Just wanted to ask about the event on the 23rd.

—on Friday the 23rd.

Is the event still on?

Replace “the event” in the last sentence with “it”.

Just wanted to ask about the event on Friday the 23rd. Is it still on?
Existing Speech-to-Text Systems

Most current systems do not support editing through voice.

Ones that do:

- Nuance Dragon NaturallySpeaking
- Microsoft Word, Dictate
Limitation: Inflexible natural language for commanding

- Relies on wake words to activate command mode
- Users must memorize a list of commands
Limitation: Inflexible natural language for commanding

- Relies on wake words to activate command mode
- Users must memorize a list of commands

Want natural and intuitive dictation and commanding
New Task!: Interactive Dictation

1. Flexible interleaving of dictation and editing
   - No reserved trigger words for invoking commands
   - Challenge: Predicting segmentation between dictation and editing commands

Attached are the espeak events. Capitalize the S&E speak. Please review.
New Task!: Interactive Dictation

1. Flexible interleaving of dictation and editing
   - No reserved trigger words for invoking commands
   - Challenge: Predicting segmentation between dictation and editing commands

2. Intuitive and open-ended natural language for editing
   - No fixed templates for different types of command
   - Challenge: Interpreting which command to invoke and where/how

Attached are the eSpeak events. Capitalize the S&E speak. Please review.
Our Contributions

1. Introducing and formalizing a new task, Interactive Dictation

2. Designing a data collection interface and build a dataset for this task

3. Creating a baseline system for the task
Interactive Dictation: Basic Procedure
Interactive Dictation: Building a System

(a) ASR

Attached are the espeak events. Capitalize the S&E speak. Please review.
Interactive Dictation: Building a System

(a) ASR

(b) Segmentation

\[ U \]

Attached are the espeak events. Capitalize the S&E speak. Please review.

\[ u_1, \text{Dictation} \]

\[ u_2, \text{Command} \]

\[ u_3, \text{Dictation} \]

Attached are the espeak events. Capitalize the S&E speak. Please review.
Interactive Dictation: Building a System

(a) ASR

(b) Segmentation

(c) Normalization

Attached are the espeak events. Capitalize the S&E speak. Please review.

$u$

$u_1$, Dictation

Attached are the espeak events.

$u_2$, Command

Capitalized the S&E speak. Please review.

$u'_2$

Capitalized the S in eSpeak.
Interactive Dictation: Building a System

(a) ASR

(b) Segmentation

(c) Normalization

(d) Interpretation

$D_0$

$D_1$

Attached are the espeak events.

$D_2$

Attached are the eSpeak events.

$D_3$

Attached are the eSpeak events. Please review.

$u$

Attached are the espeak events. Capitalize the S&E speak. Please review.

$u_1$, Dictation

$u_2$, Command

$u_3$, Dictation

$u'_2$

Capitalize the S in eSpeak.
Annotating Commands & Transcriptions

Annotation Interface

Target document state $D_n$

Document state after selected segment $D_j$

Change in document state from segment $\Delta(D_{j-1}, D_j)$

Document state before selected segment $D_{j-1}$

Dictation Segments

Command Segments

Literal Utterance $u_s$

Normalized Utterance $u'_s$

Selected segment to edit $O_i$
Dataset: TERTIUS

11 annotators were instructed to do one of the following:

1. *Replicate doc*: exactly recreate an email from the Enron Email Dataset
2. *Elaborate doc*: expand a terse description of an email into a full email
3. *Replicate segment*: exactly recreate the effect of a single command segment sampled from annotations on the previous 2 objectives

<table>
<thead>
<tr>
<th>Trajectories</th>
<th>Segments</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dictation</td>
<td>Command</td>
<td>Total</td>
</tr>
<tr>
<td>1320</td>
<td>959</td>
<td>3225</td>
<td>4184</td>
</tr>
</tbody>
</table>
Interactive Dictation: Building a System

(a) ASR

(b) Segmentation

(c) Normalization

(d) Interpretation

\[ \mathcal{U} \]

Attached are the espeak events. Capitalize the S&E speak. Please review.

\[ u_1, \text{Dictation} \]

Attached are the espeak events.

\[ u_2, \text{Command} \]

Capitalize the S in eSpeak.

\[ u'_2 \]

Attached are the eSpeak events.

\[ u_3, \text{Dictation} \]

Attached are the eSpeak events. Please review.
Interactive Dictation: Instantiating Models

(a) ASR

1. Segmentation Model $M_{\text{SEG}}$

(b) Segmentation

$u_1$, Dictation

Attached are the espeak events. Capitalize the S&E speak. Please review.

$u_2$, Command

Attached are the espeak events. Capitalize the S&E speak. Please review.

$u_3$, Dictation

(c) Normalization

Has 2 modes:
1. \((\text{prog})\) Predict programs (which get executed by execution engine), or
2. \((\text{state})\) Direct predict end-states

T5 encoder trained for BIOES tagging to identify command boundaries.

T5 model trained to fix ASR and speech errors.

(d) Interpretation

$D_0$

$D_1$

Attached are the espeak events.

$D_2$

Attached are the espeak events.

$D_3$

Attached are the espeak events. Please review.

5. Execution Engine

$M_{\text{INT}}$ (non)

$u'_2$

Capitalize the S in eSpeak.
Results: Segmentation model

**Exact Match**: # of dialogues in which command boundary are exactly correct

<table>
<thead>
<tr>
<th>Model</th>
<th>Segmentation Exact-Match (Per dialogue)</th>
<th>Per-sample runtime (A100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T5-base Encoder only</td>
<td>85.3%</td>
<td>0.097 s/it</td>
</tr>
</tbody>
</table>
Results: ASR Repair + Interpretation Models

State Exact Match: # of commands for which the end-state is correctly predicted, whereby correctness is evaluated with exact string match.

<table>
<thead>
<tr>
<th>Model</th>
<th>State EM</th>
<th>Per-command Runtime (s/it)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T5 (prog)</td>
<td>28.3%</td>
<td>1.28</td>
</tr>
<tr>
<td>T5 (state)</td>
<td>29.5%</td>
<td>3.46</td>
</tr>
<tr>
<td>GPT3 (prog)</td>
<td>38.6%</td>
<td>3.52</td>
</tr>
<tr>
<td>GPT3 (state)</td>
<td>55.1%</td>
<td>6.92</td>
</tr>
</tbody>
</table>
Future Work

● Better evaluation of models
  ○ Human Evaluation
  ○ Partial credit
● Taking advantage of incrementality
● Include timing & prosody information in models
● Greater diversity in prompts and human voices
  ○ More open-ended prompts for more natural interactions
● Better/more flexible execution engine
● Model-in-the-loop annotation
  ○ Allows for data on clobber commands/redo/undo
Thank you!

- Code & Data: https://aka.ms/tertius
Conclusion

- We introduce a new task, interactive dictation, whereby:
  - 1. Users can naturally interleave dictation and commanding, and
  - 2. Users can flexibly invoke commands with a wide variety of utterances

- We construct a dataset TERTiUS for the task

- We build a baseline system for the task, discovering a tradeoff between speed and accuracy
  - We explore different choices and sizes of model architecture (T5 vs. GPT3)
  - We explore generating programs vs. generating document states directly