Comparing Passage Lengths and Human vs Speech Recognition Scoring of Oral Reading Fluency

Joseph F. T. Nese, Julie Alonzo, Akihito Kamata
University of Oregon, Behavioral Research and Teaching
Southern Methodist University

Purpose
The purpose of this study is to explore a computerized oral reading fluency (ORF) system that uses speech recognition software (CORE). The purposes were as follows:

1. Compare the mean WCPM scores across:
   a) three passage lengths (short=25 words, medium=50 words, long=85 words), and
   b) three scoring methods (real-time, audio recording, and ASR).
2. Compare the error rates across the passage lengths.
3. Compare the timing duration of read passages between: human assessors in real time as in traditional ORF, and computer estimates.
4. Analyze the agreement of word-level scores (correct or incorrect) across the three scoring methods.

Method

Sample. Students’ response times < 2.5 secs were removed. Students’ WCPM scores > 1.9 times different between the three scoring methods were removed. As a result, five Grade 2 and seven Grade 3 students were removed from the analysis sample. Sample sizes were 127 for Grade 2, 158 for Grade 3, and 162 for Grade 4.

Passages. Administered via computer: 18 passages (3 long, 5 medium, 10 short).
1. (Short) =25 words, read in entirety.
2. (Medium) =50 words, read in entirety.
3. (Long) =85 words, read in entirety.

Scoring. Word accuracy and words correct per minute (WCPM) were scored by:
1. (Real-time): trained human assessors as in traditional ORF.
2. (Audio Recording): trained human assessors via audio recordings.
3. (Automated Speech Recognition (ASR)).

Analyses. Mixed model approach with two within-subject variables to test the mean WCPM and error rate differences between passage length, scoring method, and their interaction. The length factor included three categories short, medium, and long. The scoring method factor included three categories: Real-Time, Recorded Audio, and ASR.

Cohen’s kappa$^2$ was used to analyze the agreement between word-level scores across the scoring methods.

Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Grade 2</th>
<th>Grade 3</th>
<th>Grade 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real-Time</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audio Recording</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASR</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Across grades, significant main effects for passage length and scoring method, no significant interaction effect, and mixed results for pairwise comparisons.

The mean error rates ranged from 3% to 10%, when they were disaggregated by grade, scoring methods, and passage length.

Real-Time scoring was higher than both the ASR and Recorded.

Error rates were lower for Real-Time scoring.

Error rates were lower for lower grade levels.

Conclusions

- The timed passage duration was consistently greater (approximately 1-2 secs) for the ASR scoring methods than the Real-Time scoring method.
- Because the ASR and Recording scoring methods used the same time duration to compute WCPM, this would lead to decreased WCPM scores compared to the Real-Time scoring method.
- Although it is not possible to determine the “true” passage timed duration, we hypothesize that the ASR computer-generated time would be the most accurate. The ASR time is the duration from the utterance of the first passage word, to the termination of the last word read, in centi-seconds.

Funding Sources
The research reported here was supported by the Institute for Education Sciences, U.S. Department of Education, through Grant P305A140203 to the University of Oregon. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Further Information
jnes@uoregon.edu | http://brtprojects.org

References

- The recorded Audio may need additional training, because scoring could take place in a quiet setting with no distractions, and the capability to reread the recording to ensure the most accurate word scores. The primary interest was to compare the Recorded Audio scores to both the Real-Time and ASR scores.
- The agreement rate was quite varied between students for ASR vs. Real-Time, and for ASR vs. Recording.
- The two human scores (Recording and Real-Time) had the highest kappa agreement.
- Average ASR vs. Real-Time Cohen’s kappa: Grade 2 = .82, Grade 3 = .90, and Grade 4 = .91.
- The ASR may need additional training, especially for lower grade levels.
- Next step is to investigate instances of low agreement.
- Is kappa = .90 “good enough” if ASR can save considerable resources (time, money, instruction)?