Multimodal Voice UI for Mobile Devices: Designing for User Confidence
Vlingo Vision

Vlingo is the starting point for anything the user wants to do on their mobile phone

- One-key access from any context
- “Unconstrained” commands
- Speaker-independent recognition paired with per-speaker and system-wide adaptation
Current Features

- **Web Search** – `[search]` American Airlines flight 257
- **Voice Dial** – `call` Mom *at home*
- **Email** – `email` Steve *subject* Catching up *message* Man, it’s been too long. When can we get dinner?
- **SMS** – `send message to` Kate *message* Hey, what are you up to tonight?
- **Notes** – `note to self` buy birthday present for Lindsay
- **Launch Application** – `open` Calendar
- **Facebook** – `facebook status` is flying to San Diego
- **Twitter** – `twitter` Fido ate my BlackBerry. Again. And an eight-pack of bulkie rolls.
User perception of speech recognition lags behind the state of the art

- Why do you want to use speech recognition?
  - Useful, Convenient, Cool

- What has been your experience with speech recognition?
  - Error-prone, Frustrating, Slow

- Our challenge: predictability and recoverability
Guiding Principles

**Mobile Voice UI Guiding Principles**

- Provide multimodal feedback
- Extend, don’t replace, familiar UI elements
- Show the user what system recognized
- Make correction fast, easy and appropriate for the range of errors that occur
## Multimodal Feedback: The Recognition Process

<table>
<thead>
<tr>
<th>User Action</th>
<th>Vlingo Action</th>
<th>Audio Feedback</th>
<th>Visual Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Press side key</td>
<td>Vlingo starts listening</td>
<td>Ascending tone</td>
<td>Listening animation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speak</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Release side key</td>
<td>Start processing audio</td>
<td>Descending tone</td>
<td>Thinking animation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Finish processing audio</td>
<td>Success tone</td>
<td>Recognized text, interpreted intent</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TTS results confirm recognized text, coach user “Search concert tickets”</td>
<td></td>
</tr>
</tbody>
</table>
Recognition Results: Extend the Familiar, Show Recognized Text

• Show recognized text to put results in context ex: *Sushi vs. Shoes*

• Take immediate action when low-risk… but always allow correction
Instead of dialog, focus on efficiency. Allow users to:

- (Constrained result set) Select from similar matches
- Select phrase alternatives (phrase nbest)
- Select from word alternatives (word nbest)
- Speak again
- Type
Inspiring User Confidence

**Predictability**
- Provide multimodal feedback
- Show what system recognized
- Extend, don’t replace, familiar UI elements

**Recoverability**
- Extend, don’t replace, familiar UI elements
- Make correction fast, easy and situationally appropriate