Voice Typing: A New Speech Interaction Model for Dictation on Touchscreen Devices

Anuj Kumar\textsuperscript{1,2}, Tim Paek\textsuperscript{1}, Bongshin Lee\textsuperscript{1}

\textsuperscript{1} Microsoft Research, Redmond, USA
\textsuperscript{2} Carnegie Mellon University, Pittsburgh, USA
Mobile devices have widely penetrated the market.
Text Input (28%) Emails, Messages

Voice Calls (25%)

Others (47%) Social Networking, Games, Maps

187.7 B text and email messages sent in Dec 2010 in North America (Wireless Facts, CTIA 2011)
Existing Techniques for Text Input

- **Typing**
  - QWERTY
  - Half-QWERTY
  - Multi-tap
  - T9 (predictive text entry)

  Lack of haptic feedback;
  Ergonomic issues e.g. “fat finger problem”

- **Recognition Oriented**
  - SWYPE
  - Handwriting recognition, etc.

  Either slow, or inaccurate
Text Input via Speech

Offers several potential advantages

With speech, interaction becomes independent of device size

If accurately recognized, speech is three times faster than QWERTY (Basapur et al. ’07)

Only plausible input modality for 800 million non-literate users

Typing Speeds

<table>
<thead>
<tr>
<th>Type</th>
<th>Average User</th>
<th>Experienced User</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Handwriting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QWERTY</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predictive Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multi-tap</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Problems with Current Dictation Systems

“Voice Recorder” type interaction style

- Users formulate utterance
- Say it aloud
- Wait for a few seconds
- See the entire output at once

Real-time presentation of output sacrificed for potential accuracy gains
Problems with Current Dictation Systems

Error identification and correction takes **75% of time** (Karat et al., ’99)

Break thought chain, verify output verbatim

Each error edit requires at least two actions: **selection & correction**

**Error editing is time intensive & frustrating**
Real-time Feedback & Speaking Style

Discrete Recognition
- 1 word at a time
- Pause after each word – does not match speaking mental model
- “Conversation with a foreign accent friend”

Voice Typing
- Each chunk is a part of a thought
- Enables real-time error identification & correction
- Chunks of 2-4 words at a time
- “Typist Secretary”

Traditional Dictation
- 1 utterance at a time
- No real-time feedback
- “Voice Recorder”

Microsoft Research
Cognitive Motivations for Voice Typing

- Real-time feedback not only promotes learning of interface, but also leads to greater satisfaction (Payne, '09)

- Similar to back-channel feedback in real conversations, real-time feedback provides “common ground” (Clark et al. '91)

- Similar to current mental models of keyboard typing where users typically monitor and correct text as-they-type
Technical Motivations for Voice Typing

- Most recognition errors occur due to incorrect segmentation
  
  **Utterance:** “It’s hard to recognize speech”
  
  **Recognition failure:** “It’s hard to wreck a nice beach”
  
  /s/ incorrectly attached to “nice”, instead of “speech”

  With Voice Typing, users likely to pause where segmentations should occur
  
  “It’s hard <pause> to recognize <pause> speech”

- Real-time user correction provides correct context. Stops error propagation (Aist et al, ’07)
Error Correction: Marking Menu

- Edit operations accessed directly from the word via a marking menu, or simple gestures
- Single operation to edit errors
**Error Correction: Marking Menu**

- Edit operations accessed directly from the word via a marking menu, or simple gestures
- Single operation to edit errors
  - Delete: swipe left
Error Correction: Marking Menu

- Edit operations accessed directly from the word via a marking menu, or simple gestures

- Single operation to edit errors
  - Delete: swipe left
  - Substitute: swipe up (respeak, spell)
Error Correction: Marking Menu

- Edit operations accessed directly from the word via a marking menu, or simple gestures

- Single operation to edit errors
  - Delete: swipe left
  - **Substitute:** swipe up (respeak, spell)
    - OR
    - swipe down (alternates)
User Study & Hypotheses

- Controlled experiment to assess correction efficacy and usability of Voice Typing
- 2 x 2 within-subjects experiment (N = 24)
  - **Speech Interaction Model**: Voice Typing vs. Traditional Dictation
  - **Error Correction Style**: Marking Menu vs. Regular
- **Hypotheses**: Voice Typing outperforms traditional Dictation, and Marking menu outperforms Regular menu
Task: Compose Emails

- 5 emails per-condition (2 for practice, 3 for analysis)
  - E.g. “Write an email to your friend recommending a restaurant you like. Suggest a plate she should order and why she will like it.”
  - “Explain to your boss why you won’t be able to come into work today.”
Procedure & Types of Data

Recognizer Training

<table>
<thead>
<tr>
<th>Condition</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>25</td>
</tr>
<tr>
<td>B</td>
<td>25</td>
</tr>
<tr>
<td>C</td>
<td>25</td>
</tr>
<tr>
<td>D</td>
<td>25</td>
</tr>
</tbody>
</table>
User Correction Error Rate (UCER)

- UCER captures the amount of effort users made to correct errors.
- In Voice Typing, users made significantly lower corrections (10%) than Dictation (14%), $F(1,46) = 4.15, p=0.04^*$
  - Users naturally slowed down to monitor real-time text output, which helped accuracy.
4 Types of Corrections

- **Substitutions**: Respeak, spell, or alternates
- **Insertions**: Insert a word b/w two existing words
- **Deletions**: Deletion of a word, one at a time
- **Uncorrected**: Words identified incorrect, but left uncorrected

<table>
<thead>
<tr>
<th></th>
<th>Substitutions</th>
<th>Insertions</th>
<th>Deletions</th>
<th>Uncorrected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dictation, Marking Menu</td>
<td>7.35</td>
<td>2.14</td>
<td>3.17</td>
<td>0.21</td>
</tr>
<tr>
<td>Dictation, Regular</td>
<td>5.90</td>
<td>1.43</td>
<td>3.32</td>
<td>0.24</td>
</tr>
<tr>
<td>Voice Typing, Marking Menu</td>
<td>7.14</td>
<td>0.78</td>
<td>2.82</td>
<td>0.15</td>
</tr>
<tr>
<td>Voice Typing, Regular</td>
<td>5.10</td>
<td>1.36</td>
<td>3.10</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Difference in Number of Substitutions

- Significantly higher substitutions for marking menu than regular correction style, $F(1,46)=5.9$, $p=0.01$*

- Possibly because users preferred to substitute the word rather than leave it uncorrected
Lower Transcription Delay for Voice Typing

\[ \text{Delay} = \sum_{\text{all emails}} \frac{\text{Time (text)} - \text{Time (speech)}}{\text{Total Number of Emails}} \]

- Voice Typing = 1.27 sec; Dictation = 12.41 sec
  - Delay in dictation includes the time that the user took to speak the entire utterance, as well as the delay time.

- Delay in Voice Typing did not vary much across emails to affect user experience
  - Most emails within one S.D. of average; all within two S.D.
Wins for Voice Typing

- Users indicated Voice Typing as having lower mental demand, effort, and frustration
- 18 participants preferred Voice Typing over Dictation:

  “It [Voice Typing] was better because you *did not have to worry about finding mistakes later on*. You could see the interaction [output] as you say; thereby *reassuring you that it was working fine*.”
Losses for Voice Typing

- 6 participants disagreed because incorrect recognition disrupted thought flow in Voice Typing:

  “I preferred Dictation, because in Voice Typing, if one word was off as I was speaking, it would distract me.”
Wins & Losses for Marking Menu

- Marking Menu had lower physical and mental demand

- 21 participants preferred Marking Menu because:

  “It [Marking Menu] was great for a beginner. It was easier mentally to see the circle with choices and not have to concern myself with where to select my [correction] choices from.”

  “It [Marking Menu] seemed to involve less action.”

- 3 participants disagreed:
  - Had larger fingers than most; gestures on smaller words was difficult (e.g. single letter words “a”)

Microsoft Research
Discussion

- Real-time transcription of speech (as in Voice Typing) reduced user corrections
  - Naturally provided segmented information
  - Plausibly, when users correct transcriptions in real-time, it prevents errors from propagating
- Marking Menu preferred by most, yet had more substitutions
  - Preferred to substitute the word rather than leave it uncorrected
Discussion

- Dictation interaction model should not be dismissed
  - Useful when “eyes-free, hands-free” interaction required
e.g. driving
  - Other modalities like traditional keypad typing still useful for use in public spaces
Thank You

Questions?

Comments?

Feedback?