Sub-sequence Matching Algorithm
Improving Automated Speech Recognitions for ORF Assessment
Background
Oral Reading Fluency

The ability:
- to read a text quickly, accurately, and with proper expression

A guideline to:
- evaluate and student’s progress
- guide the teachers’ intervention

Three dimensions:
- accurately word decoding
- quick and automatic recognition
- expressive interpretation

Automaticity theory
Challenges
Three problems in Recent Research

students’ pronunciations & phrasing
sudden changes in reading speed, mumbling, or soft speech

speech recognizers exacerbate
language model corrects these odd phrasings or suggests non-relevant words

“passage-specific” limitation
To evaluate fluency with arbitrary words.
Goals

• Enable automated evaluation of oral reading fluency using modern speech recognition systems, such as Apple and Amazon transcription services.

• The system should allow inputting any student response with a reference text.
  
    ▶ i.e., no passage-specific calibration needed, no student-specific calibration needed

• Obtain voice characteristics to increase accuracy at both sentence level and word level.
Dataset
Nese and Kamata (2014-2020)

Dataset Contains human scoring of “words correct” in passages from approximately 2,000 second through fourth-grade students.

<table>
<thead>
<tr>
<th></th>
<th>Total Records</th>
<th>Total Students</th>
<th>Different Passages</th>
<th>Unique Words</th>
<th>Median Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid Response</td>
<td>14339</td>
<td>662</td>
<td>319</td>
<td>1674</td>
<td>22 sec</td>
</tr>
<tr>
<td>Reasonable Response</td>
<td>13029</td>
<td>653</td>
<td>319</td>
<td>1674</td>
<td>22 sec</td>
</tr>
<tr>
<td>(&gt;10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Methodology
Sequence Matching, Baseline

1. Use Apple Speech Recognizer
   - to generate transcripts and timestamps related to each word and potential options.

2. Use the LCSS script
   - to match and calculate the difference between the reference script and the student response.

3. Calculate the statistics from matching results
   - such as accurate words per minute.
Methodology
Feature Extraction

- Use Apple Siri Voice Text to Speech to generate reference speech with different voice options.
- Use Apple ASR to generate the transcription texts and timestamps of the Siri Voice reference speech.
- Use the same LCSS to match the reference transcription and student transcription. Paring the time slot has exact words.
- Generate the Mel-Spec of both students and reference.
Methodology

Network Setup

- The first part of the network is an auto-encoder that generates feature maps for different resolutions. A region propose network proposes an interval that is a candidate for a word.

- The content Vector is related to each word, which performs Region of Word Pooling based on the candidate interval. It will generate a scale-invariant feature.

- Content Vector can be considered as a word-dependent, speaker-dependent feature.
AutoEncoder
Detail of AutoEncoder

- The Auto Encoder is a U-Net style Network consisting of Encoder and Decoder series. At the end of each encoder, the resolution will change to create multi-resolution hierarchically.

- Region proposed network generates the interval of each word.

- All Decoder features will pool into a content vector according to the region proposed.

- The output of the decoder can be considered as a frame-based feature.
TC1(W2) is an attention-based network try to predict the between-word relationship.

Deep Feature is word related feature that students responses and reference Siri Voice. And the contrastive loss will be applied, which creates word-dependent, speaker-independent features.
Methodology
Network Setup

- TC1(W3) is a recovery network that recovers text information (Text itself or Phono embedding) from reference text.
- This is used to recover a text that is known. Which can help network for training.
- Deep Feature work for a word that is unknown.
Metrics

Word Correct per minutes

- Dynamic Indicators of Basic Early Literacy Skills (DIBELS) using Word correct vs. total word correct in a minute as an Oral Fluency score
  - https://dibels.uoregon.edu/sites/dibels1.uoregon.edu/files/2021-05/dibels_8_admin_and_scoring_guide_05_2020.pdf (page 76)

- Curriculum-Based Measurement (CBM) counts the number of words that reads correctly in 1 minute as a score.
Results
Sub-sequence matching only

- Use Word Correct Per Minutes to estimate the student response.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Original</th>
<th>Stemmed</th>
<th>Original</th>
<th>Stemmed</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>19.44</td>
<td>17.97</td>
<td>16.56</td>
<td>15.14</td>
</tr>
<tr>
<td>MAPE</td>
<td>40.03</td>
<td>37.96</td>
<td>15.75</td>
<td>14.59</td>
</tr>
<tr>
<td>MSE</td>
<td>1263.34</td>
<td>1165.95</td>
<td>579.36</td>
<td>516.63</td>
</tr>
<tr>
<td>MSPE</td>
<td>23255.87</td>
<td>23234.14</td>
<td>6.16</td>
<td>5.90</td>
</tr>
<tr>
<td>MAE middle 60%</td>
<td>12.75</td>
<td>11.28</td>
<td>12.13</td>
<td>10.62</td>
</tr>
<tr>
<td>MAPE middle 60%</td>
<td>13.53</td>
<td>12.11</td>
<td>12.10</td>
<td>10.67</td>
</tr>
</tbody>
</table>
Result

Sub-sequence matching only (Plots)
Future Works

• Apply the model to estimate the word output that the student and reference disagree with. To confirm if the model can correct the output of the Commercial ASR.

• Apple model and using feature similarity on unknown words when given reference voice to check how the model will generalize to off-domain words.

• Using a better Region Proposal method that combines with Attention to get a better region to propose.

• Investigate the Frame-Based Feature and Content-Vector in other applications.

• Investigate Phonetic related features.
Oral Reading Fluency

• Oral Reading Fluency is defined as the ability to read a text quickly, accurately, and with proper expression (National Reading Panel, 2000)

• Provides a guideline to evaluate and monitor the students’ progress and guide the teachers’ intervention for the students who are at risk of reading difficulties (Al Otaiba & Fuchs, 2006; Al Otaiba et al., 2009; Fuchs, Fuchs, & Compton, 2004; Hagans, 2008; Zutell & Rasinski, 1991)

• Rasinski identified that reading fluency has three dimensions: accuracy in word decoding, quick and automatic recognition of words in connected text, and expressive and meaningful interpretation of the text (Rasinski, 2004; Rasinski et al., 2009).

• Automaticity theory (Rasinski, 2004)
Background
Recent Research and Challenges

1. Many approaches employ “passage-specific” methods to evaluate fluency with arbitrary words.

2. Many students manifest unique pronunciations and phrasing while reading aloud (i.e., sudden changes in reading speed, mumbling, or soft speech) that need to be better modeled by commercial speech recognizers. (Hudson et al., 2005).

3. Modern commercial speech recognizers exacerbate these issues because they typically employ a language model that accidentally corrects these odd phrasings or suggests non-relevant words.
Reference
Others Work

• Benjamin et al., 2013; Schwanenflugel & Benjamin, 2017; use Praat software to match pitch and pause patterns (http://www.praat.org/, Boersma & van Heuven, 2001; Boersma & Weenink, 2021)

• Black et al. (2011); Tepperman, Lee, Narayanan, and Alwan (2011); focus on automatically scoring the student’s reading skills based on their pronunciation evaluation at the word level.

• Bolaños et al. (2013); Sabu and Rao (2018)

• Morrison and Wilcox (2020) pointed out that some researchers concern that current automated assessments of prosody have limited to measuring only pitch and pauses, but less assessing smoothness and phrasing of prosody that human evaluators may be better to do.