Madelaine Plauché, Udhyakumar Nallasamy, Joyojeet Pal, Chuck Wooters, and
Divya Ramachandran.

M

Speech Recognition for Illiterate Access to Information and Technology

***Abstract*—In rural Tamil Nadu and other predominantly illiterate communities throughout the world, computers and technology are currently inaccessible without the help of a literate mediator. Speech recognition has often been suggested as a key to universal access, but success stories of speech-driven interfaces for illiterate end users are few and far between. The challenges of dialectal variation, multilingualism, cultural barriers, choice of appropriate content, and, most importantly, the prohibitive expense of creating the necessary linguistic resources for effective speech recognition are intractable using traditional techniques.**

This paper presents an inexpensive approach for gathering the linguistic resources needed to power a simple spoken dialog system. In our approach, data collection is integrated into dialog design: Users of a given village are recorded during interactions, and their speech semi-automatically integrated into the acoustic models for that village, thus generating the linguistic resources needed for automatic recognition of their speech.

Our design is multi-modal, scalable, and modifiable. It is the result of an international, cross-disciplinary collaboration between researchers and NGO workers who serve the rural poor in Tamil Nadu. Our groundwork includes user studies, stakeholder interviews and field recordings of literate and illiterate agricultural workers in three districts of Tamil Nadu over the summer and fall of 2005. Automatic speech recognition experiments simulating the spoken dialog systems' performance during initialization and gradual integration of acoustic data informed the holistic structure of the design.

Our research addresses the unique social and economic challenges of the developing world by relying on modifiable and highly transparent software and hardware, by building on locally available resources, and by emphasizing community operation and ownership through training and education.

***Index Terms*—User interface, human factors, speech recognition, spoken dialog system, illiteracy, IT for developing regions.**

I. BACKGROUND

A. Agriculture, Literacy, and Information Technology (IT)

Agriculture is the main source of employment for over 40% of the labor force in developing regions around the world. To date, in the poorest nations of the world, over 80% of local households depend on some form of agriculture for sustenance [1]. In Tamil Nadu, 37.47% of full-time workers and 71.64% of marginal workers are in the agricultural sector – a majority of them small cultivators or seasonal laborers (Fig. 1).

Although the ability and inclination to base sale decisions on price information is open to question [2], [3], studies have suggested that under the right circumstances, price and market information can improve farmer welfare [4], [5], improvements in technical efficiency require information (pests, diseases, new seeds, new techniques) [3], and IT-based information networks can help raise the price of goods sold for small farmers [6]. In addition, a substantial body of evidence indicates that literacy increases farmer productivity and earning potential [7], [8]. Literacy is also an important factor in social potential, such as children's health and nutrition [9].

Fig. 1. Rural farmers in the Sempatti village plaza, Madurai district.

Speech-driven interfaces have often been suggested as key to universal access in countries like India, where two thirds of the 870 million illiterates in the world today are found [10]. In rural Tamil Nadu, illiteracy rates can be as high as 50% for men and 80% for women [11]. Traditional speech technologies and user interface design, developed for (and by) the literate in resource-rich countries, are a poor fit for users in developing regions. In this paper, we present a novel design for a spoken dialog system (SDS) that is scalable, modifiable, and designed to operate with limited linguistic resources. We hope our initial results (and failures) from a pilot study, field recordings, and speech recognition simulations will inform and encourage further explorations of speech technologies for the social and economic realities of developing regions, and even extend the knowledge to other domains, such as education and health.

B. MSSRF knowledge centers

The MS Swaminathan Research Foundation (MSSRF) is an established NGO in Tamil Nadu dedicated to a pro-nature, pro-poor and pro-women approach to fostering economic growth in rural, agricultural areas. MSSRF community centers, known as Village Knowledge Centres (VKCs),

[M]anuscript received December 20, 2001. The work was supported by the NSF, grant ANI-0335289

throughout the state provide general information on agricultural goods and techniques, weather, local news, employment, and government policies and services. Volunteers from the community, or “knowledge workers,” provide the rural poor with training and educational materials on health, social issues, and entrepreneurial development.

Much of our research has been in the VKC networks in the union territory of Pondicherry, which is monitored from Chennai, the state capital and MSSRF’s headquarters. Each morning, volunteers travel to nearby markets, collect information about market prices, and report back to MSSRF contacts. The information is consolidated, digitized, and disseminated to all VKCs through one of the following communication links: phone, dial-up voice transfer, or wi-fi. The following morning, villagers may access information at a nearby VKC by reading posted bulletins (Fig. 2), listening to loudspeaker broadcasts, or working one-on-one with a knowledge worker [12]. Knowledge workers in three districts reported a need to disseminate this information in a form accessible to illiterate populations. In Sempatti, knowledge workers reported much success in educating illiterate populations using a touch-screen kiosk, Livestock Guru [13]. A major limitation to the tool, however, was its inability to be modified.

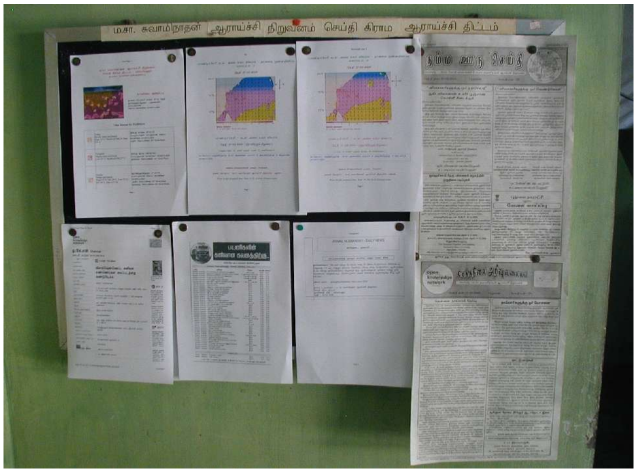

Fig. 2. Daily postings of weather, news, and market prices in the Sempatti village knowledge center. Illiterate villagers require an interpreter to gain access to this information.

C. Speech-driven interfaces for developing regions

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. Speech-driven user interfaces (UI) are cheaper than display-based UI solutions and more accessible than text-based UI solutions. Many applications for rural Indian information technology (IT) that provide information on health, weather, employment, news, and agriculture could enhance the social and economic development of the rural poor [14] and could be constructed around a limited-vocabulary speech-driven interface.

Speech-driven UI solutions are often suggested to increase access for users in developing regions, but they are rarely attempted, due primarily to the challenges of multilingualism, dialectal and cultural diversity, as well as the prohibitive costs of developing the linguistic resources needed to power speech recognition.

India has 22 “scheduled” (official) languages but estimates range from 450 [15] to 850 languages [16] overall. India is the 9th most linguistically diverse country, with a 93% probability that any two people of the country selected at random will have different mother tongues [15]. An SDS 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.

Speech technologies, such as automatic speech recognition (ASR), require the collection and hand annotation of a large corpus of speech and a dictionary of all possible words in the language with all possible pronunciations for each word. The availability of linguistic resources, such as training data and pronunciation dictionaries, which are arguably the most costly part of development, are taken for granted by developers with English-speaking users in mind, for example. Only a handful of speech technology efforts [17]-[19], have been dedicated to Tamil, however, which is spoken by over 60 million people in Tamil Nadu, Sri Lanka, and elsewhere [20]. The majority of the world's languages, spoken in developing regions, currently have no associated linguistic resources.

Power, connectivity, and limited infrastructure are all significant obstacles in developing regions. Finally, user interface design requires a familiarity with the cultural and economic context of the user. As a whole, people who have never learned to read or write are poorly understood by researchers and developers of technology.

D. User interface design for illiterate users

Illiteracy is usually used to describe an individual's competency at the tasks of reading and writing. A *functional illiterate* is an individual who may have had some exposure to education, but whose low competency at these tasks prevents the individual from wholly participating in society. Same-language subtitles and other attempts to reinforce literacy by combining text and images have met with some success [21].

Other efforts to provide immediate IT access to illiterates in developing regions, where “gaining access to target users, rural Indian villagers with literacy challenges, for experimentation would be challenging and costly,” [22] have attempted to address the inappropriateness of traditional HCI design techniques. However, their solutions are not informed by actual user studies in developing regions and heavily rely on the use of icons [23], requiring users to memorize a set of symbols and their corresponding meanings, which is essentially a literate task.

In a UI design study for a financial management system designed for micro-credit groups in rural India, Parikh *et al.* [24] conducted extensive contextual field studies and found that many of the women they worked with could read and write numbers and perform calculations. The developers used this numerical literacy as partial leverage in their graphic-based interface, which they tested iteratively to arrive at an interface that was well understood by most users. Although numbers were well understood among their numerically literate users in the context of performing calculations, when they were used to represent a concept, users experienced difficulty in remembering the intended meaning. They also reported that colors, especially reds and yellows, were useful to illiterate users in identifying relevant parts of the interface.

Cognitive and brain activation studies on literate and illiterate subjects have shown that learning the skill of writing during childhood alters the functional organization of the adult brain, in particular, the phonological processing of spoken language [25], [26]. Women who had never been to school were tested on their ability to repeat words and pseudowords (words phonetically similar to comprehensible words, but with no meaning). The illiterate women performed only slightly worse on the word repetition task than literate women but were much less adept at repeating nonsense words, often producing phonologically similar *actual* words instead. The results of this study tell us little about differences in speech *production* among literate and illiterate groups. However, the difficulty in repeating pseudowords among illiterate participants suggests that successful UI solutions for this group will rely on command words that are familiar and relevant to everyday language use.

E. Our approach

In this paper, we investigate the possibilities for a speech-driven UI design for users in developing regions who may have never been exposed to formal education. We propose a multi-modal interface that requires no training and provides users with a variety of visual and audio cues and input options, allowing each user the flexibility to interact with the application to the best of their ability and the system's ability.

This paper describes an SDS built from standard speech technologies to be accessible to illiterate users and scalable to other dialects or application domains. The rest of the paper is organized as follows: In section 2, we outline results from a pilot user study. In section 3, we present our findings from field recordings of the speech of rural Tamilians in three districts of Tamil Nadu. In section 4, we share results from a series of experiments using standard ASR techniques with limited available training data and linguistic resources. In section 5, we present the resulting design for our dialog system. Section 6 draws conclusions and discusses future work.

II. Tamil Market

A. Initial approach

Initially, we chose to restrict ourselves to the domain of speech-based UI. We developed Tamil Market, an SDS that provides weather, market prices for crops, and agricultural information to users who navigate by uttering one of 30 possible words in Tamil (Fig. 3).

The command words of an SDS like Tamil Market will vary with the needs and dialects of each village, which are difficult to anticipate in advance. We did not attempt to anticipate the exact content and command vocabulary in advance. Instead, we designed Tamil Market as a template to determine whether an SDS can be powered by limited-resource speech recognition and to initiate involvement by rural villagers in the development of an application that suits their social and economic realities.

TM: *Welcome to Tamil Market. I can tell you about crop prices, weather, and rainwater collection.*

Would you like to hear about crop prices? Please say "yes" or "no."

User: *Yes.*

TM: *Please say the name of the crop you would like today's market price for.*

User: *Wheat.*

TM: *I'm sorry, I didn't understand what you said.*

I will tell you all the crop prices in rupees per kilo. Wheat, 9, Rice, 10, Onion, 5.5, [...]

Fig. 3. An English translation of a sample dialog with Tamil Market.

Although an application-independent command vocabulary (e.g., "back," "next," "repeat") might mitigate variations in application content, it would require users to memorize an arbitrary set of command words. Tamil Market is primarily operated with the Tamil words for "yes" (*aamaam*) and "no" (*illai*), which are likely to retain their meaning and function across different dialects.

Explicit prompts for input options and a redundant structure ensure that users have access to all available information, albeit not in the most direct way, even when their input is mis-recognized or missing altogether. In this way, Tamil Market can accommodate first-time users and users unfamiliar with technology with no need for a separate training session.

B. User Study

The resulting SDS ran on a PC laptop to which participants spoke using a custom *phone mic*, a computer microphone and speaker embedded in a recycled phone handset (Fig. 4). Tamil Market was tested by 13 villagers of varying degrees of literacy in three districts of Tamil Nadu using a Wizard-of-Oz user study [27], in which one of the researchers played the role of the recognizer and sporadically injected errors approximately 2% of the time [28].

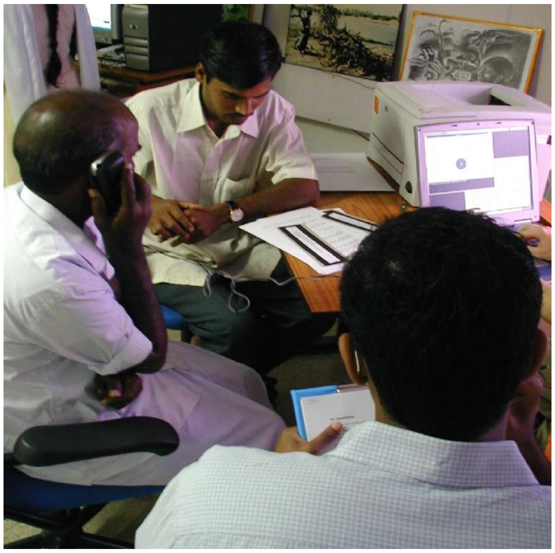

Fig. 4. Pilot user study of Tamil Market.

C. Results and Design Considerations

Participants were able to operate the system with no training, even though many had never before used a phone or computer. User errors usually occurred at the beginning of the interaction with Tamil Market and consisted of no input, complete sentences, or unexpected words or pronunciations.

Participants valued accuracy of information and time of interaction, which averaged about three minutes. Participants expressed pride and excitement at hearing a computer speak their language and were forgiving of the 2% injected recognition errors. Participants also offered many suggestions about the content of Tamil Market, including an interest in varieties of crops, grades of crops, prices at the local market, block market, and wholesale market.

Interest in Tamil Market correlated to distance from the nearest marketplace. Participants who lived far from a market reported that if the information was accurate, they would be willing to pay the price of a phone call (one rupee, equivalent to $0.02) to use such a system. In villages near a major market, however, at least one individual would visit the market each day and could easily report the market prices to other members of that village.

We found that illiterate villagers were more reluctant to participate in the study and had greater difficulty restricting their input to recognizable forms, despite explicit prompts, but reported confidence that with training they could operate such a system.

Our main finding was that an SDS like Tamil Market can be operated by rural farmers of varying degrees of literacy with little or no training. Illiterate users would likely perform better with more training or additional cues, such as graphic output. Finally, despite explicit prompts in the dialog system, participants respond using a wide variety of command words that change in pronunciation and form from district to district, with a range of variations difficult to predict in advance. In future studies we hope that Tamil Market may serve as a modifiable tool for collaborative, rapid development of applications that meet the needs of a given community.

III. Speech Recordings in Tamil Nadu

Speech recordings of rural villagers in three districts of Tamil Nadu were conducted during one field visit in 2004 and another in 2005 to quantify linguistic variations by geography and literacy and to determine what linguistic resources are necessary to power a small-vocabulary SDS, such as Tamil Market.

A. Data collection

During two field visits, the speech of 77 volunteers was recorded in three separate districts in Tamil Nadu, India (Fig. 5). 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 1) and by recruiting method.

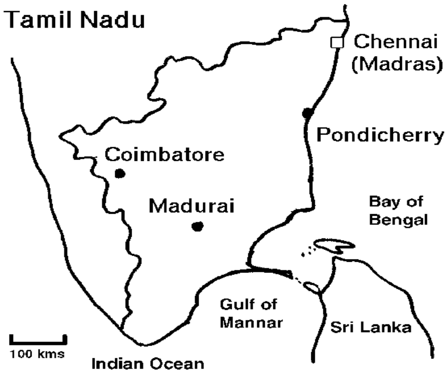

Fig. 5. Map of Tamil Nadu, India.

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 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.

Site (Year)	Number of Speakers	Average Age of		% Illiterate
		Females	Males	
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%**

Table 1. Number, age, gender, and literacy of speakers by site. Here, "illiterate" 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.

The age of speakers ranges from 18 to 80 years old, with an average age of 40. The men in this study average almost 10 years older than the women. University students contribute to a younger average age in Coimbatore. Eye care patients in Madurai represented an older demographic.

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, meant recording outside or in noisy rooms.

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 computer table microphone and a laptop running mouse-driven software that displayed a written word and recorded the speaker saying the target word.

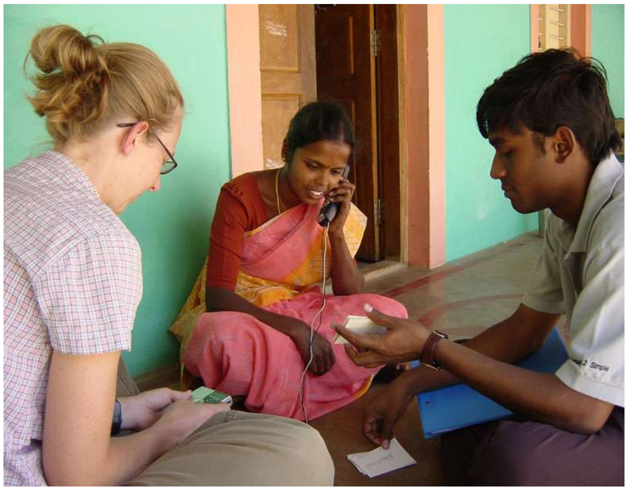

Fig. 6. Recording a literate woman in Ettimadai, Coimbatore district.

Data collection in rural Tamil Nadu (2005) relied instead on flashcards and the phone mic connected to a Sony MD Walkman (MZ-NH900) (Fig. 6). 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 phone mic avoided the need to clip or fasten equipment to the speaker's clothing and did not require the use of a table.

TABLE 2

DIGITS ELICITED FOR SPEECH RECORDINGS

Numerals	Tamil Script	English Transliteration	International Phonetic Alphabet IPA
0	பூஜ்யம்	pUjyam	/pūʤam/
1	ஒண்ணு	oNNu	/^{w}oɲ:ɯ/
2	ரெண்டு	rendu	/reɲɖɯ/
3	மூணு	mUNu	/mūɲɯ/
4	நாலு	nAlu	/nālɯ/
5	அஞ்சு	anju	/aɲʤɯ/
6	ஆறு	ARu	/ārɯ/
7	ஏழு	ELu	/^{j}ēẓɯ/
8	எட்டு	ettu	/^{j}eʈʈɯ/
9	ஒம்போது	ompOthu	/^{w}ombōðɯ/
10	பத்து	paththu	/pat̪t̪ɯ/

Bilingual flashcards with digits 0-10 written in both numerical and orthographic form were randomized and shown to speakers one at a time (Table 2). 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 into a display of fingers. (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. All speech was recorded at 44kHz, stereo, then downsampled to 16kHz, mono. Finally, the speech waveform was extracted with a collar of approximately 100 ms of silence.

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.

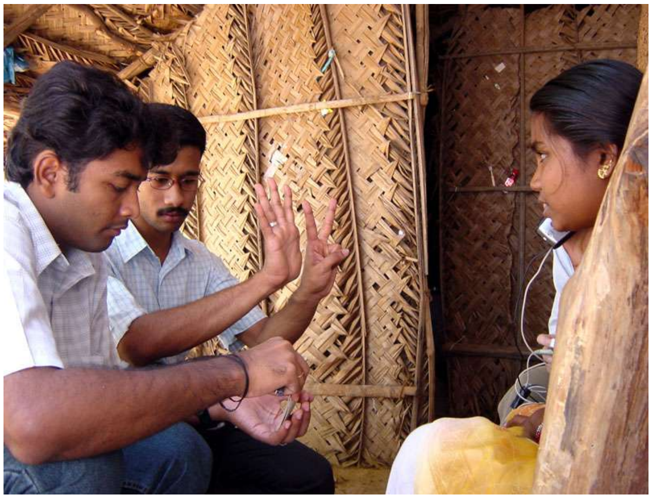

Fig. 7. Recording an illiterate woman in Ettimadai, Coimbatore district.

B.Linguistic variation by literacy

One topic we sought to explore with our field recordings was whether illiterate speech varied substantially from literate speech. Unfortunately, in our data, the differences between literate and illiterate speakers were due to differences in elicitation protocols and environmental factors rather than linguistic characteristics of speech production. We attempted an adaptive experimental protocol but failed to anticipate the cognitive difference between the task of determining a number of displayed fingers and the task of reading numbers aloud [29]. The elicitation technique developed for illiterate speakers resulted in significant implications in the speech domain, such as more disfluencies (filled pauses, false starts, repetitions), and shouting (a translated example: "um…thr...no! FOUR! Four."). This effect was compounded by the environmental conditions: Recordings often took place outside or in a crowded room, where the task of guessing the number of displayed fingers was often irresistible to helpful bystanders.

Although our data are inconclusive regarding any effect literacy has on the phonetic characteristics of speech production, we observed anecdotally that the illiterate villagers we worked with were more likely to produce words and pronunciations specific to their region than villagers who have attended school and been exposed to materials and instruction in standard Tamil.

C. Linguistic variation by geography

Like most languages, Tamil varies by geography (6 main dialects), social factors (caste, education), and register (spoken vs. written). Tamil spoken in Chennai and throughout the East Coast (including Pondicherry) has emerged as the modern standard- the Tamil of press, literature, and radio. Modern standard Tamil is considered accessible to most speakers of mainland Tamil, regardless of literacy [20].

We hypothesized that words for digits 0-10 would be similar across Tamil Nadu, though pronunciations might vary. Pronunciation of the consonant in the Tamil word for "seven," (*ELu*) for example, varied significantly ($p<0.01$, N=385) by geography. Speakers from Pondicherry use a lateral approximate, similar to English /l/, whereas speakers in Coimbatore and Madurai districts use a phoneme unique to Tamil, the retroflex palatal fricative (similar to English /z/, but with the tongue curled backwards). Age, gender, and education level were not predictive factors in this phonetic variation.

Unexpectedly, we found significant variation ($p<0.01$, N=379) in word choice for the Tamil word for "zero." Even among literate speakers reading a flashcard explicitly marked, பூஜ்யம் (*pUjyam),* three speakers in Coimbatore and 6 in Madurai said, *zeero* or *jeero*. In Madurai district, 11 speakers said *saiber*. All three forms are borrowings, from Hindi, English, and Arabic, respectively, since Tamil does not have its own term for "zero." Age, gender, and education level were not predictive factors in this variation.

D. Results and design considerations

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 different words and different pronunciations for everyday concepts that are difficult to predict in advance. An SDS that adapts to its users by learning their choice and pronunciation of command words would avoid the challenge of anticipating appropriate command words. 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. Instead, 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.

Further benefits to an SDS with integrated data collection and more detailed design considerations are discussed in the following section.

IV. Limited-resource ASR

Traditional ASR techniques depend on the existence of pronunciation dictionaries for the target language and solve the problem of dialectal variation by training robust acoustic models from very large amounts of manually transcribed speech and sophisticated adaptation techniques. Currently, promising solutions to the development of large-vocabulary ASR of Tamil, for which there are no equivalent linguistic resources, are being proposed in the form of techniques for bootstrapping and mixing models from acoustically similar dialects or languages [17]. Instead of pursuing a long-term goal of large-vocabulary recognition for all Tamil speakers, we limit our scope to a single village and focus our efforts on a modifiable, transparent SDS powered by a small-vocabulary whole-word recognizer that is robust to noise, microphone changes, and dialectal variation by virtue of its simplicity.

In this section, we present a series of speech recognition experiments trained on the annotated speech we collected from our field recordings to demonstrate the performance of a simple, whole-word Tamil recognizer trained on limited available linguistic resources.

A. Phoneme versus whole-word models

Generally, for large vocabulary recognizers, acoustic models are developed at the level of the phoneme, which enables the recognition of words that are not included in the training data by extrapolating from phoneme models of words that are. Although phoneme models are more flexible to the addition of new words, they require the costly development by a linguist of a pronunciation dictionary: a list of each word ASR may encounter and all of its possible pronunciations. As previously mentioned, pronunciation dictionaries are currently unavailable for most languages of the world and may be too costly to develop for a specific application. One innovative approach to generating the necessary linguistic resources to power recognition for unsupported languages is DictionaryMaker, a tool that allows semi-skilled native speakers to efficiently create a pronunciation dictionary [30], [31].

For this study, we adopt whole-word models instead, which yield equivalent recognition performance for small-vocabulary ASR and do not require the development of a pronunciation dictionary, thus facilitating rapid, low-cost SDS deployment for languages with limited available linguistic resources.

B.Task complexity

We trained a whole-word recognizer using the Hidden Markov Tool Kit (HTK) on speech from 22 speakers in Pondicherry (2004). The models used 18 states with 6 diagonal gaussians per state. The remaining speech was input for recognition in three trials of varying complexity: all words, digits only, and six command words. Word error rates dropped for tasks with fewer options for correct word identity (Fig. 8).

Automatic recognition of a small set of words achieves low word error rates when trained on very little training data. An SDS that limits word options at each node could be powered by a simple whole-word recognizer such as this.

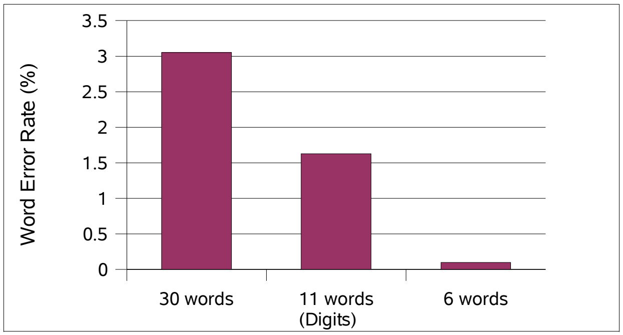

Fig. 8. Word error rate by task complexity.

C. Dialectal variation

To evaluate the influence of phonetic and lexical variations on our small-vocabulary, whole-word recognizer, we used HTK to train a system to recognize digits (equivalent to the 11 word task in Fig. 8) based on the speech of 22 Pondicherry (2004) speakers. We then tested the recognizer's performance on speakers from all three districts in our study (Fig. 9). 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=3168). Most errors were in the recognition of the words for “zero” and “seven.” Including alternate pronunciations for these digits in the whole-word dictionary improves recognizer performance.

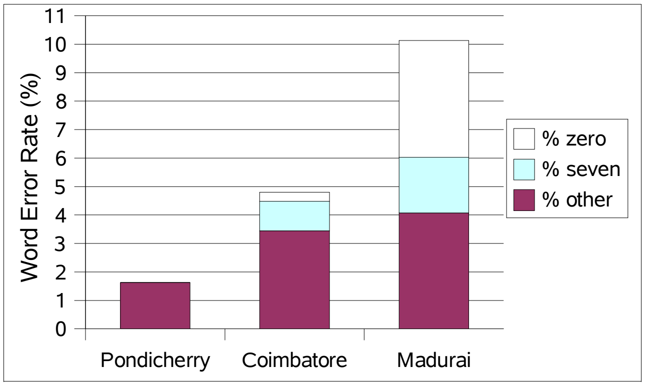

Fig. 9. Word error rate for digits by site. Errors for the words “zero” and “seven” are explicitly indicated. The recognizer was trained on data from Pondicherry (2004) speakers.

We considered the possibility that the lower word error rates were caused by a quieter environment for Pondicherry recordings, many of which took place in an office, but signal-to-noise ratios per speaker, automatically extracted using a speech/non speech detector, had no significant effect on word error rates.

D. 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 a simulation to determine how little data is needed to achieve acceptable error rates for the operation of an SDS like Tamil Market.

One speaker was randomly selected; his speech was set aside as the input speech. First, the input speech was recognized by a recognizer trained on the speech of a single speaker. The resulting word error rate was approximately 80% (Fig. 10). 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, and second, speakers from the test speaker's district were added first.

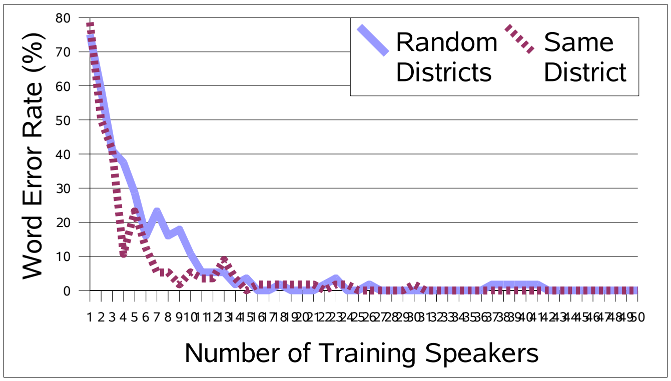

Fig. 10. Word error rate by amount of training data.

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 the parallel SDS user study described in previous sections. 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.

E. Results and design considerations

The results from these ASR experiments confirm what we already know about standard speech recognition:

- Performance improves for simpler tasks.
- Performance improves when input matches training data.
- Performance improves with more available training data.

However, the specific trials shown here for a small vocabulary whole-word recognizer trained on limited data can inform our SDS design 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 ten words or fewer at any given dialog node. Second, ASR performance by site further supports our proposal to integrate data collection into the SDS design, thus ensuring that input and training data match not only in dialect, but also in other nonlinguistic factors, such as channel and room acoustics. Finally, during deployment of an SDS like Tamil Market in a given village, ASR performance for a set of core command words (e.g., “yes” and “no”) could be cheaply and quickly initialized with the speech of only 15 volunteers, thus enabling SDS usability while acoustic models for other vocabulary words may still be impoverished.

Although the whole-word recognizer we trained on a small corpus of Tamil speech by no means represents cutting-edge speech technology, we hope that its performance in these trials, in combination with the Tamil Market user study, in which untrained, semi-literate users were able to access desired information despite errors in recognition at approximately 2%, encourages further studies in speech technologies for developing regions.

V. Current Design

The direction of our work is motivated by the desire to address the needs and conditions of the majority of the worlds' population whose lives could greatly benefit from the knowledge and services taken for granted in a handful of western countries, but whose social and economic conditions create a formidable barrier to gaining access to such information. For this study, we relied on previous field research, grassroots NGO workers, and our partner organizations to articulate the needs of populations in developing regions. In future work, we aim to use Tamil Market as a modifiable template to enable community-developed applications

Throughout Tamil Nadu, wages are low and illiteracy rates are high. However, the needs and dialects of rural villagers vary from district to district. Village knowledge centers (VKCs) in Pondicherry would benefit from an SDS like Tamil Market to provide tidal information and sea conditions to local fishermen, who currently rely on a daily broadcast over loudspeakers for this information. VKCs in Madurai district wish to distribute agricultural techniques and information over a phone line to increase access and decrease organizational costs.

The research presented in this paper represents an investigation into speech technology as a feasible means to extend access of IT to rural Tamilians of all literacy levels. Instead of requiring users to memorize a set of inappropriate command words or icons or attempting to anticipate the content of an SDS for each possible deployment, we explore a holistic design for an SDS that is powered by limited-resource ASR, can be cheaply initialized and rapidly deployed, is modifiable by a trained community member, semi-automatically adapts to the dialect of its users; and, unlike traditional ASR, scales to new dialects and new languages. Whether this technology results in increased social inclusion [32] and economic sustainability can only be determined only with field deployments and impact analysis, but our belief is that community involvement in early phases of design, community ownership and operation of the SDS and the villagers' speech recordings, and a focus on training and education during deployment will contribute to both.

A. Multi-modal input

Tamil Market was initially a speech interface designed to run over a phone line. Our field work and previous UI studies suggest that a parallel kiosk version, which combines speech with additional input (touch, typing) and output domains (graphics, local text) can provide multiple cues to unsophisticated or new users and, in the case of functional illiterates, reinforce literacy. Touch screen combined with digital photographs or graphics produced by a local artist is an ideal input for users of all literacy levels, but its cost and relative fragility make it inaccessible for a community center in rural Tamil Nadu.

We constructed a cheap and modifiable flex button system (similar to cash distribution machines) by soldering the “reset” buttons from used computers to the function keys of a dedicated keyboard (Fig. 11). The result is a system that accepts touch input as well as speech and whose construction is transparent, cheap, and easy to construct from locally available materials. The multi-modal kiosk displays digital photographs and corresponding Tamil text for command words, while still adhering to the same audio dialog structure as before, which provides relevant information even when the user provides no input or unrecognized input. Once acoustic models are adequate for recognition, voice input allows the most direct route to specific information. The multi-modal version of the SDS will be field tested in spring 2006.

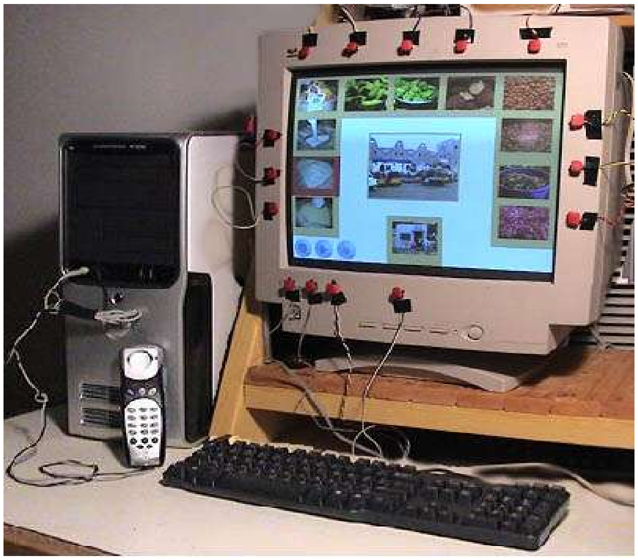

Fig. 11. Tamil Market kiosk prototype. Phone mic and modified keyboard with flex buttons are made from locally available recycled components.

B.Integrated data collection and clustering

We have provided several reasons to integrate the collection of linguistic samples into the SDS. We believe that this method for recording speech does not favor literate users over illiterate users and provides the best match of training and input data, resulting in robust ASR performance despite limited linguistic data.

The only substantial changes to the SDS initially presented is the addition of a “push-to-talk” button to the phone mic, as well as a corresponding indicator light. When the button is depressed, speech is recorded to an audio file and submitted for recognition and subsequent operation of the SDS (Fig. 12). The ability of the user to access desired information is unaffected by this modification.

TM: *Please say the name of the crop you would like today's market price for.*

User: *Wheat.* → saved as an audio file

TM: *I'm sorry, I didn't understand what you said.*
I will tell you all the crop prices in rupees per kilo. Wheat, 9, Rice, 10, Onion, 5.5 [...]

Fig. 12. An English translation of a sample dialog with Tamil Market.

The challenge of this approach is knowing which recorded audio files to incorporate into the acoustic models. Including mislabeled or inappropriate audio data in acoustic model training would be detrimental to overall recognition performance. The "push-to-talk" button will help separate intended speech from other audio data (e.g., silence, complete sentences). Automatically discarding files that are smaller than 50 ms or longer than 100 ms could also help to filter out complete sentences or truncated speech.

Suppose that most of the remaining acoustic files consist of single-word utterances; The identity of the uttered word must be known for the data to be included in ASR training. Without listening to each file, however, it will be impossible to know with certainty what the recordings contain. A trained VKC volunteer could listen to each file, discard it if necessary, and label it by its content if it is an instance of a single-word utterance. This process is time-consuming, however, so we propose a machine learning technique originally developed for speaker recognition to automatically organize the data by word identity.

In a study on speaker segmentation and clustering, Ajmera *et al.* [33] used the Bayesian Information Criterion (BIC) to perform speaker clustering without any prior knowledge of the identities or the number of speakers. We are currently investigating the use of this algorithm to cluster similar words from recorded input.

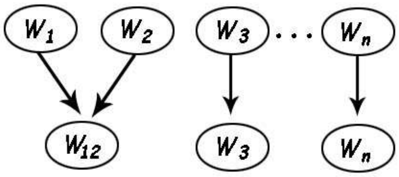

Fig.13. Merging Algorithm

In standard speech recognizer training, Mel-frequency cepstral coefficients are extracted every 10 ms. A recognizer is trained on the feature data and corresponding transcribed phonemes. Each phoneme is then represented by an n-state Hidden Markov Model (HMM), in which each state can be modeled by a Gaussian mixture. At first, HMM's of utterances of the same word are likely to be assigned to different clusters (Fig. 13). Given $\{w_1, w_2, w_3, ... w_n,\}$, the audio data to be identified, we want to find the optimal number of clusters and their acoustic models according to the likelihood of the data. The BIC is used as a merging criterion, as it imposes a trade-off between model quality and model complexity. Models are merged only if the log likelihood of the data given the combined model is higher than that of the sum of the two separate models.

Preliminary results from simulations of acoustic data are encouraging, showing that the same algorithm used to cluster speakers can be applied to cluster words, without knowing in advance what forms or what pronunciations are used in a given village. However, initialization of the models and optimization of the algorithm for small amounts of data must be explored to apply this algorithm to actual acoustic data. Recorded input (audio files) could then be clustered into a manageable number of groups (directories of files), likely to consist of the same word, greatly simplifying the amount of time needed to annotate audio input and generate acceptable acoustic models at each new deployment or modification to the system.

C. Community owned and operated

Historically, society has seldom given poor people ownership over the tools of production [32], [34]. Research and development efforts in the field of IT for developing regions suggest that involving community members in design and creation ensures that proposed solutions meet the needs of the community and provides the best chance for sustainability of technology in a community. Tamil Market was developed to investigate the role that speech technology could play in rural villages of Tamil Nadu at the request of an established NGO that serves these communities. In upcoming field visits, we will use Tamil Market as a modifiable template to allow MSSRF volunteers to produce an appropriate application, since we believe that people living in poverty can produce technological content that generates social and economic benefit.

VI. Future Plans and Conclusions

There are very few proposals for speech-driven applications in developing regions despite the need for UI solutions to universal access. The low literacy and educational levels of the rural poor, as well as their unfamiliarity with technology, makes the design of appropriate speech technologies a challenging task. We sought to investigate this problem by developing Tamil Market, a multi-modal, inexpensive spoken dialog system powered by a small-vocabulary single word recognizer. Based on user studies, field recordings, and a series of ASR experiments, we propose a design in which the SDS collects data to power a simple, whole-word recognizer that is robust in limited tasks despite minimal linguistic traning data. Unlike traditional speech technology techniques, our design addresses problems of multilingualism, literacy, and linguistic variation by restricting its scope to a single village. We focus instead on high scalability and modifiablity, resulting in a solution we believe will be rapidly deployable, useful, and well suited to limited-resource environments.

Acknowledgments

The authors thank their volunteers, their families, their hosts and partner organizations in Tamil Nadu, and colleagues and reviewers who contributed their encouragement and insights. This material is based upon work supported by the National Science Foundation under Grant No. 0326582.

References

[1] "Why Agriculture Still Matters," Chapter 3, World Employment Report 2004-04: Employment, Productivity and Poverty Reduction, International Labour Organization (ILO), December 2004.

[2] Hornik, R. *Development communication: Information, agriculture, and nutrition in the third world.* New York: Longman, 1988.

[3] Blattman, C., R. Roman, and R. Jensen, "Assessing the Need and Potential of Community Networking for Development in Rural India", *The Information Society* 19(5), 2003.

[4] Eggleston, K., R. Jensen, and R. Zeckhauser, "Information and communication technologies, markets and economic development." In *Global information technology report 2001–2002: Readiness for the networked world,* G. Kirkman and J. Sachs, eds. Oxford, UK: Oxford University Press, 2002, pp. 62–74.

[5] Prahalad, C.K. and A. Hammond, "Serving the poor profitably", *Harvard Business Review*, 2002.

[6] Kumar, R. 2004. "eChoupals: A Study on the Financial Sustainability of Village Internet Centers in Rural Madhya Pradesh" *Information Technology and International Development (ITID),* Vol. 2., Issue 1, Spring, 2004.

[7] Psacharopoulos, G. "Returns to investment in education: a global update." *World Bank PDR, Working Paper 1067*. Washington DC: World Bank, 1993.

[8] Directorate of Census Operations, *Tamil Nadu Primary Census Abstract,* Census, 2001.

[9] Borooah, V. K., "Gender Bias among Children in India in Their Diet. and Immunization against Disease" *Social Science and Medicine* 58, pp. 1719-1731,2004.

[10] United Nations. Unesco Institute of Statistics. Http://www.uis.unseco.org,, accessed April 11, 2005.

[11] Provisional Population Totals: India. *Census of India 2001*, Paper 1 of 2001. Office of the Registrar General, India. 2001

[12] Balaji. V., K. Rajamohan, R. Rajasekarapandy, S. Senthilkumaran, "Towards a knowledge system for sustainable food security: The information village experiment in Pondicherry," in *IT Experience in India : Bridging the Digital Divide,* Kenneth Keniston and Deepak Kumar, eds., New Delhi, Sage, 2004.

[13] Heffernan, C. "Fighting Poverty with Knowledge: The Livestock Guru." Report for DFID's Livestock Production Program, DFID, London. 2006.

[14] Sustainable Access in Rural India (SARI). http://www.tenet.res.in/sari, accessed February 9th, 2005.

[15] SIL International, http://www.sil.org/literacy/LitFacts.htm, accessed Februrary 9th, 2005.

[16] Noronha, F., "Indian language solutions for GNU/Linux" *Linux Journal*, 2002.

[17] Udhyakumar, N., R. Swaminathan, R., and S. K. Ramakrishnan, "Multilingual speech recognition for information retrieval in Indian context.", in *Proc. Student Research Workshop*, *HLT/NAACL*, Boston, USA, 2004.

[18] Saraswathi, S. and T. V. Geetha, "Building language models for Tamil speech recognition system." *Proceedings of AACC* 2004.

[19] A. Nayeemulla Khan and B. Ygnanarayana, "Development of a speech recognition system for Tamil for restricted small tasks"

[20] Comrie, B. *The World's Major Languages*. New York: Oxford University Press, 1987.

[21] Kothari, B. and J. Takeda, "Same Language Subtitling for Literacy: Small Change for Colossal Gains", *Information and Communication Technology in Development: Cases from India*, in S. Bhatnagar and R. Schware (Eds), Sage Publications, 2000.

[22] Huenerfauth, M. *Developing design recommendations for computer interfacees accessible to illiterate users*. Masters Thesis, University College Dublin, August, 2002.

[23] Goetze, M., and T. Strothotte, "An approach to help functionally illiterate people with graphical reading aids." *Smart Graphics Symposium* UK, 2001.

[24] Parikh, T., G. Kaushik, and A. Chavan, "Design studies for a financial management system for micro-credit groups in rural india." *Proceedings of the ACM conference on universal usability*, 2003.

[25] Castro-Caldas, A., K. M. Petersson, A. Reis, S. Stone-Elander and M. Ingvar. "The Illiterate Brain." *Brain* 121, 1053-1063. 1998.

[26] Petersson, K. M., 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, 2000.pp. 364-382.

[27] Gould, J. D., Conti, J., & Hovanyecz, T., "Composing letters with a simulated listening typewriter." *Communications of the ACM* 26, 1983, pp. 295-308.

[28] Plauché, M., and M. Prabaker, "Tamil Market: A Spoken Dialog System for Rural India," *Working papers in Computer-Human Interfaces*, April 2006.

[29] Butterworth, B., M. Zorzi, L. Girelli, A.R. Jonckheere, "Storage and Retrieval of Addition Facts: The role of number comparison." *The quarterly journal of experimental psychology*. 54A(4), 1005-1029, 2001.

[30] Davel, M. and E. Barnard, "The Efficient Generation of Pronunciation Dictionaries: Human Factors during Boostrapping" In *Proceedings of the 8th International Conference on Spoken Language Processing*, Korea, 2004.

[31] Davel, M. and E. Barnard, "The Efficient Generation of Pronunciation Dictionaries: Machine Learning Factors during Boostrapping" In *Proceedings of the 8th International Conference on Spoken Language Processing*, Korea, 2004.

[32] Castells, M. *The Power of Identity*. Malden, Mass: Blackwell.

[33] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm,"in *ASRU'03*, US Virgin Islands, USA, Dec. 2003.

[34] Gordo, B., "Overcoming digital deprivation." *IT&Society*, Vol. 1, Issue 5, Summer, 2003.