A Comparative Study of Speech and Dialed Input Voice Interfaces in Rural India

Neil Patel, Sheetal Agarwal, Nitendra Rajput, Amit Nanavati, Paresh Dave, Tapan S. Parikh
The Case for Voice Interaction
Asking a question on Avaaj Otalo

AO: <tune> Welcome to Avaaj Otalo!
You can get to the information by saying a single word.
To ask a question, say ‘question’.
To listen to announcements, say ‘announcements’.
To listen to the radio program, say ‘radio’.

User: I want to ask a question.

AO: Sorry, I didn’t understand. I can only understand single words.
Do you want to ask a question - yes or no?

User: Yes.

AO: OK, you want to ask a question.
To ask a question about agriculture, say ‘agriculture’.
To ask about animal husbandry, say ‘animal’.
Key design choice: input modality

• Application requirements
  – Inexperienced/low literacy users
  – Learnable without training

• Speech is natural
  – But speech recognition requires lots of data
Small-vocabulary, isolated word speech interfaces

Tamil Market

- 27-word vocabulary
- 18 speakers’ training data
- 98% accuracy

[Plauché et. al. 2006]
The Experiment

- Previous work for technical professionals in U.S.
- This study: low-literacy, inexperienced users
Methodology and Participants

- 45 participants, two treatments, between-subjects
- Small-scale farmers (median 10 acres)
- Native Gujarati, no English
- No experience with voice interfaces
- 73% less than 8th grade education;
  87% never used a PC
Application Features

ANNOUNCEMENTS

RADIO ARCHIVE

QUESTION AND ANSWER

Sketches courtesy of Bill Verplank
Tasks

1. Listen to announcements (1 step)
   – sequential, 30-60 second audio snippets

2. Listen to a radio program (2 steps)

3. Record a question (9 steps)
   – Categorize question (4 steps)
   – Record question (2 steps)
   – Provide personal contact information (3 steps)
Speech Recognition Accuracy

• Our method: cross-language transfer
  – Apply unmodified acoustic model using transliterated vocabulary
• Accuracy: 94% (commercial systems: ~98%)
• Alternative: model adaptation
  – Linear transformations based on GMM parameters
  – Requires some speech in target language
Testing Environment

Participants

Office

38
(half speech, half touchtone)

Village

7
(half speech, half touchtone; all women)
Overall task completion: touchtone higher than speech

- Speech: 61%
- Touchtone: 74%

(p<0.05)
Speech: slightly higher task completion with most educated users

- Speech: 81%
- Touchtone: 73%

8th and Above
Overall user satisfaction: comparable

Speech: 83%
Touchtone: 91%

‘Yes’ or ‘Definitely yes’
(5-point Likert scale)
Percentage of tasks rated “difficult” or “very difficult”

Across all tasks: 49% (speech) vs. 30% (touch), p<0.05
Why was speech less successful?

• Single-word input awkward
• Recognition errors
  – 67% to 42% drop with 1+ errors
• Touchtone benefited from simple, linear tasks
<table>
<thead>
<tr>
<th>Study</th>
<th>Technology</th>
<th>Literacy</th>
<th>Training</th>
<th>Result; User Pref</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM staff, U.S [Lee and Lai, 2005]</td>
<td>Natural language ~80% accuracy</td>
<td>High</td>
<td>None</td>
<td>Touchtone; prefer speech</td>
</tr>
<tr>
<td>Hospital staff, Botswana [Sharma et. al., 2009]</td>
<td>Wizard of Oz 100% accuracy</td>
<td>Low</td>
<td>Yes</td>
<td>No sig. diff.; prefer touchtone</td>
</tr>
<tr>
<td>Community health workers, Pakistan [Sherwani et. al., 2009]</td>
<td>Cross-language transfer 93% accuracy</td>
<td>Low</td>
<td>Yes</td>
<td>Speech; no preference</td>
</tr>
<tr>
<td>THIS STUDY</td>
<td>Cross-language transfer 94% accuracy</td>
<td>Low</td>
<td>None</td>
<td>Touchtone; no preference</td>
</tr>
<tr>
<td>Farmers in India</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Comparing speech vs. touchtone studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Technology</th>
<th>Literacy</th>
<th>Training</th>
<th>Result/User Pref</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM staff, U.S [Lee and Lai, 2005]</td>
<td>Natural language</td>
<td>High</td>
<td>None</td>
<td>Touchtone; prefer speech</td>
</tr>
<tr>
<td>Hospital staff, Botswana [Sharma et. al., 2009]</td>
<td>Wizard of Oz</td>
<td>Low</td>
<td>Yes</td>
<td>No sig. diff.; prefer touchtone</td>
</tr>
<tr>
<td>Community health workers, Pakistan [Sherwani et. al., 2009]</td>
<td>Cross-language transfer</td>
<td>Low</td>
<td>Yes</td>
<td>Speech; no preference</td>
</tr>
<tr>
<td>THIS STUDY</td>
<td>Cross-language transfer</td>
<td>Low</td>
<td>None</td>
<td>Touchtone; no preference</td>
</tr>
<tr>
<td>Farmers in India</td>
<td>94% accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TECHNOLOGY</td>
<td>LITERACY</td>
<td>TRAINING</td>
<td>RESULT; USER PREFERENCES</td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>----------</td>
<td>----------</td>
<td>-----------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Natural language (~80% accuracy)</td>
<td>High</td>
<td>None</td>
<td>Touchtone; prefer speech</td>
<td></td>
</tr>
<tr>
<td>Wizard of Oz (100% accuracy)</td>
<td>Low</td>
<td>Yes</td>
<td>No sig. diff.; prefer touchtone</td>
<td></td>
</tr>
<tr>
<td>Cross-language transfer (93% accuracy)</td>
<td>Low</td>
<td>Yes</td>
<td>Speech; no preference</td>
<td></td>
</tr>
<tr>
<td>Cross-language transfer (94% accuracy)</td>
<td>Low</td>
<td>None</td>
<td>Touchtone; no preference</td>
<td></td>
</tr>
</tbody>
</table>

**THIS STUDY**

- Farmers in India
Comparing speech vs. touchtone studies

<table>
<thead>
<tr>
<th>TECHNOLOGY</th>
<th>LITERACY</th>
<th>TRAINING</th>
<th>RESULT; USER PREF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural language</td>
<td>High</td>
<td>None</td>
<td>Touchtone; prefer speech</td>
</tr>
<tr>
<td>~80% accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wizard of Oz</td>
<td>Low</td>
<td>Yes</td>
<td>No sig. diff.; prefer touchtone</td>
</tr>
<tr>
<td>100% accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross-language transfer</td>
<td>Low</td>
<td>Yes</td>
<td>Speech; no preference</td>
</tr>
<tr>
<td>93% accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>94% accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

THIS STUDY

<table>
<thead>
<tr>
<th>TECHNOLOGY</th>
<th>LITERACY</th>
<th>TRAINING</th>
<th>RESULT; USER PREF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-language transfer</td>
<td>Low</td>
<td>None</td>
<td>Touchtone; no preference</td>
</tr>
<tr>
<td>94% accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Farmers in India</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Current Status - Pilot

• Live with 50 farmers; over 3500 hits/month
• 70% of calls in first month used touchtone
Thanks to...

• Our partners
  – Development Support Center, Gujarat, India
  – IBM India Research Laboratory

• Our funding sources
  – Stanford SOE
  – IBM India Research Laboratory

• The farmers of Gujarat!

• My advisors: Scott Klemmer and Tapan Parikh

http://hci.stanford.edu/research/otalo/