A Hindi Speech Recognizer for an Agricultural Video Search Application

Kalika Bali
Microsoft Research
Bangalore, India
kalikab@microsoft.com

Sunayana Sitaram
Carnegie Mellon University
USA
sunayana.sitaram@gmail.com

Sebastien Cuendet
École Polytechnique
Fédérale de Lausanne
Switzerland
sebastien.cuendet@gmail.com

Indrani Medhi
Microsoft Research
Bangalore, India
indranim@microsoft.com

ABSTRACT

Voice user interfaces for ICTD applications have immense potential in their ability to reach to a large illiterate or semi-literate population in these regions where text-based interfaces are of little use. However, building speech systems for a new language is a highly resource intensive task. There have been attempts in the past to develop techniques to circumvent the need for large amounts of data and technical expertise required to build such systems. In this paper we present the development and evaluation of an application specific speech recognizer for Hindi. We use the Salaam method [4] to bootstrap a high quality speech engine in English to develop a mobile speech based agricultural video search for farmers in India. With very little training data for a 79 word vocabulary we are able to achieve >90% accuracies for test and field deployments. We report some observations from field that we believe are critical to the effective development and usability of a speech application in ICTD.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: *Evaluation/methodology*, *Interaction styles (e.g., commands, menus, forms, direct manipulation)*, *Natural language*, *Voice I/O*

General Terms

Measurement, Performance, Design, Human Factors,

Keywords

Speech Recognition, Mobile Application, Hindi, Less-resource language.

DEV'13, January 11–12, 2013, Bangalore, India.
Copyright 2013 ACM 978-1-4503-1856-3 11/01/2013 …$15.00.

1. INTRODUCTION

The potential of voice interfaces for ICTD applications have been widely discussed in a number of studies where speech technologies such as automatic speech recognition (ASR), spoken dialogue systems, and text to speech (TTS) applications are applied to problems in the developing world [19]. The primary benefit of using Spoken Language Technologies (SLT) is in their reach to a large illiterate or semi-literate population in these regions where text-based interfaces are of little use. Further, many of the countries in the developing regions, for example, India and South Africa, are multilingual societies where the official language may be quite distinct from the languages spoken in the region. A Voice User Interface using SLTs can potentially cater to the needs of speakers of less-resourced languages including those without a formal script.

Around 60% of the world's 5 billion mobile phone subscribers live in the developing world [20].The high penetration of mobile phones in the developing world provides the ideal opportunity to use SLT to access ICTs. Mobile phones offer many advantages over any PC-based interface. Many users are already familiar with the device and hence less intimidated using it. Mobile network connectivity also covers a larger area than internet and costs much less in terms of power as well as subscription charges.

While SLT deployed on the mobile phone has great potential to benefit end users, an area that has received less attention is that of its use by organizations working in these regions. A number of NGOs use video and other multimedia content for training and dissemination of information. Digital Green [6] has used agricultural extension videos effectively for training farmers in rural India. Studies have shown the benefits of using SLT and multimedia content [13, 17, 18]. Ramchandran et al [13] show in their study on India's rural maternal health system that screening short videos on mobile phones helped health workers engage women in villages much more. Similarly, video content has been used for educational outreach programmes and applications [8, 14] with varying success. The use of such multimedia content by NGOs could greatly benefit from SLT deployments for indexing and searching such content on the phone.

The biggest challenge that faces any such deployment remains the absence of fundamental speech technologies for the languages used in the developing regions. As pointed out by Qiao et al [12], while the developing world would be the biggest beneficiary of

SLT applications, there are hardly any commercial grade ASR and TTS systems available for these languages. There have been attempts however, to adapt existing speech systems to under-resourced languages with minimal training data [11] or cross-language mapping [4, 12, 17]. These methods have shown some encouraging results even though they are usually limited to very small application-specific vocabulary. Salaam [4, 12] in particular, has shown promising results even though their data has been confined to experiments in a laboratory setting.

In this paper, we discuss the development of a Hindi speech recognition module for VideoKheti, a mobile video search application for farmers. In the next section we describe the application with respect to speech navigation of the menu. Section 3 discusses some of the challenges in building an ASR for a low resource language like Hindi, and describes the SALAAM algorithm as a means for overcoming some of these challenges for small vocabulary applications. The development of the speech module is presented in Section 4. We achieve results of above 90% accuracies on data collected from end-users as well as a field-study. We conclude the paper with a discussion on some of the real challenges in building and deploying a small vocabulary speech based system for a low-literate novice population.

2. VIDEOKHETI FOR FARMERS

Previous research [9] has shown that videos with a full contextual presentation of a new usage prove an effective learning medium for novice low-literate population. E.g. Digital Green, an NGO in India, focuses on training small and marginal farmers through video screenings of targeted agricultural information. The videos typically feature a local farmer demonstrating an agricultural technique to a mediator. Once shot, the videos are edited at a Digital Green facility and the mediator goes from village to village in order to screen them to farmer groups. The videos are usually stored on an SD card and projected against a wall by means of a handheld pico-projector. There are two main issues with this set-up. First, once the mediator has left the village, there is currently no way for the farmers to watch the videos again. Second, there are logistical and usability challenges faced by the mediator-- he/she has to physically visit the Digital Green facility to fetch the SD card, and then access them directly on the card without a way to filter them. A mobile based application that would let the mediator screen a video on the phone would greatly benefit the NGO as well as allow the farmers to replay videos in the absence of the mediator. VideoKheti is a multimodal application that allows low-literate farmers and mediators to access videos related to farming on a mobile phone.

VideoKheti was developed with the co-operation of Digital Green and used 147 videos produced by DG for farmers in the state Madhya Pradesh in Central India. These videos were arranged hierarchically into four levels by agriculture experts with four category labels based on their content, namely, crop, crop cycle, action type and method. A hierarchical navigation tree allowed the users to access a list of relevant videos by making a maximum of four choices. Figure 1 shows a schematic example of the navigation tree for the application with the active vocabulary at each level. The application was built as a part of a larger study that explored both Graphical as well as Speech User interfaces for low literate novice users. In this paper we specifically focus on the building and testing of the speech module for the application.

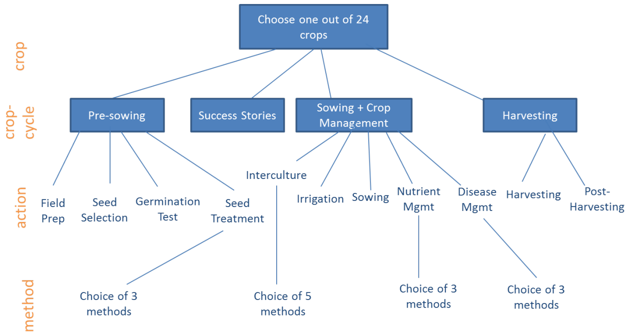


Figure1: Schematic representation of the navigation menu of VideoKheti

Figure 2 shows the text-free user interface of VideoKheti. Recorded prompts by native Hindi speakers asked the user about the information required by them. Except for the crop class, recorded prompts also informed the user of the choices available to them on a particular page. Each page also displayed the graphical representation of the choices available. A beep signaled that the system was in the "listening" mode where the system started recording user response automatically for a 7 second interval. On the correct recognition of the response the system displayed the page for the next level until the desired videos were displayed. An error loop was also implemented for incorrect recognition or out of vocabulary input that allowed the users to input their choice again. A help module repeated the choices available and provided example inputs. Given the constraints on the speech recognition discussed later in the paper, the navigation was essentially system directed.

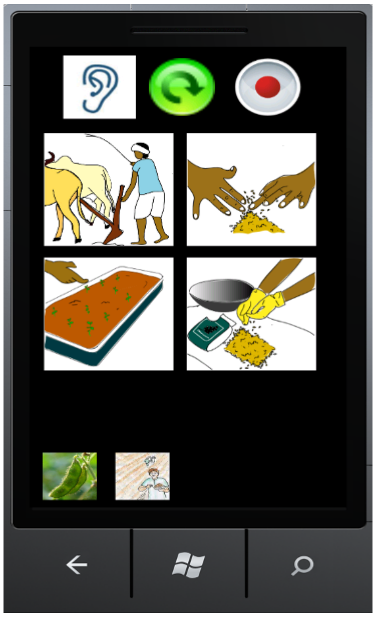
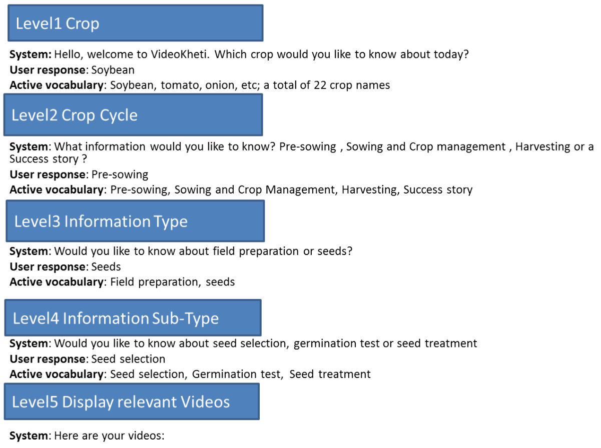


Figure2: Text Free UI of VideoKheti

Figure 3 shows an example of a path taken by a user through the menu to access the videos of her choice. The vocabulary active at each level indicates the words accepted by the system at each level.



Level1 Crop

System: Hello, welcome to VideoKheti. Which crop would you like to know about today?
User response: Soybean
Active vocabulary: Soybean, tomato, onion, etc; a total of 22 crop names

Level2 Crop Cycle

System: What information would you like to know? Pre-sowing , Sowing and Crop management , Harvesting or a Success story ?
User response: Pre-sowing
Active vocabulary: Pre-sowing, Sowing and Crop Management, Harvesting, Success story

Level3 Information Type

System: Would you like to know about field preparation or seeds?
User response: Seeds
Active vocabulary: Field preparation, seeds

Level4 Information Sub-Type

System: Would you like to know about seed selection, germination test or seed treatment
User response: Seed selection
Active vocabulary: Seed selection, Germination test, Seed treatment

Level5 Display relevant Videos

System: Here are your videos:

Figure 3: An example path taken by a user through the menu

3. CHALLENGES IN BUILDING HINDI ASR

Hindi is the primary official language of India and official language of ten of its states. According to the 2001 Census [3], 41% of the Indian population speaks Hindi as its first language, and more than 70% Indians can understand and speak Hindi to a certain level. Hindi is the lingua franca in many non-Hindi speaker states, such as the north eastern Indian state of Arunachal Pradesh, and is second most spoken language after Bengali in Andaman Islands and north eastern states [3]. 49 dialects of Hindi are listed in the census of India 2001 though the actual number far exceeds the list.

Given the large number of people speaking Hindi and its dialects and the potential role that SLT can play in ICTD in the region, the effort on developing Hindi ASR is negligible. There have been research prototypes [1, 15] and application specific systems [2] in the past but the only large vocabulary system developed has been the IBM dictation system for Hindi [10]. Thus, for an application developer wanting to use speech as an input modality there is no off the shelf system available and the alternative is to build a system from scratch; no mean feat given the resources required in terms of data and expertise.

Building an ASR for a language like Hindi would require in addition to the speech data from a large number of speakers of many dialects, a pronunciation lexicon and language models. As none of these resources are currently available, it makes the task even more difficult especially if the goal is to provide a natural speech interface where a user can expect to say anything and be understood. Thus, a viable solution is to work with adaptation techniques that allow a speech application to be developed using existing off the shelf commercial grade systems in another language. These techniques while not appropriate for a free flow user-directed dialog system, can work quite adequately for small application specific vocabulary where the system needs to recognize a limited set of words or phrases. Even then some expertise is required for creating language specific lexicons.

Many previous researchers have attempted to circumvent the need for copious amounts of data and technical expertise to build ASRs for new languages where such resources are unavailable. Some of the notable efforts in the recent past include the GlobalPhone project [16] where a multilingual database is used to train universal phone models that are then used as the basis for training language-specific models with little data. Also, Plauche et al [11] demonstrate a way to adapt existing ASRs to a new language by bootstrapping the acoustic models with some data from the target language. While both these efforts require very little data as compared to building a large vocabulary speech recognition system from scratch, the level of expertise and resources required are still substantial. The final accuracies achieved in these methods are reasonable though not fully adequate for a real world field deployment.

The Salaam approach is based on all-phone decoding [12] where target language words are decoded using the source language phonemes. Where Salaam has an advantage over other methods is that it uses the cross-language phoneme mapping with a high quality commercial engine which are black-boxes that do not allow a means for all-phone decoding.

Given a word in the target language, the Salaam algorithm iteratively generates possible phoneme sequences for that word. The initial seeding uses a small number of recorded samples of the word (or word type). A commercial ASR system then outputs a list of probable decoded phoneme sequences. The first phoneme of each of the sequence in the previous output is fixed as the candidate for the final result and a new set S1 of phoneme sequences is built. Another candidate phoneme sequence is then decoded for each of the sequence in S1. The process is repeated till no new phoneme sequences are output by the ASR for three cycles. The decoded phoneme sequence with the high score is then taken as the best candidate for that word. A later improvement on the algorithm [4] uses heuristics to avoid errors due to words with similar pronunciations. A combination of rank and confidence scores is used to select best possible sequences. Candidate sequences which are never used for recognition in the test runs as well as candidates that are matched to incorrect word types are rejected. The authors report around 5% decrease in word error rate using this discriminative training.

4. VIDEOKHETI SPEECH MODULE

As mentioned earlier, VideoKheti was designed to use a speech based navigation that allowed farmers to access farming related videos. In the absence of a speech recognition system for Hindi and any of its dialects the choice for developing a speech recognition module for the application was between building a limited vocabulary ASR from scratch or using one of the techniques reported in the literature for bootstrapping an existing high quality ASR in another language. The Salaam approach [4, 12] where an existing high quality speech recognition engine is used to automatically generate pronunciations through cross-language phoneme mapping seemed a reasonable choice for our application given that a) it required very little target language data both in terms of number of speech samples per user as well as the number of users, b) the results reported on Urdu[1] word recognition using Salaam were quite promising. Salaam used a vocabulary of 100 words with 5 samples per speaker for 5 speakers. The size of the vocabulary was comparable to that of VideoKheti. Unlike Salaam that has been primarily used in a laboratory setting as a proof of concept, VideoKheti is designed to be used in field with actual end-user population.

Figure4: Data collection for VideoKheti

4.1 Vocabulary and data collection

The data collection for building the speech module was done in the villages of the Rajgarh district in Madhya Pradesh, home to the target population of the application. While the Digital Green screenings in this area are done in standard Hindi (the official language of the state), the people of the region spoke Malvi, a sub-dialect of the Rajasthani branch of Hindi [7]. Almost all speakers spoke some standard Hindi.

A recording application was deployed using a Windows Phone mobile where the user was prompted the first time by a recording done by a native Hindi speaker. In order to prevent any pronunciation effects from mimicking or by consecutive repetition of the same word, the second prompt was in the user's own voice. The order of the prompts was randomized and a 4 second pause was inserted between the end of the prompt and the start of the recording.

A total of 3500 samples were recorded for the 79 word-vocabulary by 13 males and 11 females. An average of 2 samples per user was collected. It should be mentioned that "word" here is used to denote both single words as well as 2-4 word long phrases. The recording was done in an open environment and this meant that it was not possible to exclude background noise The vocabulary of the application included many technical jargons that were used in the original videos and were being promoted by the organization. This vocabulary was designed in collaboration with agriculture experts at Digital Green, However, many speakers, especially women, were not familiar with these terms. This and the 4 second lapse between the prompt and the recording meant that some speakers could not remember the word to be recorded. In such cases, the speakers were allowed to leave the recording blank.

4.2 Generating pronunciations using SALAAM

To train the Salaam algorithm for our application we required the audio recordings from the data collected and a list of the words in the vocabulary. We used a vocabulary of 79 words consisting of single words and short phrases. Since we had only two samples per person, we trained the algorithm with ten samples from multiple speakers; otherwise too often the Salaam algorithm would stop if the two samples happened to be of lower quality. We split up the words into 8 sets to parallelize the training and speed up the process. Salaam generated 10 pronunciations for each word in the vocabulary for each set of audio files used for training. Salaam could not come up with pronunciations for 8 words out of the 79 words in our vocabulary. We observed that these words were typically the longer phrases. A native Hindi speaker manually created 10 pronunciations for each of these words. These pronunciations were then merged with the pronunciations generated automatically for the other words.

We ran discriminative training described in [4] which reduced the number of rules in the grammar from ~600 to ~500. However, we observed that for words that were very close in pronunciation, for example, "barchi" and "batli" (names of fodder crop), many rules were removed especially if the samples were of lower quality. The discriminative training considered a pronunciation as "bad" when it was matched to a wrong word. When the samples were of lower quality then the matches were more random leading to the removal of many more rules, including sometimes the correct one.

A native speaker went through all the collected data manually and removed the audio files that had incorrect words or that only had silences because of blank recordings

We tried various combinations of speakers for our training sets and ended up using all the pronunciations generated by them for the grammar used in the audio search application. The grammar was trained on 20 speakers, 10 females and 10 males. Since each training set of 5 people had 10 rules per word, this could lead to up to 40 pronunciation rules per word. We removed all duplicate pronunciations from the grammar. The resulting grammar contained 2817 rules for the 79 words, that is, about 35 rules per word present in the vocabulary. Each word in the vocabulary had a minimum of 10 rules in the grammar.

5. RESULTS

In this section, we describe results of running the Microsoft Speech Server for Indian English with grammars generated by

[1] Urdu and Hindi are considered structurally the same language by many linguists, differing only in their scripts and the use of certain high-register vocabulary.

SALAAM both on collected held-out data and in the voice search application in the field. Some samples were rejected by the speech recognizer due to poor quality or noise. In addition, we also deployed heuristics that rejected samples in two cases: if the ASR confidence was less than 0.5 or if the confidence was between 0.7 and 0.5 and the difference between the confidences of the first and second best candidate was greater than 0.1. The accuracy of recognized words was the accuracy of the ASR on the words that it accepted. The overall accuracy took into account the fraction of words rejected by the ASR, which is the rejection rate.

Figure 5: Field testing of VideoKheti

5.1 Results from the collected data

Initially, we ran the SALAAM algorithm on a training set of 5 males and 5 females, and tested the grammar on the rest of the speakers, 7 males and 6 females. Then, we used the same grammar and tested on males and females separately. Table 1 shows the results of these three experiments. The letters 'm' and 'f' denote male and female respectively.

Training	**Testing**	**Overall accuracy**	**Accuracy of recognized words**	**Rejection rate**
10 m+f	13 m+f	0.71	0.93	0.23
10 m+f	6 f	0.60	0.88	0.32
10 m+f	7 m	0.80	0.95	0.16

Table 1: ASR accuracies on male and female speech

The overall accuracy and accuracy of recognized words was higher for males than females. The rejection rate was higher for women than for men. So, we tried training SALAAM using speech from 5 female speakers. Table 2 shows the comparison of results on the same 6 women when trained using speech from both men and women, and only women.

Training	**Testing**	**Overall accuracy**	**Accuracy of recognized words**	**Rejection rate**
10 m+f	6 f	0.60	0.88	0.32
5 f	6 f	0.53	0.87	0.40

Table 2: ASR accuracies on female speech

As shown in Table 2, the accuracies on female speech were lower when the grammars were trained on female speech only. This could be because there was more hesitation in female speech, and in general, female speech was not well-articulated compared to male speech.

We also manually removed some audio files that contained the wrong word or only silences. We did not remove audio files that contained background noise, hesitation or mispronunciations because we expected to deal with these during field testing.

Training	**Testing**	**Overall accuracy**	**Accuracy of recognized words**	**Rejection rate**
10 m+f	13 m+f	0.73	0.95	0.23
10 m+f	6 f	0.61	0.90	0.32
10 m+f	7 m	0.82	0.97	0.16

Table 3: ASR accuracies on clean data

Results on the clean data are shown in Table 3. In comparison to results on the original data shown in Table 1, overall accuracies and accuracies of recognized words are slightly higher on clean data, which is to be expected. The rejection rate remained the same as before.

We ran discriminative training [4] on the grammar generated on the original data. Table 4 compares the ASR accuracies on the original grammar and the grammar generated after running discriminative training. There was a slight increase in the accuracy of recognized words, but the rejection rate was also higher for the discriminative training grammar.

In tests on prompted data we got accuracies of >93% on male and female speech, with overall accuracies of >71%. Accuracies on male speech were consistently higher than female speech.

Training	Testing	Overall accuracy	Accuracy of recognized words	Rejection rate
10 m+f	13 m+f	0.71	0.93	0.23
10 m+f, discriminative	13 m+f	0.71	0.96	0.26

Table 4: ASR accuracies on original grammar and discriminative training grammar

5.2 Results from the field study

So far, we looked at results on recorded prompted speech and now we shift our attention to the actual application deployed in the field. We trained a grammar on 20 males and females and used it with the Microsoft Speech Server in the field. The field study was carried out on 20 farmers (12 males and 8 females) in the Rajgarh district of Madhya Pradesh. All but one participant were regular users of mobile phone. However, none had used a touch based phone before. An attempt was made to balance the user group across age and education level. However, the level of education presented some challenges as the users ranged from those with no formal education to some with undergraduate level degrees.

The study was conducted in a closed room environment to minimize noise and other distractions. Figure 5 shows the field test in progress. A Samsung GT-I8350 running a Windows Phone 7.5 OS was used for both the data collection and the field test. A wireless network was setup locally on a Lenovo T400 laptop which also ran the Microsoft Speech Server.

We tested the application on the farmers one by one by giving them pre-specified tasks. The researchers conducting the study spent a few minutes initially training the farmers on the application where the researcher explained the goal of the application and demonstrated a simple task using speech based navigation. Each user was asked to complete four scenario-based tasks, the end of goal of each of which was to find a related video. A typical scenario involved. Some example scenarios are shown in figure 6. All speech data was logged by the system as the users carried out the tasks. A researcher collected the demographic data as well as user feedback pre and post experiment, respectively.

Task 1: You notice that some of your soybeans crop have a disease due to some insect. Find a video that will demonstrate how to do insect control on your crops.

Task2: You now want to cultivate oranges. You have the seeds and the field is ready, but you have never planted oranges before. Find a video that will explain to you how to do it.

Figure 6: Example tasks for field-testing

Audio recordings from the field not only contained words spoken by the farmers, but also contained a large number of silences, Out of Vocabulary words, discussions between the farmers and the researchers and noise. Because of this variation in the quality of the data, it is difficult to get an accurate estimate automatically of how well the ASR did in the field. So, we manually went through the ASR logs and all the audio files collected in the field to identify legitimate words spoken by the farmers. In some cases, the farmers said phrases which contained one or more vocabulary words, but were not exact phrases in the vocabulary. We labeled these as "partial phrases. A total of 157 utterances were selected with 17 of them being partial phrases. Table 5 shows the results on field data including and excluding partial phrases, when compared with the best results we obtained on prompted data.

Training	Testing	Overall accuracy	Accuracy of recognized words	Rejection rate
Including partial phrases, 20 m+f	20 m+f	0.56	0.90	0.38
Excluding partial phrases, 20 m+f	20 m+f	0.65	0.94	0.30
Best result on prompted data, 13 m+f	13 m+f	0.71	0.96	0.26

Table 5: Results from the field

We observe that the overall accuracy and accuracy of recognized words of the ASR was a little lower when we included partial phrases. The heuristics used for the field study set the confidence threshold slightly higher than that used for the prompted data. Thus, a candidate was rejected in all cases with confidence smaller than 0.5; if between 0.5 and 0.7, the candidate was rejected if difference with the second best result is smaller than 0.1 This resulted in comparatively higher reject rates for the field study. However, the accuracy of recognized words >90% and comparable to the results obtained on prompted data. What this implies is that many more pronunciations were deemed unacceptable by the system and hence, there was a slight increase in the number of false negatives leading to a drop in overall accuracy. However, since the accuracy of the accepted (correct) candidates was high, the effect on the usability of the system in the field was negligible.

Cuendet et al [5] provide a more detailed discussion on the user study vis-à-vis usability and user feedback on the design and navigation modalities of the application.

6. CHALLENGES AND ISSUES

While the ASR accuracies of the field testing of VideoKheti are reasonably high the actual building and deploying of a speech based video search in Hindi for farmers posed a number of issues which need to be kept in mind while designing any speech based application for such a target population.

6.1 Data

Collection of data for training and testing ASR models always is a resource intensive activity with its own inherent challenges related to noise, speaker variability, device, naturalness, accent etc. In this case, the data collection took place in an open environment as we wanted to replicate the environment in which a mobile phone would be typically used by the farmers or an NGO mediator. The obvious effect of this was the different types of noise that we encountered in our data. This included people chatting in the background, an occasional vehicle passing by, sounds made by cattle and other animals as well as persistent low frequency insect sounds.

Another aspect of data collection has to do with the dialect used by the users. As our application was built on top of the Digital Green video database, we used the technical terms employed in those videos. These terms were typically in a highly Sanskritised standard Hindi. Most users spoke the Malvi dialect and even though they could converse in standard Hindi, for many, these terms were unfamiliar and difficult to pronounce and remember. How feasible it is to customize video content and applications to all dialects of a language, especially languages like Hindi which have over a hundred dialects, is debatable. However, collecting data for unfamiliar terms did cause problems with low quality data due to blank recordings, incorrect words, incomplete words, hesitations and bad articulations.

6.2 Technical Issues

The Salaam algorithm worked quite well with our data both from prompted speech and the field. The accuracy rates for the accepted words were above 90%. However, the rejection rates were quite high. A large part of this had to do with the quality of data as described above. But since Salaam is aimed at the developers of SLT for development, low quality data and unexpected noise is something that we have to account for.

Another issue we faced was that of similar sounding words where for relatively low quality data, the discriminative training actually threw away many of the pronunciation candidates. This meant that many times the correct pronunciation was also discarded. Also, for longer phrases of >4 words, Salaam did not work too well and sometimes did not generate any pronunciation candidates at all. For such phrase we had to resort to manual creation of the pronunciations by an expert. This might be language and application specific but again a real issue that should be kept in mind while designing a speech application.

6.3 Socio-cultural Aspects

One of the consistent results we observed was the low levels of accuracies and higher rejection rates for female users as compared to male users. An analysis of the female data showed that some of the issues discussed in Section 6.1 are more prevalent for female users than male. A look at the demographics as well as field observation points us to the role socio-cultural aspects play in this. While women in the farming community are equally if not more involved in the actual farming practices, they tend not to interact with the external forums like the NGOs, government agencies, etc. In general, they also tend to have less schooling than the men and hence be less familiar with the jargon used.

Another observation with regard to the female results has to do with the position of the women in this community. The women in general, were shy and less confident with lower education levels. They were very aware of people gathered outside the room especially male relatives. This made them more hesitant and less articulate. Even when they did know a particular term, they did not speak out loudly and clearly in front of others. It should be noted that the researchers conducting the field test and collecting the demographic data from the women were of the same gender. While the women were not incapable of using the device and completing the task, in fact, almost all of them were able to complete all the tasks, they felt less comfortable speaking into the phone or in some cases, even touching the device. This gender divide in target communities which can seriously affect the impact and uptake of an SLT application can only be observed through a field study and is hard to catch in a laboratory setting. Our results provide consistent evidence of how the accuracies of a system can suffer due to such socio-cultural differences.

7. CONCLUSIONS

In this paper, we built a Hindi speech recognition module for VideoKheti-a video search application for farmers using the Salaam method. The application allows farmers to access farming-related videos on a mobile phone using a speech based navigation system. We were able to achieve above 90% accuracies for both the prompted data as well as in-field deployment. We found that the Salaam method works well for system directed isolated word applications and did surprisingly well in field conditions.

We used the Microsoft Speech Server for English in our study. One direction to explore would be to use another ASR for a language that might be closer to Hindi. Also, work towards handling longer phrases as well as low quality noisy data would help to make speech applications more robust and hence, useful in the SLTD context.

8. ACKNOWLEDGMENTS

We would like to thank Digital Green and the field mediators and coordinators at ACCESS, Madhya Pradesh for their help in conducting the field study.

9. REFERENCES

[1] Aggarwal, R. K., and Dave, M. "Implementing a Speech Recognition System Interface for Indian Language," *Proceedings of the IJCNLP-2008 Workshop on NLP for Less*

Privileged Languages, Hyderabad, January 2008, pp. 105-112.

[2] Arora, S., Saxena, B., Arora, K. and Agarwal, S.S. “Hindi ASR for Travel Domain,” *Oriental COCOSDA 2010 Proceedings*, Centre for Development of Advanced Computing, Noida, 24-25 November 2010.

[3] Census of India website – Census 2001, Statement 3, http://censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language.html

[4] Chan, Hao Yee and Rosenfeld, R. Discriminative Pronunciation Learning for Speech Recognition for Resource Scarce Languages. *Proc. ACM DEV 2012, Annual ACM Symposium on Computing for Development*, March 2012,Atlanta, GA.

[5] Cuendet, S., Medhi, I., Bali, K. and Cutrell, E. “VideoKheti: Making Video Content Accessible to Low-Literate and Novice Users” Submitted to *The ACM SIGCHI Conference on Human Factors in Computing Systems, 2013* (under review)

[6] Gandhi, R., Veeraraghavan, R., Toyama, K., and Ramprasad, V. Digital green: Participatory video for agricultural extension. *Information and Communication Technologies and Development, 2007*. ICTD 2007.

[7] John, Matthew. "The Malvi-speaking people of MadhyaPradesh and Rajasthan: a sociolinguistic profile." *SIL Electronic Survey Reports 2009-011: 280*. http://www.sil.org/silesr/abstract.asp?ref=2009-011

[8] Kumar, A., Reddy, P., Tewari, A., Agrawal, R., and Kam, M. Improving literacy in developing countries using speech recognition-supported games on mobile devices. *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI ’12, ACM (New York, NY, USA, 2012), 11491158.

[9] Medhi, I., and Toyama, K. Full-context videos for first-time, non-literate PC users. In *Proc. 2007 International Conference on Information and Communication Technologies and Development* (2007), 1–9.

[10] Neti, C., Rajput, N. and Verma, A. “A Large Vocabulary Continuous Speech Recognition System for Hind,” *IBM Research and Development Journal*, September 2004.

[11] Plauche, M., Nallasamy, U., Pal, J., Wooters, C., & Ramachandran, D. 2006. Speech Recognition for Illiterate Access to Information and Technology. *Proc. 115. International Conference on Information and Communications Technologies and Development*, 2006.

[12] Qiao, F., Sherwani, J.,Rosenfeld, R. Small Vocabulary Speech Recognition for Resource-Scarce Languages. *Proc. ACM DEV 2010, Annual ACM Symposium on Computing for Development*, December 2010, London, UK.

[13] Ramachandran, D., Canny, J., Das, P. D., Cutrell, E. Mobile-izing health workers in rural India, *Proceedings of the 28th international conference on Human factors in computing systems*, April 10-15, 2010, Atlanta, Georgia, USA [doi>10.1145/1753326.1753610]

[14] Sahni, U., Gupta, R., Hull, G., Javid, P., Setia, T., Toyama, K., and Wang, R. Using Digital Video in Rural Indian Schools: A Study of Teacher Development and Student Achievement. Annual Meeting of the American Educational Research Association. March 2008.

[15] Samudravijaya,K. “Hindi Speech Recognition,” *Journal Acoustic Society of India*, Vol. 29, No. 1, 2009, pp. 385-393.

[16] Schultz, T., Westphal, M., and Waibel, A. The globalphone project: Multilingual lvcsr with janus-3. In *Proc. SQEL*, pages 20–27, 1997.

[17] Sherwani, J., Ali,N., Mirza, S., Fatma, A., Memon, Y.,Karim, M., Tongia, R. and Rosenfeld, R. Healthline: Speech-based access to health information by low-literate users. In *Proc. Information and Communication Technology and Development*, 2007.

[18] Sherwani, J. Speech Interfaces for Information Access by Low Literate Users. PhD thesis, CMU, May 2009.

[19] Weber, F., Bali, K., Rosenfeld, R., and Toyama, K. Unexplored Directions in Spoken Language Technology for Development, in *Proceedings of The 2nd IEEE Workshop on Spoken Language Technology. 2008*, December 2008

[20] UNCTAD. 2008. Information economy report 2007-2008: Science and technology for development—The new paradigm of ICT. In *United Nations Conference on Trade and Development*.