Maximizing the Efficiency and Usability of Speech Based Applications

Jennifer Lai
IBM Research
Pervasive Computing
lai@watson.ibm.com
Overview of today's talk

- the design of an effective voice based system
  - why use speech
  - the importance of prompts
- research findings in speech
  - how well is synthetic speech understood?
Effective Speech Design

The key is in the prompts
Benefits of Speech

- Natural for Humans
- No Keyboard Required
- Direct Access
- Hands & Eyes Busy
Challenges:
- Asymmetric
- High expectations
- Immature technology
- Transient
- Variable input
- Awkward Output
- Error Prone
- Lack of grammars
Types of Synthesizers

- Parameterized
  - Formant-based
  - Articulatory

- Concatenative
  - Concatenative words
  - Concatenative phonemes

human  synthesized
ibm_tts
Conversation

- Speaking and listening are collaborative human activities
- Even a simple Speech UI is a conversation
- Adhere to conversational conventions
  - establish a common ground
  - use discourse cues
  - avoid repetition
## E-mail Example

<table>
<thead>
<tr>
<th>Sender</th>
<th>Subject</th>
<th>Date/Time</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arlene Rexford</td>
<td>Learn about Java</td>
<td>Mon Oct 28 11:23</td>
<td>2k</td>
</tr>
<tr>
<td>Shari Jackson</td>
<td>Re: Boston rumors</td>
<td>Fri Jul 18 09:32</td>
<td>3k</td>
</tr>
<tr>
<td>Hilary Binda</td>
<td>Change of address</td>
<td>Wed Jul 16 12:59</td>
<td>1k</td>
</tr>
<tr>
<td>Arlene Rexford</td>
<td>Class Openings</td>
<td>Tue Jul 21 12:35</td>
<td>5k</td>
</tr>
<tr>
<td>George Fitz</td>
<td>Re: Boston rumors</td>
<td>Thu Jul 24 10:18</td>
<td>1k</td>
</tr>
</tbody>
</table>

### long mail

### conversational mail
Prompts: Implicit & Explicit

- **Explicit or directive prompts**
  - indicate exactly what to say
  - useful for constraining user input
  - referred to a "directed dialog" systems

- **Implicit or conversational prompts**
  - are open ended
  - rely on conversational conventions
  - provide more natural interaction
  - allow more room for error
  - known as "conversational" systems
Prompts: Continuum

Continuum from explicit to implicit:

Welcome to ABC Bank. You can check an account balance, transfer funds, or pay a bill. Say "check balance," "transfer funds," or "pay bills."

Welcome to ABC Bank. You can check an account balance, transfer funds, or pay a bill. What would you like to do?

Welcome to ABC Bank. What would you like to do?
Prompts: Be Brief

- Tempting to make prompts wordy
- Remember, speech output is slow
- Allow barge-in
  - ability to interrupt TTS
  - double-edged sword
    - speeds interaction
    - users may miss important information
Prompts: Taper

- Tapering can reduce speech output
- Taper presentation of data
  - remove unnecessary words
- Taper prompts over time
  - make same prompts shorter
  - remove "hints"
Tapering Examples

As of 15 minutes ago, Sun Microsystems was trading at 45, up 1/2
Motorola was at 83, down 1/8, and
IBM was at 106, up 14

Start recording after the tone. Pause for several seconds when done.
Record after the tone, then pause.
Record then pause.

You have 2 new messages. The first is from Rich Miner. I'd recommend saying What's-it-say, Describe it, or Next-item.
You have 2 new messages. The first is from Rich Miner.
Welcome to ABC Bank. What would you like to do? <timeout>

You can check an account balance, transfer funds, or pay a bill. <timeout>

Say "check balance," "transfer funds," or "pay bills."
Errors: cooperative behavior

✓ Aim for cooperative behavior
  ➔ eliminate repetitive messages
  ➔ try to move conversation forward
  ➔ provide progressive assistance

1st prompt: "What?"
2nd prompt: "Sorry. Please rephrase"
3rd prompt: "Still no luck. Speak clearly, but don't overemphasize"
Design for Errors

✓ Aim for cooperative behavior
  ➤ eliminate repetitive messages
  ➤ provide progressive assistance

  1st prompt: "What?"
  2nd prompt: "Sorry. Please rephrase"
  3rd prompt: "Still no luck. Speak clearly, but don't overemphasize"

✓ Improve Odds
  ➤ Reprompt with explicit choices
  ➤ Switch to more constrained grammar
  ➤ Provide backup mode
    ● touchtone or human
Consequences of Errors

- Errors have serious consequences:
  - destroys user mental model
  - user must initiate correction
  - error can change meaning of text
Research Findings

How well is synthetic speech understood
Two Studies Examining TTS Comprehension

Why

- big move in the industry to deliver personalized news and data such as email messages and reminders
- most studies were done at the word level
- large subjective differences

Lab study

In car study
First Study - In Lab

What

- measured comprehension levels for 5 commercial synthetic speech engines
- focused on longer messages
- varied message type and task conditions
  - short, medium and long email
  - always listened on a telephone

Who

- 78 subjects - IBM employees
- native English speakers
- no reported hearing problems
- screened out those with significant TTS experience
Findings

- No significant difference for comprehension performance among the 5 engines
- Subjects performed significantly worse with synthetic speech for all tasks
- Comparable times for human and synthetic

<table>
<thead>
<tr>
<th></th>
<th>Av. Time w/ notes</th>
<th>Av. Time w/o notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>199 seconds</td>
<td>174 seconds</td>
</tr>
<tr>
<td>Synthetic</td>
<td>202 seconds</td>
<td>175 seconds</td>
</tr>
</tbody>
</table>

- No effect observed for age of subjects
- Education level did not effect comprehension accuracy with synthetic speech
Comprehension Performance

- Performance for TTS degraded as task got longer and less familiar (in non-note-taking condition).
- Prior experience with TTS gave a small advantage in the note-taking condition.
- Subjects listening to a male synthetic voice had higher comprehension levels.

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Synthetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male TTS Group</td>
<td>72%</td>
<td>65%</td>
</tr>
<tr>
<td>Female TTS Group</td>
<td>72%</td>
<td>51%</td>
</tr>
</tbody>
</table>
Comprehension by note-taking and voice type

Accuracy

Notes

No Notes

Human

Synthetic
Second Study - In Car

What

- measured comprehension of synthetic speech and driving performance
- used simulator at University of Michigan
- 3 types of messages
- measurements:
  - accuracy
  - subjective
  - standard deviations
Driving Study (continued)

Who

- 24 subjects
- drivers represented two extremes of the population
  - 12 (ages 21-28)
  - 12 (ages 65-71)
- equally divided between males and females
- subjects underwent both a hearing and a vision test
Findings

Comprehension of synthetic speech messages was lower than for recorded human messages and this was consistent across message lengths.

There was a significant difference in comprehension levels for the different types of messages.

<table>
<thead>
<tr>
<th>Navigation</th>
<th>Email</th>
<th>News</th>
</tr>
</thead>
<tbody>
<tr>
<td>95.5%</td>
<td>59%</td>
<td>74%</td>
</tr>
</tbody>
</table>

Overall, younger subjects performed better than older subjects.
Findings (continued)

- Driving performance did not degrade as a result of listening to messages - regardless of voice type
  - caveat: the simplest possible driving conditions

- There was no significant difference between the genders
Maximizing the Efficiency and Usability of Speech Based Applications

Jennifer Lai
IBM Research
Pervasive Computing

lai@watson.ibm.com