is organized as follows. In Section 2, criteria relevant to speech technologies and likely to be found in developing regions (literacy, local content, and local language) are described. Section 3 provides an overview of speech recognition and user interface design, especially as they pertain to literacy and limited resources. Section 4 describes our experiment in Tamil Nadu in which a prototype for a simple, scalable SDS was used to conduct recordings of participants which would allow the recognizer that powered it to gradually adapt to the participants' speech. Finally, in Section 5, results from ASR experiments are presented in which conditions of limited resources are simulated and cross-language transfer and language adaptation techniques are used to make the most of available data. We conclude in Section 6, with major findings and recommendations for further study in simple, scalable spoken dialog systems.

2 Equitable Access

The “digital divide” is the concern that access to computers, the Internet, and other information technologies (IT), despite

their potential to bring about massive social changes, are likely instead to entrench existing social patterns and hierarchies, especially relationships of power, authority, and wealth [Boas *et al.*, 2005]. For example, IT holds great promise for the physically disabled, due to its malleability, however development of technology to serve the disabled is not pursued to the extent that it could be [Johnson, 2001].

The inequity in power and wealth across nations (developing vs. developed), correlates with a lack of necessary IT infrastructure and software/hardware designs that fail in developing regions [Brewer *et al.*, 2006]. The main hurdle in the access of computing technologies has been the prohibitive cost of computing devices. Across India, for every 100 people, less than two PCs and three Internet access points are available [Fonseca, 2006]. Cell phones are affordable, however, and are extensively used throughout the developing world; often they are shared by multiple users of varying degrees of literacy [Donner, 2004]. Therefore, automated telephony services (an SDS) are attractive channels for information dissemination in developing regions as they depend on infrastructure that is more widely available, do not require literacy, and in principle, can run on any language or dialect, even those for which there is no written form.

2.1 Literacy

Literacy is usually used to describe an individual's competency at the tasks of reading and writing, or her exposure to formal schooling. A person may be considered illiterate because she is a recent migrant or mentally disabled [Deo *et al.*, 2004]. It is often mistakenly assumed that a dependent continuum exists from numeric literacy to print literacy, then to computer literacy. In fact, it is difficult to separate the ability to memorize arbitrary assignments of abstract symbols to sounds and meaning, from more salient factors such as amount and type of schooling, age, and life experiences.

Comparative cognitive studies show that formal schooling alters the functional organization of the adult brain [Castro-Caldas *et al.*, 1998; Petersson *et al.*, 2000]. Cross-cultural studies involving verbal logic problems, indicate that unschooled adults rely on *empirical, situational* reasoning (Table 1) rather than *abstract, categorical* reasoning [Scribner, 1977], which likely stems from the day

to day life of their subjects, rather than a lack of ability [Dias *et al.*, 2005].

Verbal Logic Problem

“All Kpelle men are rice farmers.
Mr Smith is not a rice farmer.
Is he a Kpelle man?”

Kpelle Farmer’s Response

“I don’t know the man in person. (...) If I know him in person, I can answer that question, but since I do not know him in person I cannot answer that question.”

Table 1. Example of situational reasoning [Scribner, 1977].

It is important to note that definitions of literacy only apply to people who cannot read and write and *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, keep alive cultural and societal history through oral methods, though these individuals would be considered non-literate by current definitions.

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 *et al.*, 2000], where health, nutrition, and earning potential positively correlates with literacy [Psacharopoulos, 1994; Tamil Nadu Census, 2001; Borooah, 2004]. Even in Sweden, where the highest overall literacy rate is enjoyed, one out of ten adults suffers from a severe literacy deficit in everyday life and work [Lievesley *et al.*, 2000].

The information needs of oral communities and non-literate people in developing regions are likely to be vast and varied, stemming in great part from rich traditional channels of communication, which might include family visits, marriage ceremonies, festivals, disputes, harvest, markets, schools, churches, and village squares.

2.2 Local content

Relevant, local content creation in local languages is a large concern for developing regions [Chisenga, 1999]. Attempts by local or national government or non-governmental 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, accessible forms of mass media, which can be effective at creating initial public awareness. However, mass media is less effective in influencing people to improve practices in health, agriculture, or education than traditional, *oral* methods and content that stem from within a community [Soola, 1988].

Today, public libraries are a potent source of information for the rural non-literate person and multimedia documents such as audio cassettes and digital footage do not require that the user be literate [Soola, 1988; Deo *et al.*, 2004]. In

our own work in southern India, we saw a wealth of short videos that had been prepared locally by universities and community organizations on recommended agricultural and health practices. In addition, video conferences were used by non-profits to allow farmer groups or self-help groups to share experiences and local innovations, and to offer recommendations to one another across districts of Tamil Nadu.

Information and Communication Technologies (ICT) offer the opportunity and infrastructure for publishing and distributing all types of information in the shortest possible time and at the lowest cost. Could speech and ICT play a role in guaranteeing equitable access to information, regardless of physical ability or formal education? We believe they can if both *content* (locally created graphical, audio, and video media) and *access* (searching and browsing in local language) are customizable at the community level.

2.3 Local Language

Depending on how one defines a language, over 6,000 languages are spoken in the world today, though that number is likely to drop by half in the expected lifetime of the authors [Crystal, 2000]. Language death is mainly due to pressures for indigenous groups to assimilate into dominant cultures; often this takes the form of institutionalized coercion or repression in educational or governmental policy. In India, there are two official languages (Hindi and English), 22 scheduled languages (including Tamil), and an estimated 400 more [Ethnologue, 2006].

Each of the world’s languages, uniquely groups a subset of all possible speech sounds into contrastive phones. Languages vary in segmentation; meaning may be represented with distinct words (e.g., English), or as suffixes and prefixes (e.g., Turkish, Tamil, Swahili). Syntactic (subject-object-verb vs. subject-verb-object) and prosodic structure (tone, stress) also vary across languages. Even within each language, factors such as age, geography, gender, education, socio-economic status, and task of both the speaker and the listener combine to produce an infinite number of dialects and registers. Speech technology that facilitates local language authoring at the community level, will bypass the challenges of this complexity by relying on native speaker expertise.

3 Speech Technologies

Successful speech-driven applications can accommodate a sight-impaired user’s access to computer output (screen) and a mobility-impaired user’s access to computer input (keyboard) [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. Speech-driven solutions to user interface are often suggested for developing regions, where a strong oral tradition remains and where costs must be low [Barnard *et al.*, 2003], though others [Jelinek, 1996] predict that they will likely not enhance the quality of life of those that rely on them. In any case, speech technologies must handle the challenges of multilingualism, dialectal and

cultural diversity, and the limitations of conveying speech versions of 2D information, such as graphs, or hierarchies.

The last 50 years of research in speech technologies have primarily (1) focused on a handful of languages, (2) assumed the availability of costly language resources, such as annotated corpora and pronunciation dictionaries, and (3) strived for performance that mimics a human’s capacity to speak and understand (e.g., large-vocabulary, continuous speech recognition, and expressive, “natural-sounding” speech synthesis). This section presents an overview of speech recognition and User Interface (UI) design with a focus instead on how current techniques and technologies might meet the criteria for equitable access outlined in the previous section: literacy, local media, and local language. Our belief is that small, easily modifiable systems that will scale to new domains and new dialects more quickly and more affordably than large vocabulary, continuous speech systems, can accommodate oral populations with access to digital, local resources.

3.1 Speech Recognition

Automatic speech recognition (ASR) is the process of algorithmically converting a speech signal (audio) into a sequence of words (text). This is usually achieved by training hidden markov models (HMMs) for phones, diphones, or triphones from training data, a hand-labelled speech corpus. 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 non-trivial task 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.

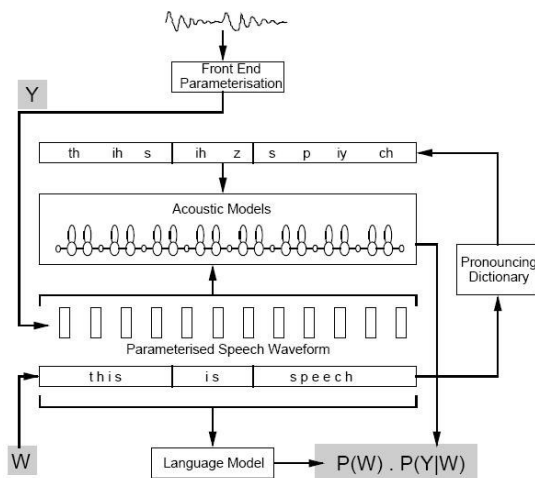


Figure 1. This diagram shows the computation of the

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 (taken from [Young, 1996]).

ASR systems are evaluated along several dimensions [Raman, 1997]:

- **Accuracy**– percentage of words correctly recognized
- **Reliability**– degradation under noisy conditions
- **Speaker Dependence**– for one or multiple speakers
- **Vocabulary Size**– words that can be recognized input can be a few utterances or several thousand words
- **Resources**– computation, memory, and speed

The basic principles of ASR performance are:

The more data, the better.

The more input matches training data, the better.

The simpler the task, the better.

For large vocabulary connected speech, the first principle has dominated the direction of automatic speech recognition research. Moore [2003] examines questions the wisdom of the credo *There’s no data like more data*, by estimating the amount of training data required to bring the performance of ASR up to that of a human listener. Given that word error rates drop *logarithmically* with increasing quantities of training data, Moore estimates that current systems would require 3 to 10 million hours of acoustic training data which is equivalent to between 4 and 70 human lifetimes of exposure to speech to achieve human levels of recognition. We should forgo the singular drive for more training data, which requires substantial cost, time, and expertise to collect, in favor of methods that better exploit information in existing data.

Adaptation Techniques

Adaptation refers to tuning the speech recognizer using target speech data obtained from the field to better match the training and test conditions. *Unsupervised adaptation* consists of training a recognizer on very little transcribed data and improving its acoustic models by gradually integrating new, un-transcribed speech [Kemp and Waibel, 1999; Lakshmi and 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. As in unsupervised adaptation, *lightly supervised* or *active learning* begins by training acoustic models on available, transcribed data. Then the target data is automatically recognized with a certain confidence score. The utterances are ranked by confidence and those with the *lowest* scores are deemed to be the most informative, so they are selected for hand transcription and added to the training data. These techniques can be combined with one another and result in reducing the amount of labeled data needed by as much as 75% [Lamel *et al.*, 2000; Riccardi and Hakkani-Tür, 2003].

ASR for Limited Resource Languages

The techniques explained in the previous section are used to adapt an ASR system from a (source) language, for which training resources (transcribed speech data) are available, to a (target) language for which there are limited or no annotated corpora. Conditions and recommended strategies include [Waibel *et al.*, 2000]:

Technique	Condition
Cross-language Transfer	no data
Language Adaptation	very limited data
Bootstrapping	large amounts of data

Cross-language transfer is used when no existing data is available for a language. It involves training the recognizer on one or more source languages. Linguistically similar models and multilingual models offer the best results.

Language adaptation is used to generate linguistic resources for a target language by initializing acoustic models on available source language data and adapting the models to the target language using a very limited amount of training data. Performance correlates to the amount of data available in the target language. Also, the number of different speakers used for training is found to increase performance more than the number of utterances.

During *bootstrapping*, acoustic models are initialized from a small amount of transcribed source data. The ASR system is then iteratively rebuilt, using increasing amounts of training data and adaptation [Schultz *et al.*, 1998; Udhayakumar *et al.*, 2004; Kumar and Wei, 2003].

A strategy that takes advantage of the substantial overlap of speech sounds across languages is to train acoustic models on IPA (International Phonetic Alphabet) representation [Schultz *et al.*, 1998]. Training acoustic models directly on graphemes (script characters) or automatically converting graphemes to phonemes omits the need to create a pronunciation dictionary by hand [Kanthank and Ney, 2003]. This method works best for languages with a close grapheme to phoneme relationship.

Another unique research effort, Dictionary Maker [Davel and Barnard, 2004] incorporates native speaker knowledge while minimizing the required human effort and expertise in the creation of a pronunciation dictionary for a new language. The system presents a native speaker with the script and audio representation of a 'best word' from a list of words to extrapolate the grapheme to phoneme rules. This approach requires simple speech synthesizer to generate the pronunciations, but it does not require linguistic expertise.

3.2 User Interface Design

The human-computer interaction community for UI design in interactive systems offers these two guidelines [Del Galdo and 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. *Question & 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). A 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) [Lansdale and Ormerod, 1994].

In menu systems, the options may form a hierarchical, tree-like structure. A balance between depth (the number of menus to be traversed before arriving at the required information) and breadth (the number of choices allowed per menu), must be found. 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.

When interfaces are designed for different languages and cultures, the following are subject to change: fonts, color, currency, abbreviations, dates, register, icons, concepts of time and space, value systems, behavioral systems. In order to access models of culture, questionnaires, storyboards, and walkthroughs with a large sample of potential users are recommended [Schneiderman, 1992; Delgado and Araki, 2005]. These techniques assume literacy and thus are ill suited for work in predominantly oral cultures. Time, location, available resources, and remuneration also challenge traditional UI design [Plauché *et al.*, 2006; Brewer *et al.*, 2006]. Accessing and deeply understanding the cultural model of each community is not necessary, however if the system is designed by members within the community or culture.

Researchers agree on the following guidelines for user interface designs that accommodate all literacy levels [Deo *et al.*, 2004; Plauché, 2006; Medhi, 2006]:

- ease of learning, ease of remembrance
- no textual requirements
- graphics (and possibly speech)
- support for internationalization
- accommodates localization
- simple, easy to use, tolerant of errors
- accurate content
- robust in (potentially distracting) public spaces

Additional recommendations include the use of a side-bar of available options, instead of a home button [Deo, 2004]. Design features considered standard or intuitive, such as hierarchical browsing, and icons that represent concepts present a challenge to individuals with no formal schooling.

We predict that designs that build on existing means of information transfer in oral communities will likely prove most successful.

4 Open Sesame: A Simple, Scalable SDS

In the following sections, we describe a template spoken dialog system (SDS) which can be run over the phone or on a kiosk (with accompanying graphics). The overall goal of the project is to encourage the development of low cost, affordable SDS that can be created and maintained by small communities, regardless of language and literacy.

4.1 Open sesame recordings in Tamil Nadu

In a previous field study (in 2005) conducted in rural Tamil Nadu, we found that speech-enabled applications powered by simple speech recognizers, could be quickly built with limited language resources. In addition, we found that traditional methods of SDS development (Wizard-of-Oz and speech collection) yielded different results for subjects of different levels of literacy [Brewer *et al.*, 2006]. The better the training data matches the input data, the better a recognizer's performance will be. Thus, we hypothesized that by initializing the recognizer with whatever data was available (same language or different language) and gradually adapting it to the *speech of people using the system*, we could quickly localize to specific dialects and domains. This required an SDS that contained relevant, accurate information. As we discussed in the first section, the information needs of each community vary widely, and are difficult to anticipate in advance.

The design of Open Sesame SDS can be easily and quickly localized to a specific domain and dialect. Local stakeholders in a rural district of Tamil Nadu provided expertise and connections to help develop content that was accurate and relevant to nearby communities. Villagers from the district were then recorded while navigating the application to retrieve information. Recordings from the system are categorized by input type and used to adapt a simple speech recognizer (Section 5).

Dindigul District

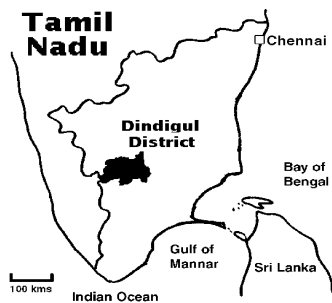


Figure 2. Political map of Tamil Nadu.

Our fieldwork was conducted in May 2006 at the VRC in Sempatti (Figure 3), in the Dindigul district of Tamil Nadu (Figure 2). The district has a mostly rural population of 2 million; only 60% of whom can read and write. The main languages spoken are Tamil, Telugu, and Kannada [Dindigul Statistics, 2004].

M.S. Swaminathan Research Foundation (MSSRF)

MSSRF is a non-profit network of village centers in rural Tamil Nadu that enable information exchange among experts and villagers in order to address the social and economic needs of the rural poor. Trained community members in villages across southern India operate village *knowledge* centers (VKCs) and regularly communicate the needs of their neighbors to village *resource* centers (VRCs) through weekly meetings, a user registry, and door-to-door surveys. The four VRCs, in turn, communicate needs to MSSRF headquarters.

The VRCs and MSSRF headquarters then work to provide materials (text, audio, video), additional training, video conferencing, and workshops to VKC operators and community members. Since its founding in 2004, the Sempatti VRC is responsible for 9 nearby VKCs (Figure 3), each one representing and serving the needs of 2 to 11 thousand people.



VRC Sempatti

VKC Panzampatti

Figure 3. MSSRF village resource and knowledge centers

Also, VRC staff meticulously document the crops grown in different regions, including varieties, planting techniques, soil properties, and fertilizers. One result of this network is a static, text-based website called *Valam* (Tamil word for “resources”) available in all VKCs. It is used by the trained community members to retrieve locally-relevant information on health, education, jobs, micro-credit, self-help groups, and agriculture (Figure 4). The VRC Sempatti helped us port one unit, Banana Crop, to the Open Sesame SDS template.



Figure 4. Valam website

Banana Crop SDS

The Banana Crop SDS adhered to the design guidelines for user interface design described in Section 3.2. The 28 command vocabulary options corresponded to the Valam website subheadings (*soil preparation, varieties, etc...*) and do not assume previous text or computer experience (*back, next, 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 when no input was provided. The SDS output consisted of pre-recorded (*canned*) speech by a native Tamil speaker (one of the authors). Input was a hand-made touch screen and a small vocabulary, Tamil ASR.

Researchers and MSSRF staff created a multi-modal SDS for one unit of their website in only three weeks. The process involved identifying appropriate content, verifying the text version, taking and 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 document local varieties and sites and stage demonstrations of recommended techniques. Their understanding of needs and practices informed graphical content that best illustrated the recommended practices. For example, farmers identify banana varieties primarily by their fruit, not by the tree. Photos were prepared accordingly. The researchers provided design and schedule recommendations based on the limitations and strengths of the technology.

For good growth and high yields, suitable saplings must be selected and planted. The ideal sapling has grown for 2 to 3 months close to the mother plant to a height of 2 to 3 feet. It should weigh 1.5 to 2 kilograms and not be affected by diseases or insects. First, the lower roots are removed, then ...

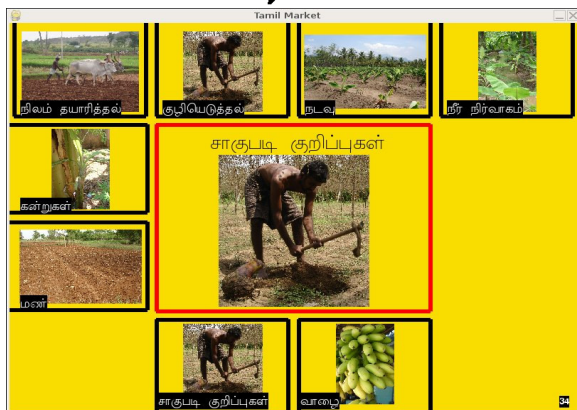


Figure 5. Screen shot of Banana Crop Application. Large square correlates to audio output, the smaller square indicate available options, also accessible via button panels.

At first, we tried to gather banana crop pictures from the Internet. The slow connection, variations in agricultural terminology, and a lack of images reflecting local varieties and conditions, proved this technique to be useful only for pictures of diseased plants and insects (five pictures in two days). Taking pictures locally, including staging and documenting cultivation techniques was much more effective (remaining 50 pictures in one day). Integrating the images and synchronizing the audio output was the most time consuming part of development.

4.2 Tamil Speech Recognizer

The Tamil speech recognizer that powers the Banana Crop SDS was trained on a database referred to as *Tamil 2005*. Tamil 2005 contains transcribed speech recordings of 80 speakers, collected across three districts and representing a range of literacy levels in 2005. Native Tamil speakers in offices, schools, and fields were shown flash cards or a number of fingers to elicit digits (*one, two, three, ... ten*).

The speech recognizer used triphone HMM models (single Gaussian) and state-based parameter tying for robust estimation. Decision tree based, state-tied, triphone models easily accommodate new words and contexts by traversing through the tree and synthesizing the triphones from a cluster of acoustically similar models.

Once the 28 vocabulary options for the Banana Crop SDS had been selected, we recorded five MSSRF staff members reading the words out loud three times each. The recordings constitute the test database, *Tamil 2006*. The triphone model was chosen to power the SDS based on its performance on the *Tamil 2006* database. Monophone models yielded 73% accuracy, while triphone models performed at 97%. Whole word models were not an option, as they do not accommodate new vocabulary words.

The recognizer for the SDS must recognize multiple speakers and be robust to noisy conditions, and it must do so under conditions of limited linguistic data. By reducing the task of the recognizer to recognizing one or two words at a time, we use the third basic principle of ASR to our advantage:

The simpler the task, the better.

SDS Speech Recordings

The Banana Crop SDS was evaluated by rural villagers in three different conditions across six different sites. 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 and gradually integrate speech for training, thereby pursuing performance according to the second principle: *The more input matches training data, the better.*




Conditions	Users	Site Description
Controlled User Study 	3 men (literate)	Sempatti VRC <ul style="list-style-type: none"> •one user at a time •group feedback •30 min. sessions •speech only
	8 women 5 men (literacy varied)	Panzampatti VKC <ul style="list-style-type: none"> •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 <ul style="list-style-type: none"> •group use •group feedback •5 min. sessions •speech and touch
	10 women 20 men (literacy varied)	Gandhigram <ul style="list-style-type: none"> •group use •group feedback •5 min. sessions •speech and touch
Village Outreach 	5 men (literacy varied)	Athoor <ul style="list-style-type: none"> •one user at a time •group feedback •10 min. sessions •speech only
	8 men 4 women (literacy varied)	P.Kottai <ul style="list-style-type: none"> •one user at a time •group feedback •10 min. sessions •speech only

Table 2. SDS Recording Conditions

Results

The overall categories of input recorded across all sessions are shown in Figure 6. 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-Vocab). 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.

Types of Input & Recognizer Performance

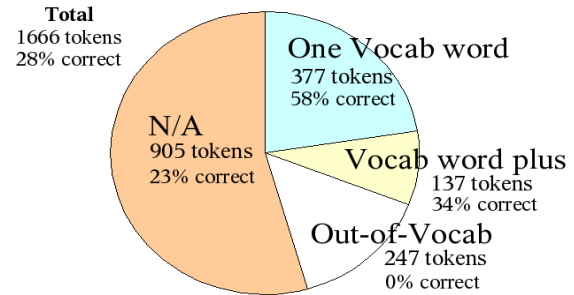


Figure. 6: Categories of input from all six sessions. The percentages shown do not total 100, as they refer to the recognition results within each category, not the portion of all data the category represents. One Vocab word is referred to as the *SDS Tamil 2006* database in later sections.

Recognition performance on isolated vocabulary words was much worse for speech recorded during SDS interactions than for the speech recorded from MSSRF staff as they read words aloud in a quiet office (58% vs. 97%). Although ASR is known to degrade in noisy environments, *Tamil 2006* 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 speaking style and a dissimilarity to the training database conditions (read out loud vs. commands to a machine).

Figure 7 shows the recognition performance on input comprised of a vocabulary word either alone (One Vocab word) or with other speech (Vocab Word plus), by site. 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 two onlookers for every participant.

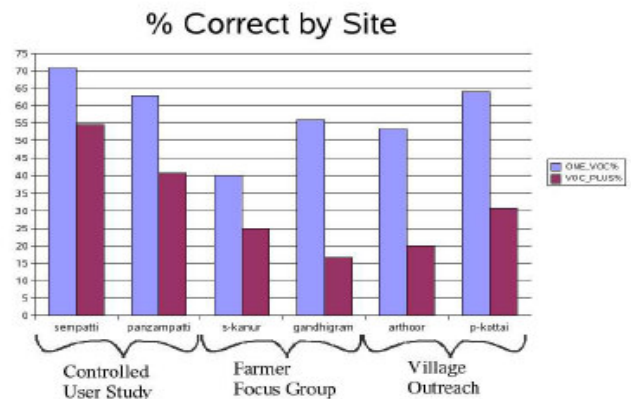


Figure 7. Recognition performance by site

Although 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 participants preferred to listen to the system for several minutes before speaking to it. Others offered corrections and suggestions for our pictures and content, especially the addition of more crops to the system. When prompted explicitly, some subjects reported preferring the touch screen as a means of input, others preferred speech.

The three recording conditions (Table 2) 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 (village outreach). We tried to balance controlled user studies with existing methods of community information flow used by MSSRF (farmer focus groups). We observed that in both the village outreach and farmer focus groups, the MSSRF staff played a larger role in the evaluation, which enabled lively and informed debates about recommended agricultural practices and farmer’s successes and failures with current practices. In addition, villagers had the opportunity to learn where and what services and materials were currently available to them at their VKC and VRC Sempatti. MSSRF staff heard feedback from villagers about the services and information they desired. The farmer focus groups also emphasized the advantage that software with speech and graphical output can have in a multiple user setting, which was not apparent in traditional user study conditions.

5 ASR Experiments

Based on the speech we collected from within the Banana Crop SDS, we ran a series of recognizer experiments. In the field, the recognizer performed at only 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. Simple noise robustness methods such as cepstral mean subtraction are important to apply for factoring out environmental noise and generalizing across tasks and speakers. Sections 5.1 and 5.2 present our findings from simulations of conditions of no available training data (cross-language transfer) and very limited training data (adaptation). For all experiments, we used the following databases: SDS Tamil 2006, Tamil 2006, Tamil 2005, and English TIMIT (Table 3).

Data Sets	Size	Dictionary Size	Description
Tamil (2006)	<i>very small</i> 170 words	<i>very small</i> 28 words	agricultural words <i>read</i> out loud by MSSRF staff in a fairly <i>quiet</i> office in Dindigul district

Data Sets	Size	Dictionary Size	Description
SDS Tamil (2006)	<i>very small</i> 377 words	<i>very small</i> 28 words	same agricultural words <i>spoken</i> by application villagers <i>indoors and out</i> in Dindigul district
Tamil (2005)	<i>small</i> 10K words	<i>very small</i> 50 words	digits and verbs <i>read</i> or guessed out loud by speakers of all literacy levels <i>indoors and out</i> in three districts
English (TIMIT)	<i>medium</i> 50K words	<i>medium</i> 6K words	phonetically balanced sentences <i>read</i> out loud in a <i>quiet</i> laboratory setting

Table 3. English and Tamil annotated corpora.

5.1 Cross-language Transfer

When an annotated corpus is unavailable, the options are to build one by collecting and transcribing speech, as we did for Tamil 2005, or to train a recognizer on an available corpus of speech in another language, such as the English TIMIT corpus.

First, we 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 context-dependent, three-state, left-to-right HMMs with a mixture of diagonal-covariance Gaussian component output distributions. The acoustic models are trained by first starting with a default initialization. Then, a set of monophone HMMs with single-Gaussian output distributions are trained, which are used to tie the context-dependent triphone states using decision trees. Using these tied triphone states, the number of Gaussian components at the output distributions is grown to 16, with a splitting-and-training algorithm. The decoding is first-pass using a simple grammar, which outputs a single word. No language model is required.

ASR Tasks	Test on	
Train on	SDS Tamil (2006)	Tamil (2006)
English (TIMIT)	30.2%	66.1%
Tamil (2005)	68.7%	97.1%

Table 4. ASR performance by language and data quality

Recognition results were significantly better when trained on a small amount of same language data than when trained on a medium amount of English data. An SDS powered by a recognizer trained only on English would only predict the correct word 30% of the time. The accuracy on the Tamil

2006 recordings is significantly higher than for the speech recorded within the SDS (97.1% vs. 68.7%). There is a substantial mismatch in the properties of read speech and live, interactive speech, suggesting that a recognizer that is initially trained on or adapted to speech collected from interactions with the SDS will show improved performance.

5.2 Language Adaptation

For the remaining experiment, we test ASR performance on the SDS Tamil 2006 corpus, as it represents the speech and context of participants retrieving information from a real application. The Tamil 2006 corpus represents an available, yet very limited language corpus. We train a recognizer on either English or Tamil 2005 using the recognizer recipe previously described in Section 5.1, and then adapt the recognizer to the Tamil 2006 corpus using maximum likelihood linear regression [Young, 1997]. Results are presented in Table 5.

IWR Tasks	All Tests on Field Tamil (2006)	
	No Adaptation	With Adaptation to Careful Tamil (2006)
English (TIMIT)	30.2%	80.4%
Tamil (2005)	68.7%	82.2%

Table 5. Recognizer performance in conditions of no or limited available Tamil resources.

Adaptation to Tamil 2006 improves performance for both the recognizer trained on English, and the recognizer trained on Tamil 2005. Adapting the system trained on a medium sized English database yields higher accuracy than a system trained on a small amount of Tamil (80.4% vs. 68.7%). Adaptation to Tamil 2006 improves accuracy for both the Tamil-trained system (68.7% vs. 82.2%) and the English-trained system (30.2% vs 80.4%) improves accuracy. Interestingly, the results are comparable (82.2% vs. 80.4%). There is very little gain to be had by collecting and annotating a corpus like Tamil 2005, which took an estimated 100 hours of expert time, when adapting to a very small corpus from an English-trained system yields similar results. Note that there is a substantial mismatch in the properties of Tamil 2006 and SDS Tamil 2006. If a system were trained on a medium amount of English and adapted to speech collected from within a live SDS, accuracy would substantially improve for the SDS recognizer.

6 Conclusion

This paper reviews literature on the language and information requirements likely to be found in primarily oral, limited resource environments. Speech technologies and techniques that are small, scalable, and easy to be modified and updated by local stakeholders in community development can be constructed to deliver accurate, locally-relevant information to individuals regardless of their literacy level. Language adaptation and integrated data

collection proved to be very useful techniques for exploiting available language resources. Recording speech from within an SDS, meets the user’s needs (gaining access to relevant information) and the system’s needs (gathering speech instances to enable recognition) simultaneously.

In future studies, we would like 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, pre-filtered, and gradually integrated (automatically or semi-automatically) to adapt to the dialect and speaking style.

We encourage further work in simple, affordable designs for speech synthesis and UI designs for text-free browsing and searching across libraries of audio and digital media.

Acknowledgments

The authors wish to acknowledge the following: TIER group, Richard Carlson, Chris Oei, Joyojeet Pal, Chuck Wooters, Dilek Hakkani-Tür, Srinivasan Ramaswamy, Amrita University, and our reviewers. The authors gratefully acknowledge the staff of the MSSRF VRC Sempatti for their enthusiasm, ideas and hospitality. This material is based upon work supported by the National Science Foundation under Grant No. 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

- [Barnard *et al.*, 2003] E. Barnard, J. P. L. Cloete, and H. N. Patel. Language and Technology Literacy Barriers to Accessing Government Services. In *EGOV*, Heidelberg, Germany, 2003.
- [Boas *et al.*, 2005] T. Boas, T. Dunning, and J. Bussell. Will the Digital Revolution Revolutionize Development? Drawing Together the Debate. *Studies in Comparative International Development* 40(2):95-110, 2005.
- [Boroovah, 2004] V. K. Boroovah. Gender Bias among Children in India in Their Diet and Immunization against Disease. *Social Science and Medicine* 58:1719-1731, 2004.
- [Brewer *et al.*, 2006] E. Brewer, M. Demmer, M. Ho, R. J. Honicky, J. Pal, M. Plauché, and S. Surana. The Challenges of Technology Research for Developing Regions. *IEEE Pervasive* 5(2):15-23, April-June, 2006.
- [Castro-Caldas *et al.*, 1998] A. Castro-Caldas, K. M. Petersson, A. Reis, S. Stone-Elander, and M. Ingvar. The Illiterate Brain. *Brain* 121:1053-1063, 1998.
- [Chisenga, 1999] J. Chisenga, Global Information infrastructure and the question of African content. In *Proc. of the 65th IFLA Council and General Conference*. Bangkok, Thailand, 1999.
- [Crystal, 2000] David Crystal. *Language Death*. Cambridge University Press, 2000.

- [Davel and Barnard, 2004] M. Davel and E. Barnard. The Efficient Generation of Pronunciation Dictionaries: Machine Learning Factors during Bootstrapping. In *Proc. of ICSLP*, Korea, 2004.
- [Delgado and Araki, 2005] R. L.-C. Delgado, and M. Araki. *Spoken, Multilingual and Multimodal Dialogue Systems: Development and Assessment*. John Wiley & Sons, Ltd., Northwest Sussex, England, 2005.
- [Del Galdo and Nielsen, 1996] E. M. Del Galdo and J. Nielsen. *International User Interfaces*. John Wiley & Sons, Inc., New York, USA, 1996.
- [Deo et al., 2004] S. Deo, D. M. Nichols, S. J. Cunningham, and I. H. Witten. Digital Library Access for Illiterate Users. In *Proc. of the 2004 International Research Conference on Innovations in Information Technology*, Dubai, U.A.E., October, 2004.
- [Dias et al., 2005] M. Dias, A. Roazzi, and P. L. Harris. Reasoning from Unfamiliar Premises. *Psychological Science*, 16(7):550-554, 2005.
- [Dindigul Statistics, 2004] Assistant Director of Statistics of Dindigul. District Statistical Handbook, 2004.
- [Tamil Nadu Census, 2001] *Tamil Nadu Primary Census Abstract, Census, 2001*, Directorate of Census Operations.
- [Donner, 2004] Jonathan Donner. Microentrepreneurs and mobiles: An Exploration of the Uses of Mobile Phones by Small Business Owners in Rwanda. *ITID*, 2(1):1-21, 2004.
- [Ethnologue, 2006] Ethnologue. Web Publications. www.ethnologue.com. Accessed July 2006.
- [Jelinek, 1996] F. Jelinek. Five speculations (and a divertimento) on the themes of H. Bourlard, H. Hermansky, and N. Morgan. *Speech Communication*, 18:242-246, 1996.
- [Johnson, 2001] D. G. Johnson. *Computer Ethics*. Prentice-Hall, Inc., New Jersey, USA, 2001.
- [Kanthak and Ney, 2003] S. Kanthak and H. Ney. Multilingual Acoustic Modeling Using Graphemes. In *Proc. of the European Conference on Speech Communication and Technology*, Switzerland, 2003.
- [Kemp and Waibel, 1999] T. Kemp and A. Waibel. Unsupervised Training of a Speech Recognizer Using TV Broadcasts: Recent Experiments. In *Proc. of EUROSPEECH'99*, pages 2725-2728, Budapest, Hungary, September 1999.
- [Kumar and Wei, 2003] C. S. Kumar and F. S. Wei. A Bilingual Speech Recognition System for English and Tamil. In *Proc. of the Fourth Information, Communications and Signal Processing and Fourth Pacific Rim Conference on Multimedia*, Singapore, December, 2003.
- [Lakshmi and Murthy, 2006] A. Lakshmi and H. A. Murthy. A syllable-based speech recognizer for Tamil. In *Proc. of INTERSPEECH 2006 -ICSLP*, pages 1878-1881, Pittsburgh, USA, 2006.
- [Lamel et al., 2000] L. Lamel, J.-L. Gauvain, and G. Adda. Lightly Supervised Acoustic Model Training. In *Proc. of ISCA ITRW ASR2000*, pages 150-154, Paris, September, 2000.
- [Lansdale and Ormerod, 1994] M. W. and T. C. Ormerod. *Understanding Interfaces: A Handbook of Human-Computer Dialogue*. Academic Press, Harcourt Brace & Company, Publishers, London, 1994.
- [Lievesley and Motivans, 2000] D. Lievesley and A. Motivans. *Examining the Notion of Literacy in a Rapidly Changing world*, UNESCO Institute for Statistics, Paris, August, 2000.
- [Medhi et al., 2006] I. Medhi, A. Sagar, and K. Toyama. Text-Free user interfaces for illiterate and Semi-literate users. In *Proc. of ICTD*, Berkeley, USA, May, 2006.
- [Moore, 2003] Roger K. Moore. A Comparison of the Data Requirements of Automatic Speech Recognition and Human Listeners. In *Proc. of EUROSPEECH'03*, pages 2582-2584, Geneva, Switzerland, September, 2003.
- [Petersson et al., 2000] K. M. Petersson, A. Reis, S. Askelöf, A. Castro-Caldas and M. Ingvar. Language processing modulated by literacy: a network analysis of verbal repetition in literate and illiterate subjects. *Journal of Cognitive Neuroscience*, 12(3):364-382, 2000.
- [Plauché et al., 2006] M. Plauché, C. Wooters, D. Ramachandran, J. Pal, and N. Udhayakumar. Speech Recognition for Illiterate Access to Information and Technology. In *Proc. of ICTD*, Berkeley, USA, May, 2006.
- [Psacharopoulos, 1994] G. Psacharopoulos. Returns to investment in education: a global update. *World Development*, 22(9):1325-1343, 1994.
- [Riccardi and Hakkani-Tür,] G. Riccardi and D. Hakkani-Tür. Active and unsupervised learning for automatic speech recognition. In *Proc. of EUROSPEECH'03*, Geneva, Switzerland, September, 2003.
- [Schneiderman, 1992] B. Schneiderman, *Designing the User Interface: Strategies for Effective Human-computer interaction*. Addison-Wesley, Boston, USA, 1992.
- [Schultz and Waibel, 1988] T. Schultz and A. Waibel. Multilingual and Crosslingual Speech Recognition. In *Proc. DARPA workshop on Broadcast News Transcription and Understanding*, pages 259-262, USA, 1998.
- [Schumacher, 1973] E. F. Schumacher, *Small is Beautiful: Economics as if People Mattered*. Blond & Briggs, Ltd., London, U.K., 1973.
- [Scribner, 1977] S. Scribner. 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*, pages 483-500, Cambridge University Press, New York, USA, 1977.
- [Soola, 1988] E. O. Soola, Agricultural communication and the African Non-Literate Farmer: the Nigerian Experience. *Africa Media Review*, 2(3):75-91, 1988.
- [Udhayakumar et al., 2004] N. Udhayakumar, R. Swaminathan, and S. K. Ramakrishnan. Multilingual Speech Recognition for Information Retrieval in Indian Context, In *Proc. of HLT/NAACL Student research workshop*, Boston, USA, 2004.
- [Waibel et al., 2000] A. Waibel, P. Guetner, L. Mayfield Tomokiyo, T. Schultz, and M. Woszczyna. Multilinguality in Speech and Spoken Language Systems. In *Proc. of IEEE*, 88(8):1297--1313, 2000.
- [Young, 1996] Steve Young. A review of large-vocabulary continuous speech recognition. *IEEE Signal Processing Magazine*, pages 45-57, September 1996.
- [Young, 1997] Steve Young. *The HTK BOOK*. Version 3.2. Cambridge University, U.K., 1997.