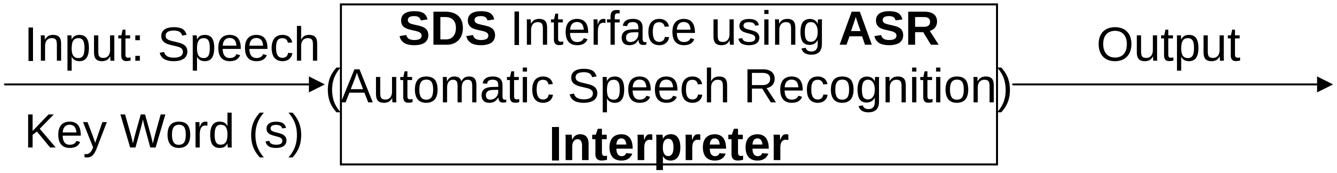
Speech Interfaces for Equitable Access to Information Technology

Review of a low cost, scalable, speech-driven application providing agricultural info to rural India's farmers.

Anastasios Dimas

Images from Plauché, Madeline and Udhyakumar Nallasamy. "Speech interfaces for equitable access to information technology." *Information Technologies and International Development* 4, no. 1 (2007): 69-86.
Courtesy of the USC Annenberg School of Communications.

Speech Interfaces or Spoken Dialog systems (SDS)



Input: Speech

Key Word (s)

SDS Interface using **ASR** (Automatic Speech Recognition) **Interpreter**

Output

Automated Telephony

Pros:

•enhancing access to IT services for the visually or mobility impaired by replacing / enhancing traditional computer input (mouse, Keyboard) and output (screen).

Cons:

•Prohibitive cost of computing devices

•Required IT infrastructure

•Software design that assumes:

-literacy

-computer savvy

Speech Interfaces or Spoken Dialog systems (SDS)

Mobiles Phone applications in Rural India provide **info** on:

- Health
- Weather
- Employment
- News
- Agriculture

Pros:

- Affordable
- Infrastructure is more readily available
- Used extensively

Speech Interfaces or Spoken Dialog systems (SDS)

Challenges for SDS applications in rural India:

•Noisy Environment

•Multilingualism (over 420 languages spoken in India)

•Dialectal Variation (dialects change dramatically within a few hundreds of Km)

•Annotated corpora nor other costly linguistic resources exist for ASR use

•Design techniques developed for accessing sociocultural models developed for the Western world are not effective in poor communities
-Leisure & Formal education are spare

•Local content created by local providers is rare
-News
-Events
-Innovations

Radio and TV are less effective in influencing people to improve their practices in health agriculture etc than traditional oral methods of info dissemination

Speech Interfaces or Spoken Dialog systems (SDS)

Design Requirements for successful SDS:

Front-end dialogue interface should be:

- Interactive
- Easily adoptable
- Able to accommodate illiterate users
- Able to accommodate technology agnostic users

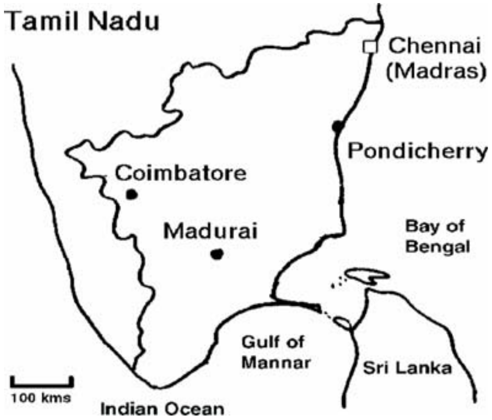
IT design and creation should involve community members. This ensures that:

- Proposed solutions meet community needs
- The best chance for technological sustainability in the community is provided
- Community partners can provide accurate, locally created info to illiterate adults

  -quickly

  -cheaply

Speech Interfaces or Spoken Dialog systems (SDS)



Plauché and Nallasamy, 2007.

Tamil Nadu

37.47% of full time workers are farmers

71.64% of marginal workers are farmers

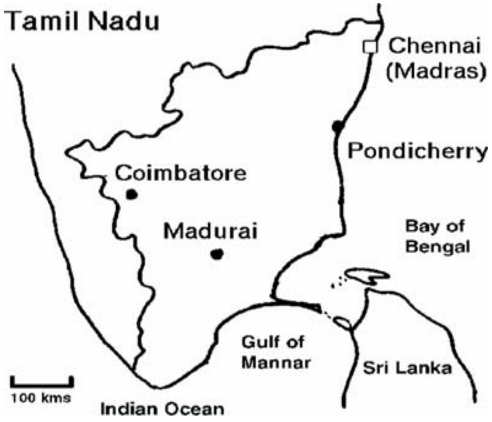
Overall 40% of the labor force in developing countries are farmers

Farmers' information needs:

- Price info (IT-based info networks can help raise the price of goods sold)
- Market info
- Techniques to improve production
  - Pest and disease prevention
  - New seeds

Speech Interfaces or Spoken Dialog systems (SDS)

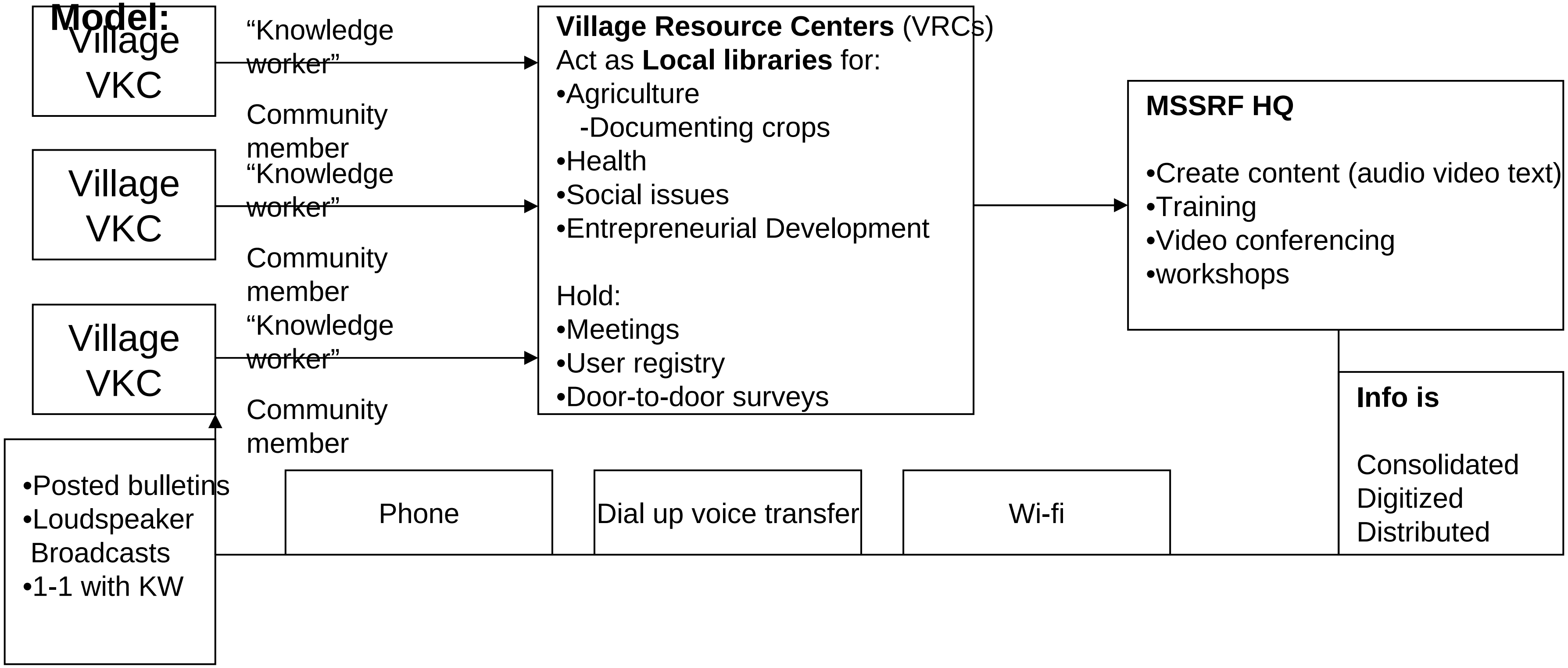
MS Swaminathan Research Foundation (MSSRF)



Plauché and Nallasamy, 2007.

Pro-nature, pro-poor, pro-women NGO fostering economic growth in rural agricultural areas. Cooperated in the study.

Model:



Village VKC

“Knowledge worker”

Community member

Village VKC

“Knowledge worker”

Community member

Village VKC

“Knowledge worker”

Community member

Village Resource Centers (VRCs)
Act as **Local libraries** for:
- Agriculture
  - -Documenting crops
- Health
- Social issues
- Entrepreneurial Development

Hold:
- Meetings
- User registry
- Door-to-door surveys

MSSRF HQ

- Create content (audio video text)
- Training
- Video conferencing
- workshops

Info is

Consolidated
Digitized
Distributed

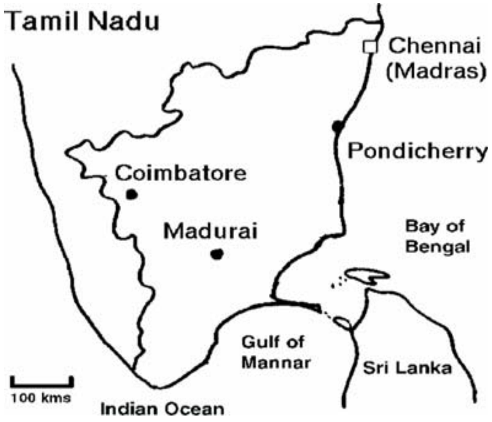
- Posted bulletins
- Loudspeaker Broadcasts
- 1-1 with KW

Phone

Dial up voice transfer

Wi-fi

Speech Interfaces or Spoken Dialog systems (SDS)



Plauché and Nallasamy, 2007.

ASR (Automatic Speech Recognition)

Algorithmically converts a speech signal into a sequence of words.

Hidden Markov Models* are trained on a large corpus of speech (training data).

Success depends on:

•the collection and annotation of the training data

•The creation of a dictionary of all possible words with all possible pronunciations in the language.

Success of 95% under optimal conditions:

•Controlled environment (quiet)

•Limited domain

•Single speaker

*A hidden Markov model (HMM) is a statistical mode in which the system being modeled is assumed to be a Markov Process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters.

Speech Interfaces or Spoken Dialog systems (SDS)

ASR (Automatic Speech Recognition)

Barriers:

To achieve human levels of recognition we require:

•Speaker-independent

•Large vocabulary

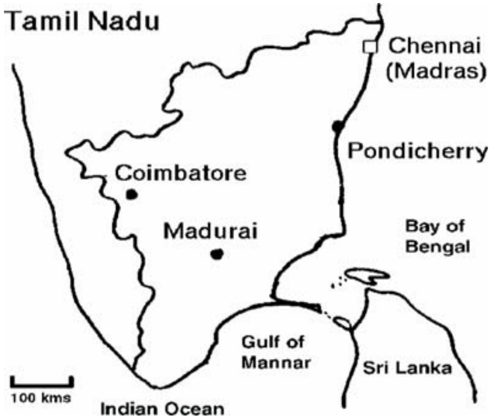
•Continuous speech recognition

These as a result lead to:

•cost of creating linguistic resources being prohibitive

•Time dedicated to acoustic training data being vast (4-70 lifetimes)

•Too much expertise required to collect training data



Plauché and Nallasamy, 2007.

Basic Principles of Speech Recognition Performance:

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

Speech Interfaces or Spoken Dialog systems (SDS)

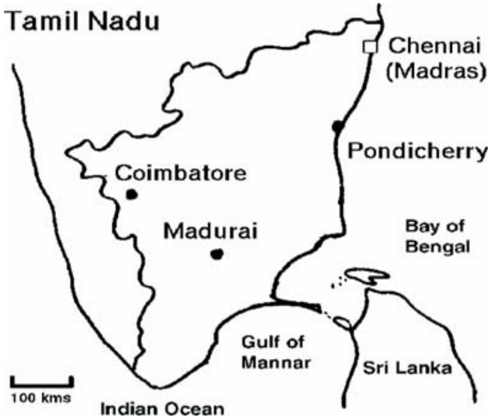
ASR (Automatic Speech Recognition)

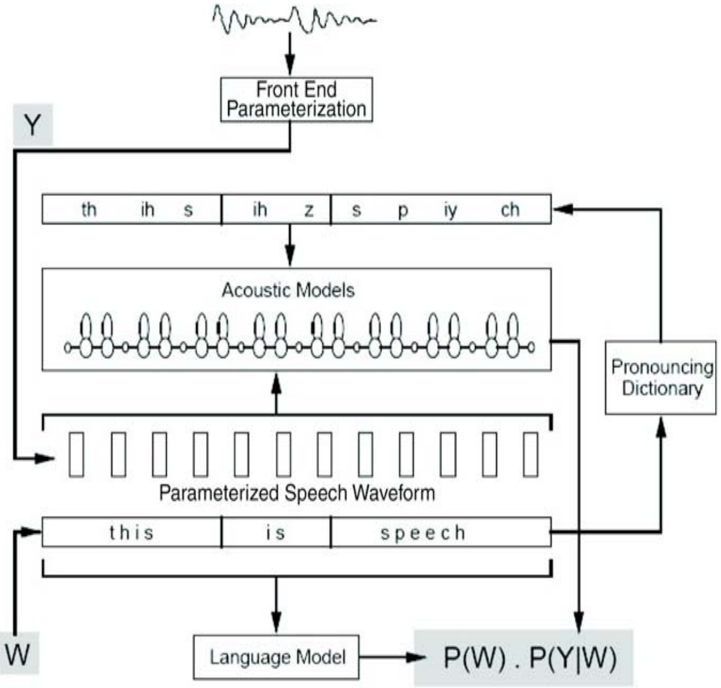
Possible Solutions:

•Simplifying the recognition task

•Adopting adaptation techniques that tune the recognizer's models to match input data.

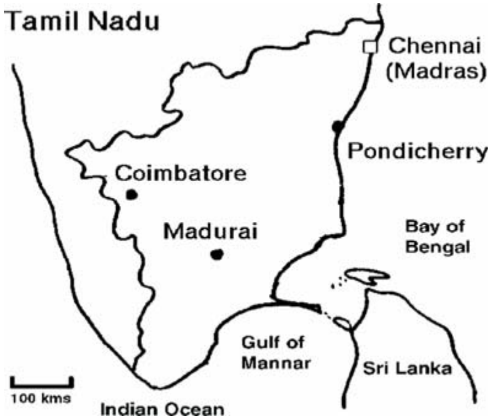
-leads to a minimal linguistic corpus required for acceptable error rates.





Images: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)



Field Study 1

•Speech recording of 77 rural villagers over 18 were conducted in 3 districts of Tamil Nadu (2004 -2005) to create adequate training data for machine recognition of a small vocabulary (<100 words). Gender education and age were balanced.

•Working alongside trusted organizations that serve the rural poor was the most efficient method for recruiting and recording villagers.

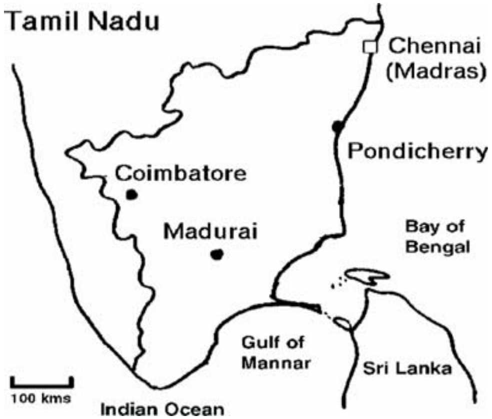
•2004 Data Collection: 30 words recorded in quiet offices via laptop and microphone

•2005 Data Collection: words for digits 0-10 using flashcards and a telephone handset with embeded microphone connected to Sony MD Walkman.

<ins>Interesting facts:</ins>
•data recordings from illiterate speakers took 6 times more!
•10000 speech samples extracted
•whole word recognizer trained on speech of 22 speakers using HTK (2004)

Images: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)



Field Study 1

Experiment 1:

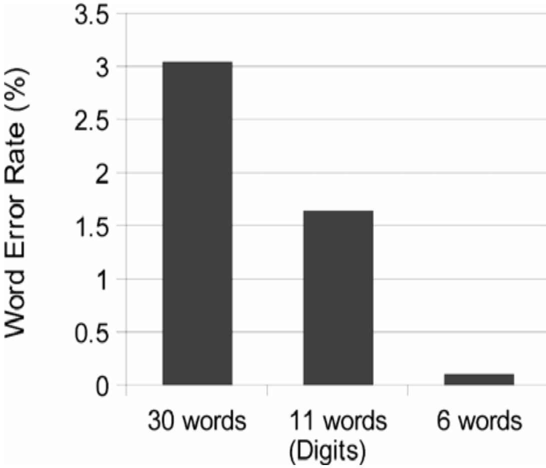
3 trials of varying complexity

- All words
- Digits only
- 6 command words

Results:

Word error rates dropped for tasks with fewer options for correct word identity. Overall error rate <2%.

An SDS of small vocabulary or limited word options per dialog node would require very little training data (<3hours) to achieve recognition



Images: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)

Field Study 1

Experiment 2:

Goal: Evaluate influence of phonetic and lexical variations on a small vocabulary recognizer.

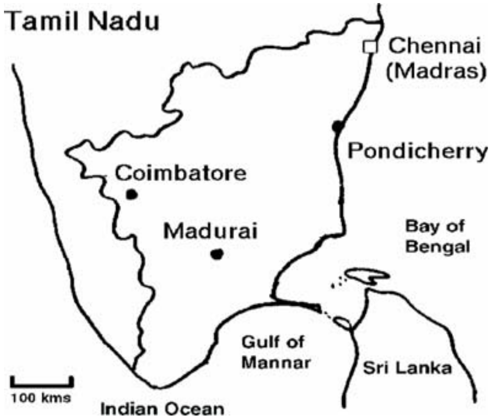
Recording words for digits in Tamil language in 3 different districts revealed that pronunciation of “7” and the choice of word for “0” varied significantly.

The trained recognizer on the speech of the 22 speakers of 2004, was used in the 3 districts of the study.

Results:

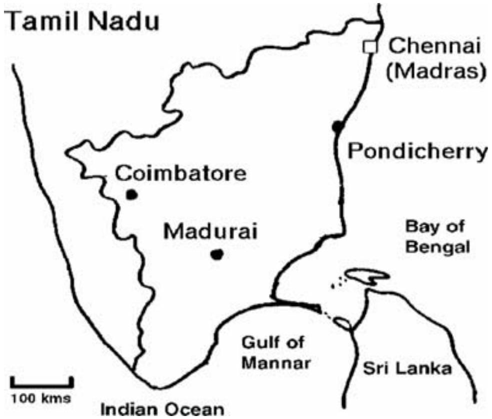
The study yielded significantly higher error rates than before.

This shows that SDS must be trained on speech from people who are potential users to ensure there will be no huge variations in dialect and choice of vocabulary between training speech and field data.



Map: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)



Field Study 1

Experiment 3:

Goal: Determine the least amount of data needed to achieve acceptable error rates for the SDS operation via simulation.

a. Organizer was trained on 1 randomly selected speaker's speech in which a second randomly selected speaker’s speech was used as input.

Results: word error rate =80%

b. Next the recognizer was trained on 2 randomly selected speakers’ speech

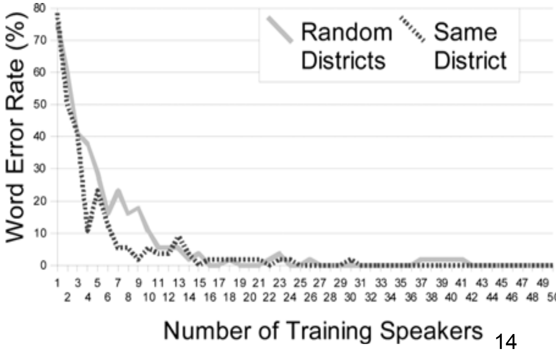
Results: word error rate dropped

c. Experiment was replicated under 2 conditions:

•More speakers were added randomly from all 3 districts.

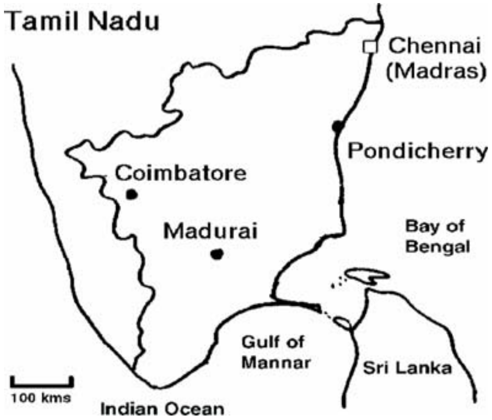
•Speakers from the test speakers district were added first.

Results: When less than 15 speakers are available for training, recognition for a given speaker is better if trained on speakers from the same district.



Images: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)



Field Study 1

Overall results:

Errors decrease with

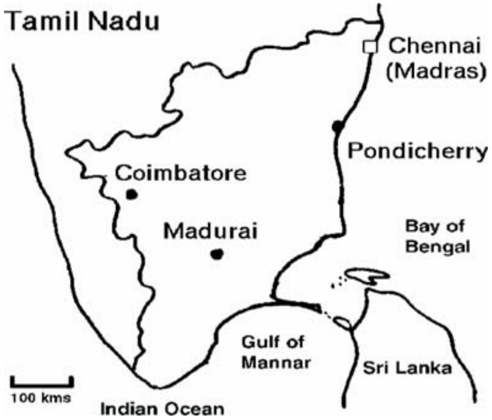
- simple tasks
- Matching input and training data
- With more training data

Proposals:

- SDS design should limit the complexity of the ASR task to approximately 10 words or fewer at any given dialog node.
- Speech collection should be integrated into the SDS design. Thus needs of the user and needs of the system are met simultaneously.
- ASR for each village could achieve adequate accuracy for SDS usability with the speech of only 15 speakers.

Map: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)



UI Design

Overall Guidelines:

- Let users feel in charge
- Spare users as much effort as possible
- An appropriate and affective user interface is one that fits the task to be accomplished.

The nature of the task should dictate appropriateness of UI style, not the level of expertise of the user.

Speech UIs

- Less expensive than display-based UI solutions
- More accessible than text-based UI solutions
- Voice feedback in the local language helps with user interest and comfort

Map: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)

UI Design - Speech

Literacy:

•In Tamil Nadu, illiteracy rates are 50% for men, 80% for women.

•Information is primarily disseminated via word of mouth.

•Unschooled adults rely on empirical situational reasoning

•Design features considered to be intuitive (hierarchical browsing, icons representing items, etc) present a challenge to illiterate user.

Design Guidelines:

•Ease of learning

•No textual requirements

•Graphics (and possibly speech in local language)

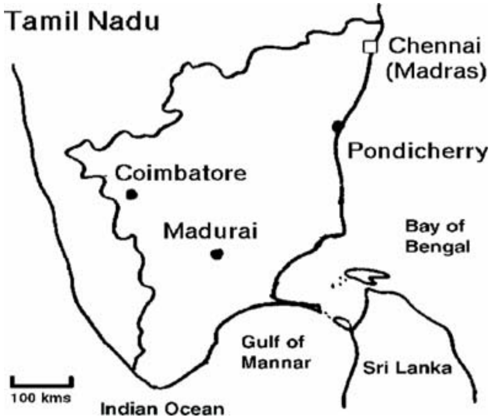
•Support for internationalization

•Accommodates localization

•Simple, easy to use, tolerant of errors

•Accurate content

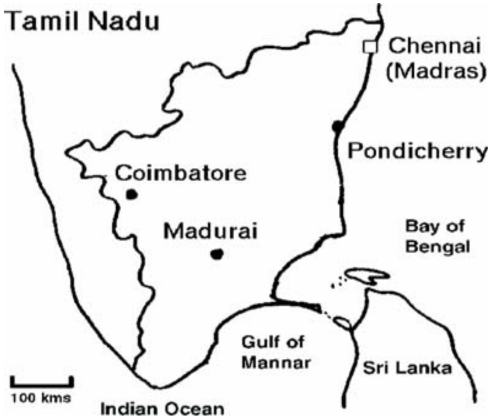
•Robust in (potentially distracting) public places



Map: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)

UI Design - Speech



Localization:

For each language and culture the following UI elements are subject to change:

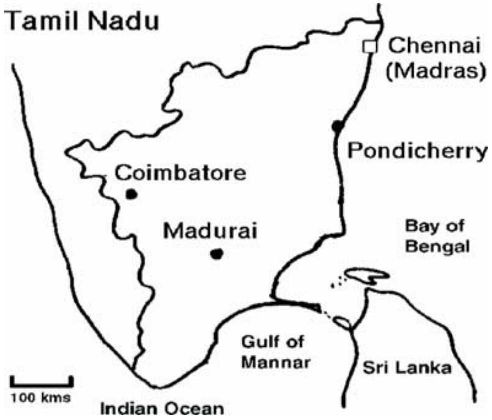
- Fonts
- Color
- Currency
- Abbreviations
- Dates
- Register
- Concepts of time and space
- Value Systems and Behavioral Systems

User study techniques such as questionnaires, storyboards and walkthroughs present difficulties for the illiterate due to their daily requirements and ambient infrastructure. **Successful UIs** should build on existing means of info transfer and existing linguistic and cultural expertise by enabling community authorship of content.

Map: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)

Field Study 2: OpenSesame SDS



Experiment:

•Development of SDS template for creating multi-modal SDS

•Collaboration with MSSRF staff

•Goal: Port 1 unit (Banana crop) of the text-based Valam website to the interactive OpenSesame application.

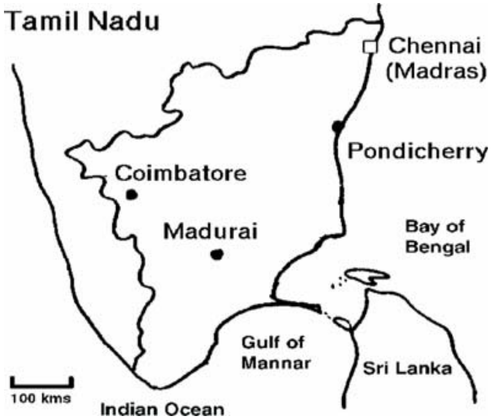
•User studies conducted using live speech recognition in Dindigul region, Tamil Nadu

•Audio input recorded during user interactions with OpenSesame SDS served to simulate integrated data collection and ASR adaptation techniques.

Images: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)

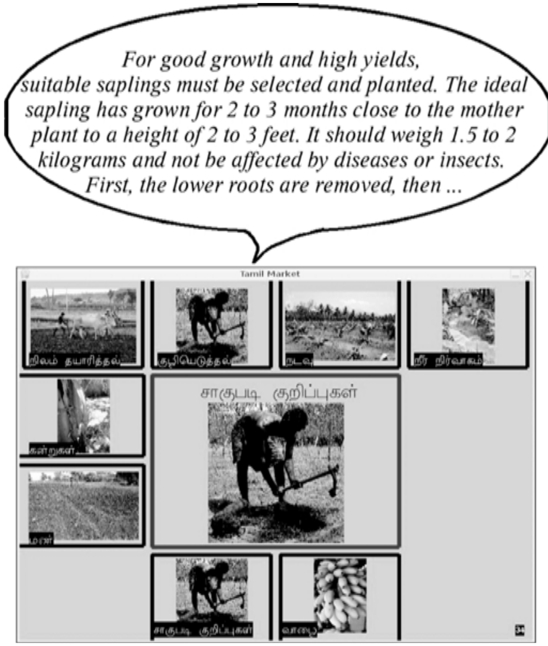
Field Study 2: OpenSesame SDS



OpenSesame – Banana Crop Application

The application was a tool to educate the user via digital photographs and a narrative in Tamil on the recommended practices for growing, harvesting and propagating a banana crop according to local conditions and needs.

- •Interactive UI that adhered to aforementioned guidelines
- •Completed in less than 3 weeks:
  - -identifying appropriate content (sites, photos etc)
  - -varying the accuracy of the text version
  - -gathering digital pictures
  - -recording speech output
  - -synchronizing all elements
- •Input: Speech and Touch screen
- •Output: Graphics, small text, pre-recorded audio files


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 ...
Tamil Market
நிலம் தயாரித்தல்
குழியெடுத்தல்
நடவு
நீர் நிர்வாகம்
கன்றுகள்
சாகுபடி குறிப்புகள்
மண்
சாகுபடி குறிப்புகள்
வாழை

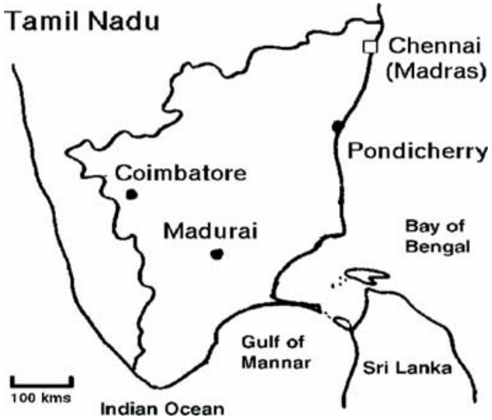
Images: Plauché and Nallasamy, 2007.

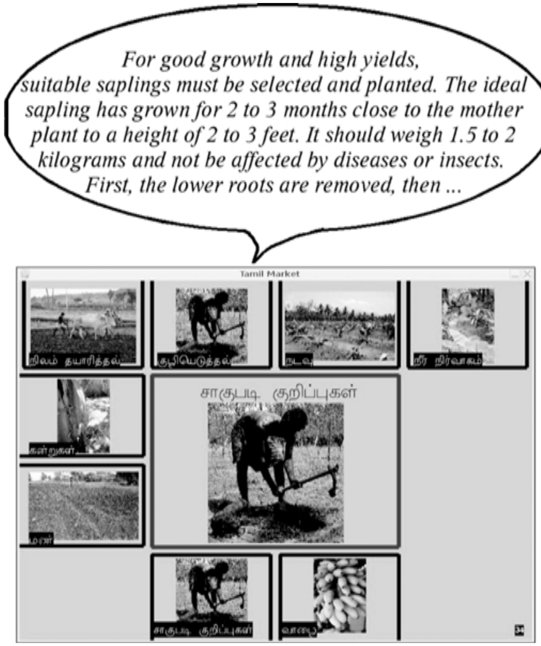
Speech Interfaces or Spoken Dialog systems (SDS)

Field Study 2: OpenSesame SDS

OpenSesame – Banana Crop Application

- 28 acoustically dissimilar and locally appropriate vocabulary words were selected to correspond to the Valam site subheadings
- Menu system was only 3 levels deep
- No more than 8 options at a time were presented
- The system was highly redundant when no input was provided
  - -listing options at every screen
  - -disseminating info in the form of audio slide show

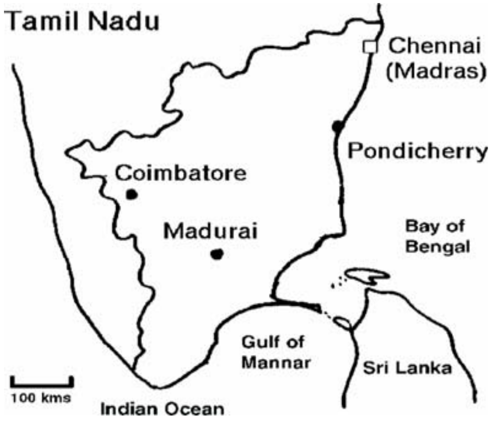




For good growth and high yields, suitable saplings must be selected and planted. The ideal sapling has grown for 2 to 3 months close to the mother plant to a height of 2 to 3 feet. It should weigh 1.5 to 2 kilograms and not be affected by diseases or insects. First, the lower roots are removed, then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

Images: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)



Field Study 2: OpenSesame SDS

OpenSesame – Banana Crop Application ASR

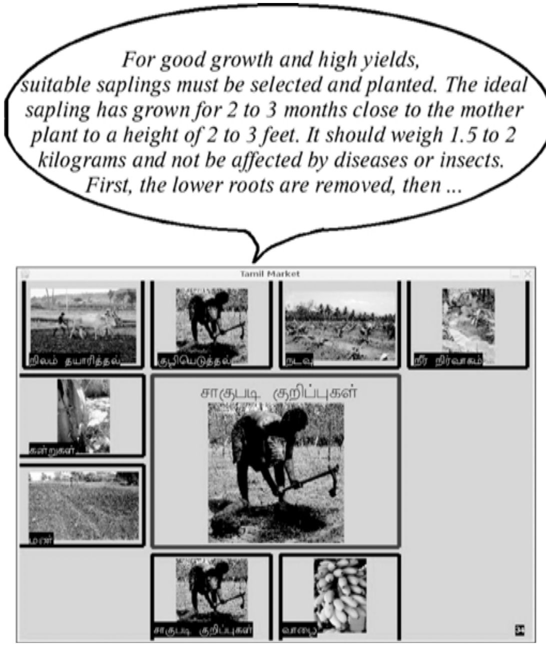
Requirements:

- Recognize multiple speakers
- Be robust to noisy conditions under limited linguistic data

Design:

- The ASR was trained in Tamil speech recordings used in Field Study 1,
- built using the HTK
- Pronunciation dictionary was prepared along with its phonemic representation (sub-word level) allowing to accommodate new words and phonetic contextual variations
- Test database was prepared by recording 5 MSSRF members

  -Triphone models (single Gaussian) performed at 97% accuracy


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 ...
Tamil Market
நிலம் தயாரித்தல்
குழியெடுத்தல்
நடவு
நீர் நிர்வாகம்
கன்றுகள்
சாகுபடி குறிப்புகள்
மண்
சாகுபடி குறிப்புகள்
வாழை

Images: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)

Field Study 2: OpenSesame SDS

OpenSesame – Banana Crop Application ASR

Field evaluations:

•System was evaluated by 50 people in 3 different conditions across 6 different sites

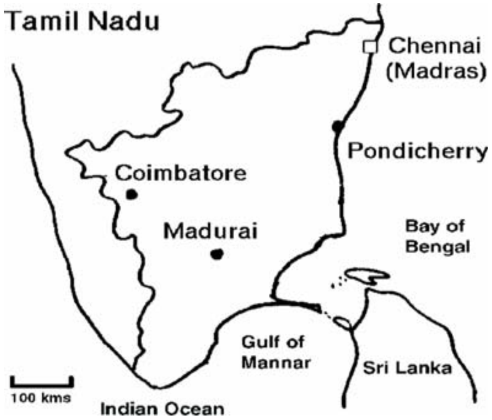
•200 more people were onlookers

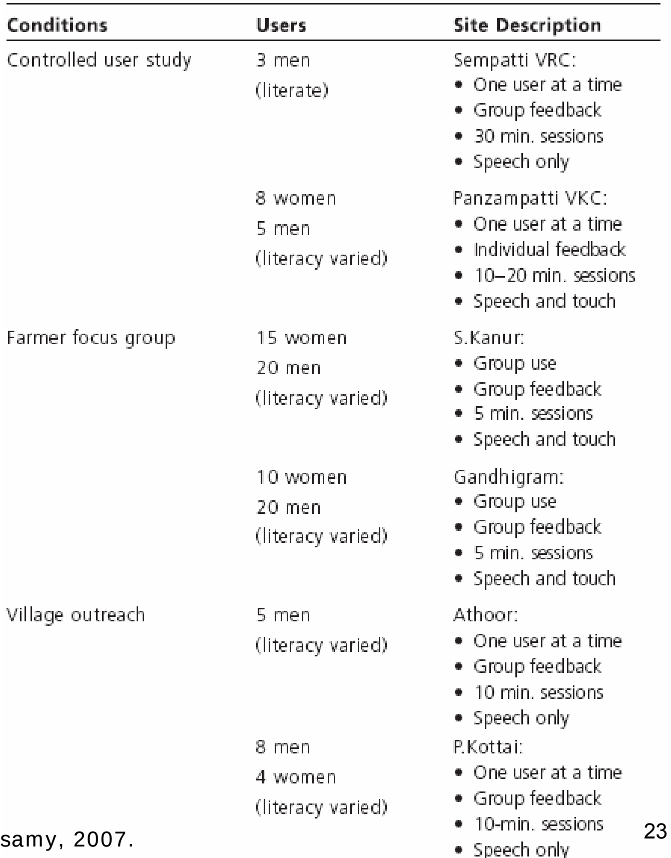
•Participant's audio commands to SDS were recorded during use

•Sessions were sort

•Involved little training

•Informal feedback was requested (content ease of use, preferred modality (audio/touch-screen))



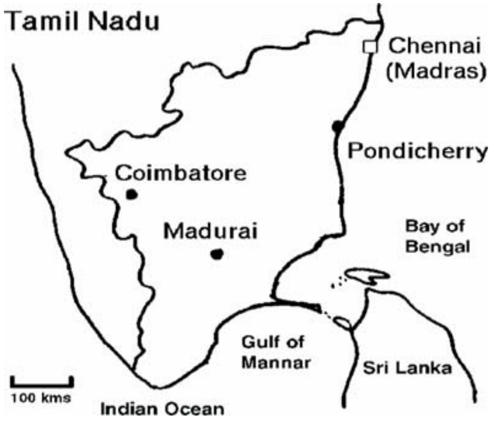


Conditions	Users	Site Description
Controlled user study	3 men (literate)	Sempatti VRC: • One user at a time • Group feedback • 30 min. sessions • Speech only
	8 women 5 men (literacy varied)	Panzampatti VKC: • One user at a time • Individual feedback • 10–20 min. sessions • Speech and touch
Farmer focus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

Images: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)

Field Study 2: OpenSesame SDS



OpenSesame – Banana Crop Application ASR

Results:

•Input categories:

-N/A: empty sound files/ no speech/ background speech (23% success- silence recognition)

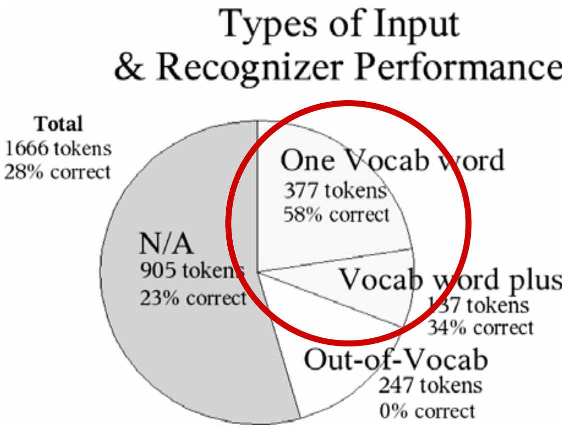
-Out-of-vocab:15% of input were utterances directed to SDS but not included in ASR's vocabulary (0% success)

-One vocab word: contained 1 ASR vocabulary word (58% success)

-Vocab word plus: 1 ASR vocabulary word + other non-vocabulary word/speech (34% success)

•Recognition performance on isolated words was much worse during SDS interactions (58%) than that of MSSRF staff (97%):

Dissimilarity probably due to difference in speaking style (reading aloud /issuing commands to a machine)



One vocab word + Vocab word plus input was <30% of total input

Images: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)

Field Study 2: OpenSesame SDS

OpenSesame – Banana Crop Application ASR

Results:

•Recognition performance by site revealed that social and environmental factors affect performance

-Controlled user study with literate subjects yielded high performance (highest)

-A not well-controlled user study with illiterate subjects yielded low performance (lowest)

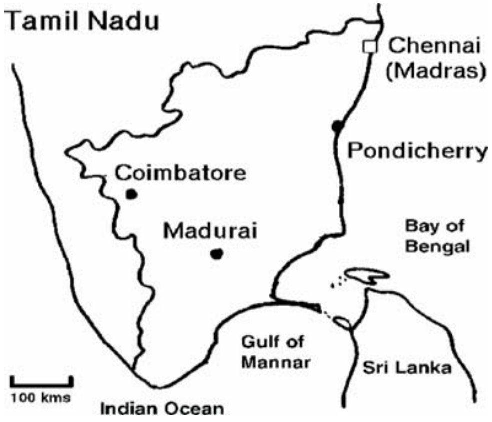
•Participants reported that the interface was easy to use

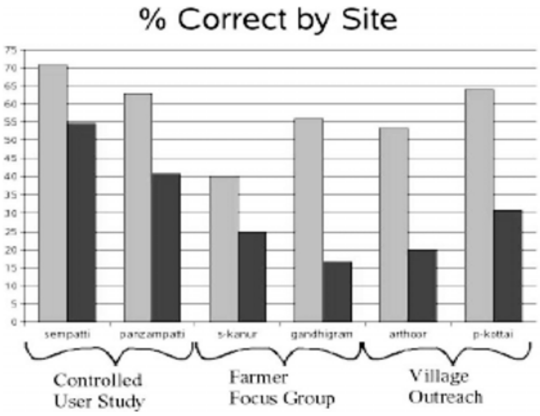
•Educated participants commented that the system would be “good for people who cannot read”.

•Some subjects preferred speech as a means of input, while others speech

•Many corrections and suggestions were proposed for the SDS, such as the addition of more crops to the system

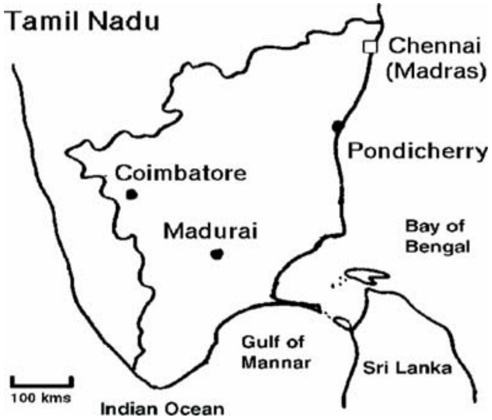
•MSSRF staff played a key role in the evaluative sessions





Images: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)



Images: Plauché and Nallasamy, 2007.

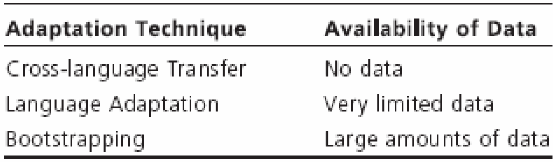
ASR Adaptation

A technique for automatically or semi-automatically optimizing a recognizer by gradually integrating new, untrascribed data into the models for speech.

Fact:

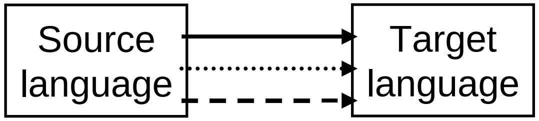
•Only 3 speech technology efforts have been directed to Tamil which is spoken by 60 million people.

Techniques to overcome this barrier:



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

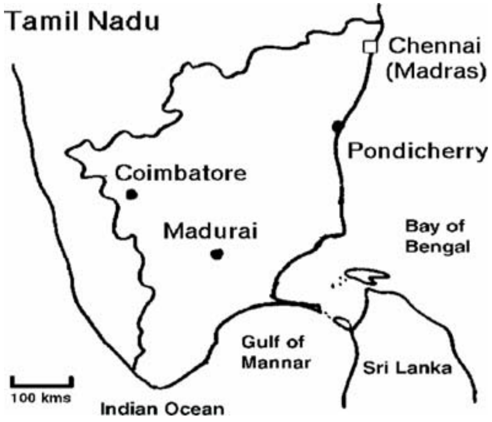
•**Cross-language transfer**: When annotated corpora is not available (as in Tamil) an ASR trained on transcribed data from one or more (source) languages can be used to recognize speech in the new (target) language.



•**Language adaptation:** ASR trained on a large source language corpus and then the acoustic models are adapted to a very limited amount of target language data. (depends on # speakers + data).

•**Bootstrapping:** Acoustic models are initialized from a small amount of transcribed source data. The ASR is then iteratively built, using increased amounts of training data and adaptation.

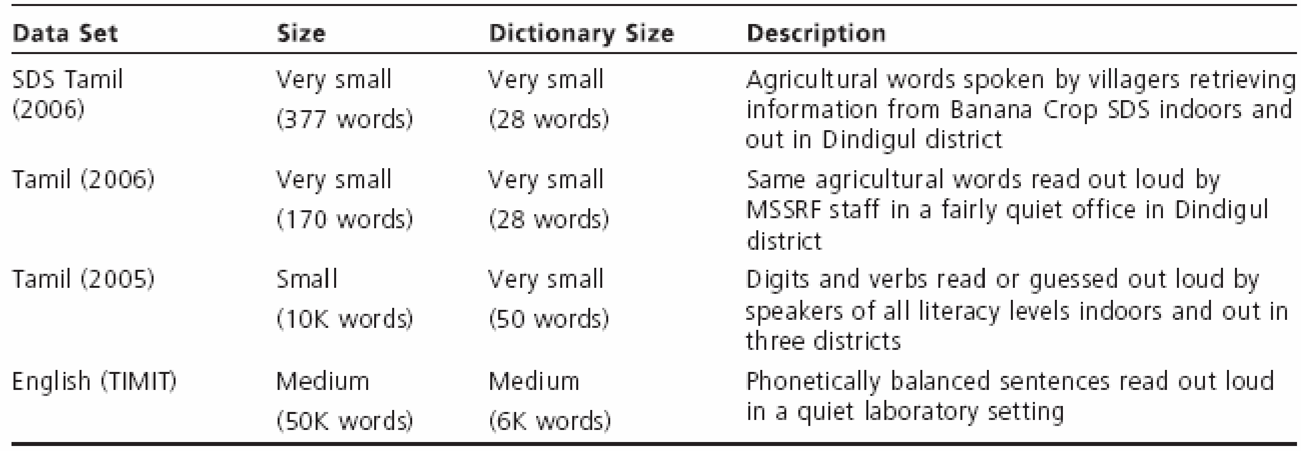
Speech Interfaces or Spoken Dialog systems (SDS)



Experiment 4:

Goal: How to optimize the small vocabulary recognizer to the speech of a particular community given no or limited Tamil training data.

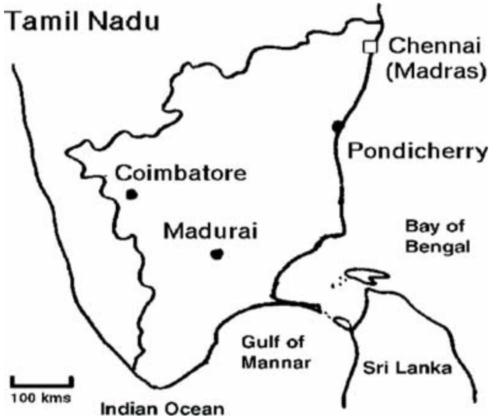
- Use of speech collected during **Banana Crop Application**
- Use of **Cross-language transfer** and **Language adaptation**
- Databases used: SDS Tamil 2006 /Tamil 2006/Tamil 2005/ English TIMIT



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

Images: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)



Experiment 4:

Process:

•In the field, the recognizer trained on the Tamil 2005 database recognized commands for Banana Crop SDS with 58.1% accuracy

•Substantial improvement in accuracy (68.7%) occurred via:

-the collapse of certain contrastive phonetic categories (long vs. short vowels)

-the addition of noise robustness method (cepstral mean subtraction) to factor out environmental noise and generalize across tasks and speakers.

•**Cross-language transfer:**

-Tamil phonemes were mapped to English as closely as possible

-Training and decoding were performed using HTK

1.Acoustic models are trained with a default flat initialization

2.Triphone models are developed based on monophone HMMS and the ASR decodes using a simple finite state grammar.

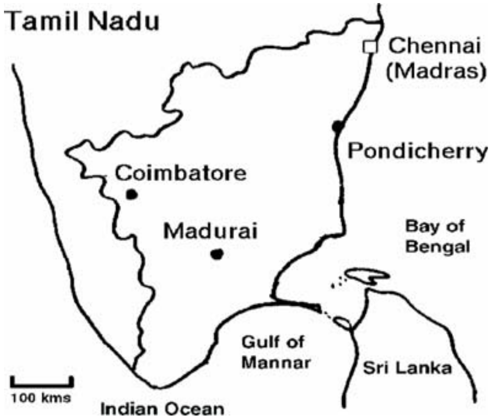
Results:

Tamil SDS powered by English recognizer had 30% accuracy

Conclusion:

Its better to train on a small amount of same language data than on a greater amount of mismatched data

Speech Interfaces or Spoken Dialog systems (SDS)



Experiment 4:

Process:

•Recognizers were initialized on either English or Tamil and then the recognizer was adapted to the Tamil 2006 database (5 volunteers -**1 hour recording**); maximum likelihood linear regression was used.

-Adaptation to Tamil 2006 improved the performance for both the recognizer trained in English and the recognizer trained on Tamil (82.2% and 80.4% accuracy rates respectively).

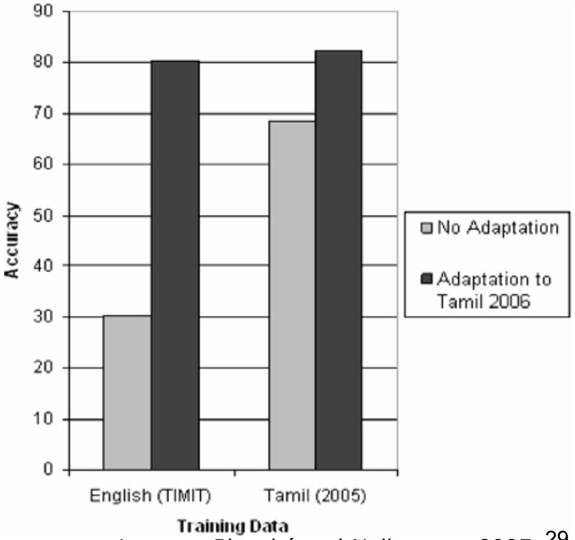
Conclusions:

•It is more sensible to use an existing English trained system with a small database (like Tamil 2006) than to use a larger database such as Tamil 2005 (**100 hours of recording**) if they are to yield similar results (**82.2% , 80.4%**). It overcomes **recording costs.**

•**Other methods of ASR adaptation mentioned:**

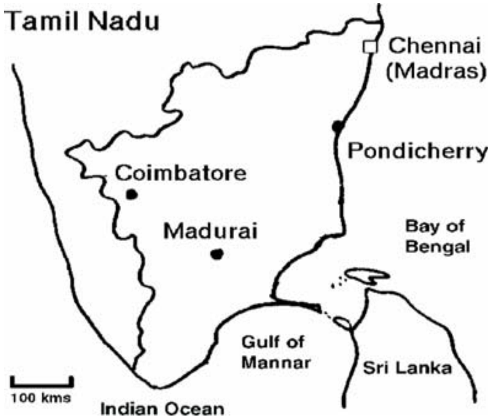
-Supervised adaptation

-Unsupervised adaptation



Images: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)



Future plans and overall conclusions:

•We reviewed Speech technologies and techniques that are:

-small

-scalable

-easy to modify and update by local stakeholders

And that can be constructed to deliver:

-accurate

-locally relevant

information to individuals regardless of their literacy level

•**Integrated data collection** and **language adaptation** are found to be useful techniques for collecting linguistic resources according to user needs and system needs

•**Future tasks:**

-determine the minimum amount of adaptation data required to reach adequate levels of ASR accuracy

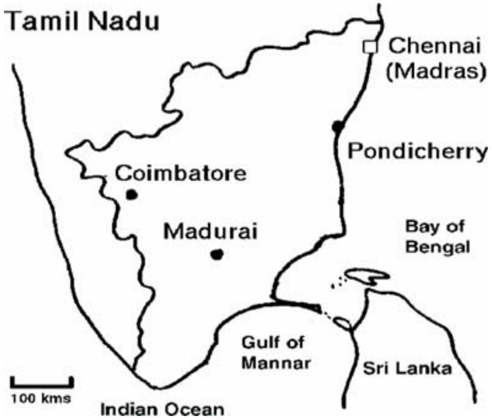
-develop speech/no speech detectors and out of vocabulary models

Map: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)

Questions:

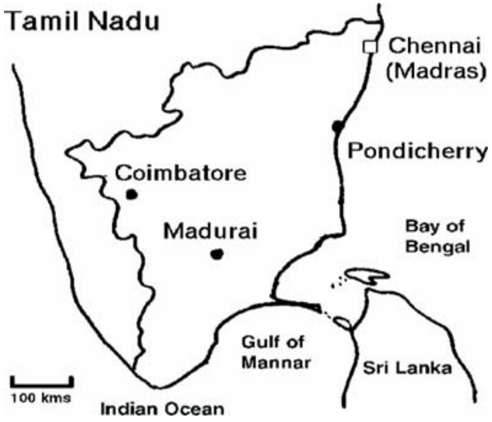
- **Ways of increasing the one vocab + “space solution”?**
- **Other ASR adaptation techniques?**



Map: Plauché and Nallasamy, 2007.

Speech Interfaces or Spoken Dialog systems (SDS)

Thank you



Map: Plauché and Nallasamy, 2007.

MAS.965 / 6.976 / EC.S06 NextLab I: Designing Mobile Technologies for the Next Billion Users
Fall 2008