The Effects of Entrainment in a Tutoring Dialogue System

Huy Nguyen, Jesse Thomason
CS 3710 – University of Pittsburgh
Outline

• Introduction
• Corpus
• Post-Hoc Experiment
• Results
• Summary
Introduction

• Spoken dialogue systems can offer students one-on-one instruction from a computer tutor

• Student entrainment to computer tutor voice has been shown to correlate with learning gain (Ward and Litman, 2007; 2008)

• A system encouraging or responding to entrainment might lead to better student performance
Introduction

• The CMU Let’s Go!! bus information system elicited user entrainment to improve speech recognition (Raux and Eskenazi, 2004)
• For tutoring systems, knowing which entrainment features are correlated with learning could inform this strategy
• We searched an existing intelligent tutoring dialogue system corpus to find such correlations
Outline

• Introduction
• Corpus
• Post-Hoc Experiment
• Results
• Summary
Corpus

• Our data comes from a 2005 experiment with ITSPOKE
• Each student interacted with either a pre-recorded or synthesized tutor voice (Forbes-Riley et al., 2006)
• Students responded to tutor questions both verbally and with written essays for 5 problem dialogues
Corpus

• We omit Students who started but did not complete a problem in a past session

• This left us with 26 students

• Effects of tutor voice, but not entrainment, were examined in (Forbes-Riley et al., 2006)
Corpus and Motivations

• Student pre- and post-test scores, satisfaction evaluations of the system, ASR word-error rate per student, and other student metadata were available.

• We investigate whether the level of student entrainment had any correlation with learning gain, user satisfaction, or word-error rate.
Corpus and Motivations

• Whether student entrainment differed significantly between the pre-recorded and synthesized voices was also of interest.

• Inspired by (Pardo, 2006), we were also interested in the relationship between user gender and entrainment.
Outline

• Introduction
• Corpus
• Post-Hoc Experiment
• Results
• Summary
Hypotheses

1. a positive correlation between entrainment and learning gain
2. a positive correlation between entrainment and user satisfaction
3. a negative correlation between entrainment and word-error-rate
4. higher entrainment coefficients for students interacting with the pre-recorded tutor voice
5. higher entrainment coefficients for males
Entrainment Features

- Lexical and prosodic
- Lexical based on coarser-grained, free-form student essays
- Prosodic based on finer-grained, exchange-level student utterances
- All entrainment scores calculated on a per-problem basis, then averaged to obtain student entrainment value
Lexical Entrainment Features

• We take word repetition as primary measurement of entrainment
  – Not counting repeated words between turns

• ITSPoke tutoring format:
  – Student reads the problem, writes initial essay
  – Computer tutor evaluates, guide to improve
  – Student re-writes the essay, submit again

Reference essay
T-S conversation
Edited essay
Observation 1

• Students' answers are typically short

- I don't know
- yeah
- yeah
- sun is stronger than earth's
- opposite
- yes
- yes
- yes
- they're equal
Observation 2

• Learning evidence are shown by occurrence of new terms, and lost of other terms

Reference essay:
No the earth does not pull equally on the sun. The mass of the earth is much smaller than the sun. So it pulls with a smaller force. This is why the earth orbits the sun.

Edited essay:
No the earth does pull equally on the sun because of Newton's Third Law. The force is gravitational. It is equal and opposite.
Lexical entrainment as understanding to suggestions

• Knowledge entrainment through language

• Consider non-stop words
  – in tutor responses
  – appear in edited essay
  – but not in reference essay

• Also, non-stop words
  – appear in reference essay
  – but not in edited essay
Three metrics

1. new-word:

\[
\text{mean} \left( \frac{\text{number of new words}}{\text{number of tutor responses}} \right)
\]

2. new+removed-word:

\[
\text{mean} \left( \frac{\text{number of new words} + \text{removed words}}{\text{number of tutor responses}} \right)
\]

3. essay-length:

\[
\text{mean} \left( \frac{\text{length(\text{reference}\_\text{essay})} - \text{length(\text{edited}\_\text{essay})}}{\text{number of tutor responses}} \right)
\]
Prosodic Entrainment Features

• Our method is inspired by the metric used to find entrainment in (Ward and Litman, 2007)
  – Itself inspired by the method in (Reitter et al., 2006)

• openSMILE to get mean, min, max, and standard deviation of the energy (RMS) and pitch (F0) of every utterance
Prosodic Entrainment Features

- Strict turn-taking offers verbal student responses to most tutor utterances
- We created progressive, exchange-level similarity scores between the student and tutor
- We used a linear regression to find the change in those similarity scores throughout each dialogue
Prosodic Entrainment Features

• For each problem dialogue and raw prosodic feature, our algorithm is implemented as follows
Student RMS mean

Tutor RMS mean

Number of exchanges (i)

$r^2$ similarity score

$i=3$

$r^2=0.0739$

$i=4$

$r^2=0.2452$

$i=25$

$r^2=0.2970$

$r = 0.8158$

$r^2=0.666$
Experimental Methods

• We looked for significance in:
  • Correlations entrainment scores and student properties relevant to hypotheses
  • Those same correlations for low and high pre-testers (using a median split)
  • Differences in mean between users’ entrainment in the pre-recorded and synthesized voice conditions and between male and female entrainment to the system
Experimental Methods - Control

• Re-performed these tests on a randomized baseline corpus

• Tutor turns remained in place as student responses were randomly paired with tutor turns from which they did not originally follow

• No relationships which appeared significant in the original corpus appeared in the randomized corpus
Experimental Methods - Metrics

• For learning gain, we considered:
  – Standard Learning Gain (SLG)
    • $post - pre$
  – Normalized Learning Gain (NLG)
    • $(post - pre) / (1 - pre)$

• User satisfaction, **UsrSat**, based on sum of survey questions in (Forbes-Riley et al., 2006)
Outline

• Introduction
• Corpus
• Post-Hoc Experiment
  • Results
• Summary
Results and Discussion

• We denote:
  • Significant (p < 0.05) results with *
  • Highly significant (p < 0.01) results with **
  • All other shown results are trending (p < 0.1)

• 12 Low pre-test student (under median)
• 10 High pre-test student (above media)
Support Hypothesis 1

• “a positive correlation between entrainment and learning gain”

• When considering all students, we found:

<table>
<thead>
<tr>
<th>Student Data</th>
<th>Entrainment</th>
<th>(r-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLG</td>
<td>new+removed−word</td>
<td>0.447*</td>
</tr>
<tr>
<td>SLG</td>
<td>essay−length</td>
<td>0.348</td>
</tr>
<tr>
<td>NLG</td>
<td>new+removed−word</td>
<td>0.382</td>
</tr>
</tbody>
</table>

• We note that prosodic features were not found indicative of learning gain
Support Hypothesis 2

• “a positive correlation between entrainment and user satisfaction”

• With respect to **UsrSat**, we found mostly positive correlations with prosodic features:

<table>
<thead>
<tr>
<th>Group</th>
<th>Entrainment</th>
<th>(r-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>RMS max</td>
<td>0.536**</td>
</tr>
<tr>
<td>Low pre-tester</td>
<td>F0 max</td>
<td>0.623*</td>
</tr>
<tr>
<td>Low pre-tester</td>
<td>RMS max</td>
<td>0.554</td>
</tr>
<tr>
<td>Low pre-tester</td>
<td>F0 mean</td>
<td>-0.533</td>
</tr>
</tbody>
</table>
Reject Hypothesis 3

• “a negative correlation between entrainment and word-error-rate ”
• WER often did not correlate at all
• When considering high pre-testers, we found:

<table>
<thead>
<tr>
<th>Student Data</th>
<th>Entrainment</th>
<th>(r-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER</td>
<td>RMS mean</td>
<td>0.771**</td>
</tr>
<tr>
<td>WER</td>
<td>RMS stddev</td>
<td>0.693*</td>
</tr>
</tbody>
</table>
Support Hypotheses 4,5

• “higher entrainment coefficients for students interacting with the pre-recorded tutor voice”
  – \textit{RMS mean}\(^*\) and \textit{RMS stddev} entrainment higher in the pre-recorded voice condition

• “higher entrainment coefficients for males”
  – \textit{F0 min} entrainment higher among males
Outline

• Introduction
• Corpus
• Post-Hoc Experiment
• Results
  • Summary
Summary

1. a positive correlation between lexical entrainment and learning gain
2. a positive correlation between prosodic entrainment and user satisfaction
3. a negative correlation between prosodic entrainment and word-error-rate
4. higher prosodic entrainment for students interacting with the pre-recorded tutor voice
5. higher prosodic entrainment coefficients for males
Summary

• We support existing claims that:
  – entrainment may affect student performance in intelligent spoken tutor dialogue systems
  – tutor voice and gender both play roles in entrainment

• Our findings suggest that:
  – dialogue-level entrainment correlates with learning gain and trends against satisfaction
  – short-term, prosodic entrainment correlates with satisfaction

• Encouraging entrainment from their users may elicit higher learning gain and user satisfaction
  – the duration of that elicited entrainment must be considered
The Effects of Entrainment in a Tutoring Dialogue System

Huy Nguyen, Jesse Thomason
CS 3710 – University of Pittsburgh
## All Correlations

<table>
<thead>
<tr>
<th>Student Data</th>
<th>Entrainment</th>
<th>(r-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLG</td>
<td>new+removed–word</td>
<td>0.447*</td>
</tr>
<tr>
<td>SLG</td>
<td>RMS min</td>
<td>-0.367</td>
</tr>
<tr>
<td>SLG</td>
<td>essay–length</td>
<td>0.348</td>
</tr>
<tr>
<td>NLG</td>
<td>RMS min</td>
<td>-0.558**</td>
</tr>
<tr>
<td>NLG</td>
<td>new+removed–word</td>
<td>0.382</td>
</tr>
<tr>
<td>UsrSat</td>
<td>RMS max</td>
<td>0.536**</td>
</tr>
<tr>
<td>UsrSat</td>
<td>new–word</td>
<td>-0.330</td>
</tr>
</tbody>
</table>

Student data correlated with entrainment features
Low pre-test correlation

<table>
<thead>
<tr>
<th>Student Data</th>
<th>Entrainment</th>
<th>(r-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UsrSat</td>
<td>F0 max</td>
<td>0.623*</td>
</tr>
<tr>
<td>UsrSat</td>
<td>RMS max</td>
<td>0.554</td>
</tr>
<tr>
<td>UsrSat</td>
<td>F0 mean</td>
<td>-0.533</td>
</tr>
</tbody>
</table>

Low pre-test student (12 total) data correlated with entrainment features
High Pre-test Correlations

<table>
<thead>
<tr>
<th>Student Data</th>
<th>Entrainment</th>
<th>(r-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLG</td>
<td>RMS min</td>
<td>-0.708*</td>
</tr>
<tr>
<td>SLG</td>
<td>F0 stddev</td>
<td>0.582</td>
</tr>
<tr>
<td>NLG</td>
<td>RMS min</td>
<td>-0.720*</td>
</tr>
<tr>
<td>NLG</td>
<td>RMS mean</td>
<td>0.554</td>
</tr>
<tr>
<td>WER</td>
<td>RMS mean</td>
<td>0.771**</td>
</tr>
<tr>
<td>WER</td>
<td>RMS stddev</td>
<td>0.693*</td>
</tr>
</tbody>
</table>

High pre-test student (10 total) data correlated with entrainment features
Tutor Voice and Gender

• Voice: $RMS \text{ mean}^*$ and $RMS \text{ stddev}$ entrainment higher in the pre-recorded (12 students) than synthesized (14 students) condition

• Gender: $F0 \text{ min}$ entrainment higher among males (11 students) than females (15 students)