Why and How Our Reading Tutor Listens

1. How is reading related to LT and CAPT?
2. What is Project LISTEN’s Reading Tutor?
3. Why does it listen?
4. How does it listen?
5. What have we learned along the way?
1. What is “reading”? ~ LT? vs. CAPT?

Skills targeted by the Reading Tutor

1. Phonics
   - TTS: Decode
   - ASR: Spell

2. Fluency
   - OCR: Say words quickly, easily, accurately as usual – so tolerate
   - TTS: Expressively = like an adult vs. correctly – so detect

3. Vocabulary
   - WSD: Retrieve word meaning

4. Comprehension
   - NLU: Make meaning from print

My cat is fat
/kæt/
2. What is Project LISTEN’s Reading Tutor?

Uses Sphinx-II speech recognizer to track reader and detect mistakes
Responds with assistance modeled after expert reading teachers
The Reading Tutor addresses key problems in children’s reading

Key Problem:

Motivation

Phonemic Awareness

Phonics (Word Attack, Spelling)

Word Identification, Fluency

Vocabulary (Word Comprehension)

Comprehension

Reading Tutor Response:

Provide an active, patient, responsive audience

Praise achievement and improvement

Engage student in mixed-initiative interaction

Use visual speech to identify individual phonemes

Build, swap, sound out, compare, spell, echo

Supply spoken word or hint

Introduce, explain, spell, and read new words

Reread halting, disfluently read sentences

Teach strategies, insert questions
Project LISTEN’s Reading Tutor
3. Why the Reading Tutor listens

Detect speech and silence to guide turn-taking
Track reader’s position in text
Detect oral reading miscues
Measure word reading time
Compute prosodic contour
Assess student skills
Support research studies
Map Prosodic Featur...
Henry lived next door to Mrs. Johnson.
4. How the Reading Tutor listens

Acoustic model estimates $\Pr(\text{phoneme} \mid \text{signal})$

Lexical model gives phoneme sequence(s) for each word
  - Look up or synthesize pronunciations: W AH N S

Language model specifies $\Pr(\text{word} \mid \text{state})$
  - Given starting point,
    - Expect correct reading, but
  - Allow for deviations

Confidence score estimates $\Pr(\text{word} \mid \text{signal})$

Alignment model matches ASR output to text

Prosodic model estimates $\Pr(\text{prosodic contour} \mid \text{text})$
Types of research studies

Test overall effects by comparing pre- to post-test gains to control treatment(s)
- Reading Tutor rivaled human tutors in some skills (JECR03)
- Reading Tutor beat Sustained Silent Reading (SSSR02, JECR04)
- Reading Tutor helped ELLs in Canada (IDEC09), Ghana (ITID10), India (Dev10)

Assess student skills and validate against gold standard tests
- Prosody predicts proficiency (CALICO04, TICL04, TSLP11)
- Cloze questions predict vocabulary and comprehension (TICL04)

Evaluate tutor interventions using randomized within-subject experiments
- Test effects of vocabulary intervention on taught vs. untaught words (ETS02)
- Compare effects of tutor help on next encounter of word (SSSR04, ICALL04)

Identify important variables using correlational analyses
- Gains decrease with time spent picking stories (ITS02)
- Labs average higher usage than classrooms (Scale-up 07)

Compare types of practice using growth curve micro-analyses
- Wide reading builds fluency faster than rereading (SSSR05, ITS08)
- Students gained more from stories they picked (AIED07)

Mine tutor logs to discover useful knowledge
- Identify prosodic indicators of reading gains (FLAIRS12 Best Paper)
Acoustic models: *Quality trumps quantity.*

1. Semi-continuous HMMs for adult female speech.
2. Adapted codebook means for children’s oral reading.
3. Trained from scratch on transcribed oral reading.
4. Switched to continuous HMMs.
5. Augmented transcripts with “ASR-transcribed” data.
   - Heuristic filters achieved ~95% transcript accuracy.
   - Yet adding “ASR-transcribed” training data hurt!
Lexical models: *Distracters detract.*

Predict phonetic truncations as distracters [AAAI94]
- E.g. START\_ONCE = /W AH/
- Detected more miscues
- Didn’t increase false alarms

Predict likely miscues as distracters
- Train $G \rightarrow P \rightarrow P'$ malrules on Colorado miscues [AIED01]
- List miscues on top 100 words, e.g. *came* $\rightarrow$ *can*
- Learn which real-word miscues are likely

... detected more miscues only by hallucinating others.
From Miscues to Malrules [AIED01]

From miscues like

\[
\begin{align*}
\text{present} & \rightarrow P + R + \text{EH} + Z + \text{EH} + N + T/ \\
\text{learn} & \rightarrow P + R + \text{IY} + Z + \text{EH} + N + T/
\end{align*}
\]

learn \( G \rightarrow P \rightarrow P' \) malrules, e.g. \( e \rightarrow /\text{EH}/ \rightarrow /\text{IY}/. \)

Top 5 errors:

<table>
<thead>
<tr>
<th></th>
<th>( G )</th>
<th>( P )</th>
<th>( P' )</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( s )</td>
<td>/S/</td>
<td>_</td>
<td>plants</td>
</tr>
<tr>
<td>2</td>
<td>( s )</td>
<td>/Z/</td>
<td>_</td>
<td>arms</td>
</tr>
<tr>
<td>3</td>
<td>_</td>
<td>_</td>
<td>/N/</td>
<td>ha_d</td>
</tr>
<tr>
<td>4</td>
<td>$</td>
<td>_</td>
<td>/Z/</td>
<td>car_</td>
</tr>
<tr>
<td>5</td>
<td>( n )</td>
<td>/N/</td>
<td>_</td>
<td>land</td>
</tr>
</tbody>
</table>

_ = null

$ = \text{end of word}
Predicting by Rote [ICSLP02]

If at least 2 readers make the same miscue in the training data, predict others will too.

- E.g. misread “came” as “can”

Limited to word types in training set.
Idea: many miscues are real words

- E.g. misread “deems” as “dreams”
- ~65% of substitution miscues

Extrapolation task:

- Given: “deems” \(\rightarrow\) “dreams,” etc.
- Predict: “grotesque” \(\rightarrow\) ?
Approach: learn a classifier

Training data
- Positive examples: word-miscue pairs in training set
- Negative examples: other dictionary words

Features to distinguish likely from unlikely miscues
- Spelling/phonemic distance from correct word to miscue
- Word frequency – the rarer, the likelier to misread
- Miscue frequency – the more familiar, the likelier to say
- Level of reader, level of text

Learn to classify miscues as likely or not likely
- Decision stumps + LogitBoost [Friedman 98]
- Predict likely miscues, e.g. “grotesque” → “ghost”
Detecting Miscues: 
Listen for text words + distractors

Miscues spoken

Transcript:    for  little    dreams  my    real    dame
Correct text:  For    little    deems    my    royal    dame
ASR output:    FOR LITTLE    DREAM    MY    DEEMS    DAY

Miscues detected (not necessarily identified)

Miscue detection rate = # detected miscues / # actual miscues
False alarm rate = # false alarms / # correct mappings
## Pilot evaluation: Compare distractor sets

Data (recorded by Reading Tutors at schools)
2675 words read correctly
145 “miscues” include minor errors, non-words, typos, ...

<table>
<thead>
<tr>
<th>Distractor set</th>
<th>Distractors per word</th>
<th>Miscue detection rate</th>
<th>False alarm rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>None (just rest of sentence)</td>
<td>0</td>
<td>16%</td>
<td>2.5%</td>
</tr>
<tr>
<td>Truncations (used currently)</td>
<td>3.5</td>
<td>18%</td>
<td>3.1%</td>
</tr>
<tr>
<td>Miscues predicted by classifier</td>
<td>9.75</td>
<td>25%</td>
<td>4.3%</td>
</tr>
<tr>
<td>Random words</td>
<td>5.00</td>
<td>22%</td>
<td>6.4%</td>
</tr>
<tr>
<td>Cheating (actual miscues)</td>
<td>0.08</td>
<td>25%</td>
<td>2.8%</td>
</tr>
</tbody>
</table>
Confidence scores: Tracking trips up scoring.

ASR confidence score can detect unpredicted miscues. But confidence scores of mistracked words are useless.
Language models: *Rely on realism.*

Allowing word $i \rightarrow$ any word $j$
- Models any trajectory
- But leads ASR astray

Allowing only word $i \rightarrow$ word $i+1$
- Minimizes temptation
- But models other behavior poorly
- So leads ASR astray

All-phone model
- Matches any trajectory or miscue
- But leads ASR astray

Model children’s oral reading better to track it better
Alignment models: *Mask mistakes.*

“No problem is too big to run away from.”

– Lewis Johnson 6/8/2012, quoting Linus

Tracking the reader is hard, so conceal mistracking:

When reader reads:

- Don’t “chase the kid” by showing (possibly wrong) position.
- Just “blame the kid” by showing which word to read next.

When reader gets stuck:

- Don’t give hint on expected word, as tracker may be wrong.
- Just prompt to click on a word for help; let reader track.
Prosodic models: *Silences are golden.*

Fluent oral reading prosody is like adult [Schwanenflugel].

Example: They have to put food on the plane so that we can eat.

- Adult narration
- Fluent child
- Disfluent child

Pauses between words help assess reading fluency.

- Word’s latency drops from first to last encounter [AAAI97].
- Distribution of word latencies predicts test scores [TICL04].
- Correlation to adult durations predicts test scores [AIED09].
- Correlation to expected durations predicts better [TSLP11].
- Scoring against adult prosody model is better yet [FLAIRS12].
5. Lessons and questions

Acoustic models: *Quality trumps quantity.*
- Adding “ASR-transcribed” training data hurt. Can it help?

Lexical models: *Distracters detract.*
- Predicting miscues hallucinates them. Which are worth it?

Confidence scores: *Tracking trips up scoring.*
- Scoring mistracked words is useless. Can it help anyway?

Language models: *Rely on realism.*
- Model children’s oral reading better to track better. How?

Alignment models: *Mask mistracking.*
- Don’t follow reader; just lead or prompt. How fix interface?

Prosodic models: *Silences are golden.*
- Pauses before words predict test scores. How predict better?
Thank you! Questions?