# Acoustic-Prosodic Entrainment and Rapport in Collaborative Learning Dialogues

Nichola Lubold
CIDSE
Arizona State University
Tempe, AZ, USA
nlubold@asu.edu

Heather Pon-Barry
Department of Computer Science
Mount Holyoke College
South Hadley, MA, USA
ponbarry@mtholyoke.edu

## ABSTRACT

In spoken dialogue analysis, the speech signal is a rich source of information. We explore in this paper how low level features of the speech signal, such as pitch, loudness, and speaking rate, can inform a model of student interaction in collaborative learning dialogues. For instance, can we observe the way that two people's manners of speaking change over time to model something like rapport? By detecting interaction qualities such as rapport, we can better support collaborative interactions, which have been shown to be highly conducive to learning. For this, we focus on one particular phenomenon of spoken conversation, known as *acoustic-prosodic entrainment*, where dialogue partners become more similar to each other in their pitch, loudness, or speaking rate during the course of a conversation. We examine whether acoustic-prosodic entrainment is present in a novel corpus of collaborative learning dialogues, how people appear to entrain, to what degree, and report on the acoustic-prosodic features which people entrain on the most. We then investigate whether entrainment can facilitate detection of *rapport*, a social quality of the interaction. We find that entrainment does correlate to rapport; speakers appear to entrain primarily by matching their prosody on a turn-by-turn basis, and *pitch* is the most significant acoustic-prosodic feature people entrain on when rapport is present.

## Categories and Subject Descriptors

K.3.1 [**Computers and Education**]: Computer Users in Education

## Keywords

entrainment; rapport; spoken dialogue; collaborative problem solving

*MLA'14,* November 12, 2014, Istanbul, Turkey.
ACM 978-1-4503-0488-7/14/11. http://dx.doi.org/10.1145/2666633.2666635.

## 1. INTRODUCTION

One characteristic that makes spoken dialogue a powerful mode of communication is that when people engage in conversation, speakers can convey metacommunicative information to their listeners through the speech signal, by how they speak. For example, they may modulate their pitch or tone, speak faster or slower, louder or softer. These changes convey information and make the speech signal itself a rich source of information for analyzing interactions. In the field of multimodal learning analytics, there is great potential for studying the speech modality in conjunction with other modalities [20].

One particular phenomenon of conversational dialogue that has been shown to be correlated to learning [11, 24] is called *entrainment.* Also known as accommodation, adaptation, or alignment, entrainment occurs when dialogue partners become more similar to each other during the course of a conversation. People can entrain both on the words they use as well as the how they say them. In this paper, we are interested in the latter which is called *acoustic-prosodic entrainment* and occurs when people adapt their pitch, loudness, or speaking rate.

In addition to learning, past studies have linked entrainment with dialogue quality, task success, and certain social behaviors [12, 15, 21]. We are interested in looking at how we can use acoustic-prosodic entrainment to support student learning in collaborative interactions by detecting interaction qualities such as whether students have rapport. Within educational interactions, the existence of rapport between students has been shown to lead to greater learning gains [19]. In collaborative interactions, which have been shown to be highly conducive to learning [6], rapport can enable stronger collaboration and engagement [4], potentially leading to greater learning.

Because of the relationship between rapport and collaborative learning, we explore using acoustic-prosodic entrainment to detect rapport in collaborative learning scenarios. Our hypothesis that there is a relationship between entrainment and rapport is motivated by Tickle-Degnen and Rosenthal's coordination-rapport theory, which posits that nonverbal coordination should correlate with the amount of liking between conversational partners [25]. In addition, Lakin et al. found that there is a bi-directional relationship between rapport and people's tendency to unconsciously adopt the postures, gestures, and mannerisms of their partners [14].

In this paper, we have two goals regarding acoustic-prosodic entrainment and rapport. The first goal is to identify whether there is evidence of acoustic-prosodic entrainment in human-

human collaborative learning dialogues. If entrainment exists, we examine which aspects and features are the most prominent. The second goal, given that acoustic-prosodic entrainment is present, is to investigate whether acoustic-prosodic entrainment can help us model the interaction between the two students by detecting the presence or absence of rapport, one social factor of the interaction.

The data set for our analysis is a novel corpus of human-to-human peer learning dialogues. We capture both self-reported feelings of rapport as well as perceptual judgments of rapport, at the level of the observer. We explore the data set for three different forms of entrainment using four acoustic-prosodic features. We find that all three exist in the dialogues and, in-line with previous findings on other corpora, that *intensity* or loudness is the most significant acoustic-prosodic feature which people entrain on when looking at the corpus as a whole. Finding that entrainment is present in the dialogues, we then explore the relationship between entrainment and rapport, and we find that *pitch* (F0) is the most significant acoustic-prosodic feature people entrain on when there is rapport.

In the next section, we introduce related work on entrainment and rapport. In Section 3, we describe collecting the dialogues, the methodology for measuring entrainment and rapport, and the statistical tests we employed. Section 4 presents our results, and we end in Section 5 with a discussion of the results and future work.

## 2. RELATED WORK

People can entrain in various ways, becoming more similar in their body language or facial expressions [2,5,17,18]; however, entrainment in speech is one of the most common forms. Within learning applications, the focus on acoustic-prosodic entrainment has been primarily on human-computer dialogue [7,28], which is why we choose to focus here on human-human collaborative learning dialogues.

Methods for quantifying the level of entrainment in speech are varied. Looking specifically at acoustic-prosodic entrainment, it can be measured by how aligned the speakers are at each turn compared to any other point in the conversation, sometimes called *proximity*. The amount of similarity between the speakers may change in tandem, *synchronously*, or it might *converge*, with the speakers becoming increasingly similar over time. To measure entrainment in this paper, we choose to leverage the methodology proposed by Levitan and Hirschberg which utilizes all three measures [16]. Previous works have utilized one or two of these measures; we investigate all three aspects of entrainment on a turn-by-turn level and describe this more in the methodology.

One contribution of this paper is the exploration of the detection of rapport using acoustic-prosodic features of speech. Past works investigating rapport have primarily focused on building rapport through non-verbal signals such as nodding or smiling [3,9] or through acoustic-prosodic features such as loudness and speaking rate [1,23]. Detecting rapport using acoustic-prosodic features of speech has received less attention.

## 3. METHODOLOGY

This section describes the data collection process, preparing the speech data, and introduces our approaches for measuring entrainment and rapport.

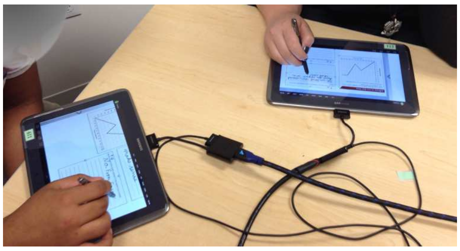


Figure 1: *Collaborative problem-solving with FACT.*

### 3.1 Data Collection

To model entrainment and detect rapport in collaborative learning, we collect a set of eight 30-40 minute dialogues from 16 undergraduate college students. The students work together in pairs as peers. We give each student a tablet containing a version of the Formative Assessment with Computation Technologies (FACT) application.[1] The application encourages collaborative interaction through the use of a shared workspace, shown in Figure 1, where students can simultaneously write and see each other's changes.

The application is designed to support and provide formative assessment for K-12 students solving mathematical problems. The mathematical problems available in the FACT application are part of the Mathematics Assessment Project.[2] The problems are designed with a goal to make knowledge and reasoning visible; the iterative refinement required to solve the problem is intended to generate conversation and drive collaboration between the students as seen in the sample dialogue is below.

A: Ohhh ... negative. Wait, this doesn't help anything
B: Well it's just a bad equation because it's a fraction
A: I clearly can't do this
B: No it's okay we can do it. So $y$ equals 10 minus $x$
...I mean negative $x$ plus 10

The student volunteers were undergraduate students with basic knowledge of algebra and geometry. They do not receive any mathematics-based training before the experiment. Given the knowledge level of the students, we gave them problems at grade level nine and above. Figure 2 shows an example problem.

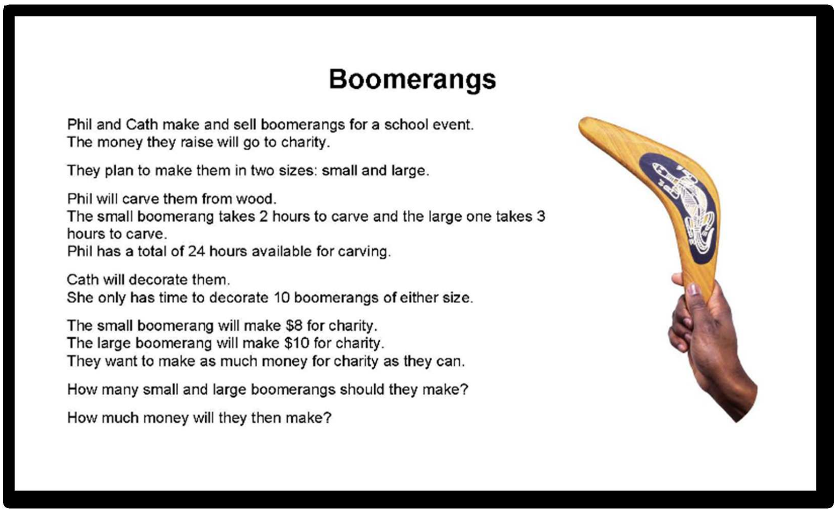
To ensure the subjects are able to use the application, they first individually complete a ten minute introduction to the tablet, the application, and its capabilities. We then give them two FACT application math problems and instruct them to work together.

### 3.2 Data Preparation

We record high-quality audio data, using unidirectional microphones with two separate audio channels, one channel for each speaker. For the analysis at hand, we manually select four two-minute segments from each dialogue (32 segments in total). These segments optimize the amount of dialogue pertaining to the math problems and minimize the amount of silence. We manually annotate turn boundaries in each two-minute segment, defining a turn as a continuous speech utterance by a single speaker, including filled pauses and laughter [26].

[1] http://fact.engineering.asu.edu/

[2] http://map.mathshell.org/materials/index.php



Boomerangs

Phil and Cath make and sell boomerangs for a school event.
The money they raise will go to charity.

They plan to make them in two sizes: small and large.

Phil will carve them from wood.
The small boomerang takes 2 hours to carve and the large one takes 3 hours to carve.
Phil has a total of 24 hours available for carving.

Cath will decorate them.
She only has time to decorate 10 boomerangs of either size.

The small boomerang will make $8 for charity.
The large boomerang will make $10 for charity.
They want to make as much money for charity as they can.

How many small and large boomerangs should they make?

How much money will they then make?

Figure 2: *Screenshot of an example MAP problem from the FACT application.*

We then further segment each turn into inter-pausal units or IPUs. An IPU is a pause-free unit of speech separated from any other speech by at least 50ms (see [16]). Turns are composed of one or more IPUs. For example,

> B: No it's okay we can do it. So $y$ equals 10 minus $x$ . . . I mean negative $x$ plus 10

is composed of two IPUs where the first IPU is "No it's okay we can do it. So $y$ equals 10 minus $x$" is the initial IPU of the turn followed by a pause greater than 50ms and the final IPU of the turn "I mean negative $x$ plus 10."

### 3.3 Acoustic-Prosodic Features

To determine whether entrainment exists in the dialogues, we focus on four acoustic-prosodic features: **intensity**, **pitch**, **voice quality** and **speaking rate**. We extract all four of these features from each IPU.

To extract intensity, pitch, and voice quality, we use a tool called openSMILE [10]. Using openSMILE, we actually extract one feature to represent intensity, one feature for pitch, and three for voice quality (local jitter, differential jitter, and shimmer)[3]. In addition, we extract several functionals for each feature, like the mean, maximum, and minimum. Table 1 describes these features and the functionals extracted using openSMILE. For the speaking rate, we apply the approach from de Jong and Wempe, which automatically detects syllables and estimates speaking rate based on syllables per second [8].

When comparing speakers with different vocal tracts, we need to ensure that the features affected by the vocal tract lie in the same range. This is primarily an issue with gender so we normalize the female pitch mean and maximum by scaling them to lie in the same range as the male values; all other non-pitch features are raw.

Table 1: *Acoustic-Prosodic features and their functionals extracted from the dialogues.*

| Feature | Description | Functionals |
|---|---|---|
| **Pitch** | F0: The fundamental frequency | mean<br>maximum value<br>max value position<br>min value position<br>standard deviation |
| **Intensity** | The normalized intensity | mean<br>maximum value<br>minimum value<br>max value position<br>min value position<br>standard deviation |
| **Voice Quality** | *Local Jitter*: frame-to-frame jitter (pitch period length deviations) | mean<br>maximum value<br>max value position<br>min value position<br>standard deviation |
| | *DDP Jitter*: Differential frame-to-frame jitter (the 'jitter of the jitter') | mean<br>maximum value<br>max value position<br>min value position<br>standard deviation |
| | *Shimmer*: (amplitude deviations between pitch periods) | mean<br>maximum value<br>max value position<br>min value position<br>standard deviation |
| **Speaking Rate** | Measured in estimated syllables per second | N/A |

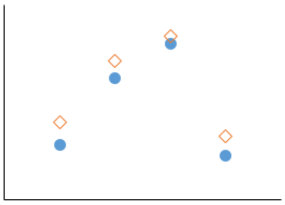
### 3.4 Entrainment Measures

Entrainment can be measured many ways. We investigate three different measures of entrainment, which we call proximity, convergence, and synchrony.

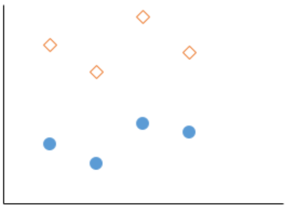
**Proximity** we define as the similarity between two speakers' speaking styles at each turn. A graphical representation

[3] openSMILE configuration file is located at `http://www.public.asu.edu/~nlubold/publications/entrainment_config.html`

of proximity can be seen below where the circles represent one speaker and the diamonds represent their conversational partner. The x-axis is time while the y-axis is the raw feature value. The distance between the two speakers' raw acoustic-prosodic features is indicative of how much the two speakers entrain by proximity.



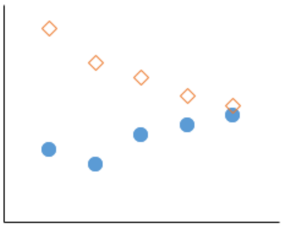
(a) High Proximity



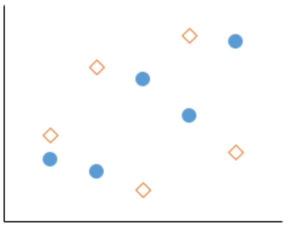
(b) Low Proximity

This measure of entrainment looks at how close the two speakers are to each other at a specific point in time as compared to the rest of the conversation. To determine proximity, we run a paired samples t-test where each pair is composed of two differences. The first difference is the absolute difference between a speaker and their partner at an adjacent turn. The second difference is the absolute difference between a speaker and their partner at ten *other* non-adjacent turns.

**Convergence** is the degree to which speakers become more similar to each other over the course of the entire conversation. Related to proximity, convergence measures the change in similarity between two speakers' speaking styles over time. As seen below, where the circles represent one speaker and the diamonds represent their conversational partner, when convergence exists we will see the difference between the two speakers shrink over time. In the representation below, the x-axis is time while the y-axis is the raw feature value.



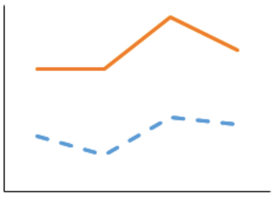
(c) High Convergence



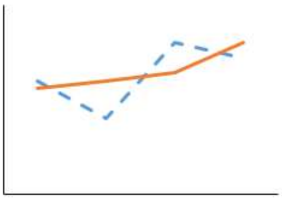
(d) Low Convergence

If convergence does not exist, the two speakers may grow further apart over time (diverge) or there may be no pattern. We find convergence by using Pearson's correlation in a two-tailed t-test between time and the absolute difference between a speaker and their partner at an adjacent turn.

**Synchrony** is the quality of interaction which occurs when speakers stay "in sync" as they converse. With this particular measure of similarity, the speakers may have widely different acoustic-prosodic feature values at each turn, but as they converse, they modulate these values in tandem. This can mean that if one speaker increases how loud they are speaking, the other speaker reacts by also increasing their loudness. In the graphical representation of synchrony below, each line represents the change in a particular speaker's acoustic-prosodic feature value over time.



(e) High Synchrony



(f) Low Synchrony

If two speakers are not in sync, this means there is no pattern in how they modulate their voices. To find synchrony, we compute Pearson's correlation coefficient with a two-tailed t-test on the speakers' feature values at adjacent turns.

Figure 3 visibly depicts these phenomenon as they occur in our corpus for a two-minute sample with particularly high-levels of entrainment. Following Levitan and Hirschberg [16], we run a series of statistical tests to determine significant acoustic features for these three measures. In all three tests, we follow Levitan and Hirschberg in considering results with $p < 0.01$ to be statistically significant and the results with $p < 0.05$ to approach significance.

## 3.5 Rapport Measures

In this paper, we primarily look at rapport from a perceptual perspective. We validate our measure of perceptual rapport by comparing it to self-reported rapport obtained from five of the eight dyads.

To obtain a measure of *perceptual rapport*, three annotators listen to only the audio for each 32 two-minute-long conversational segment. Since there are four segments per dyad, we mix the order of all of the segments to ensure that the annotators listen to each two-minute conversation in a random order. For each segment, the three annotators respond to the following statement using a three-point Likert scale (Agree, Neutral, Disagree):

> "There is a sense of closeness between Student A and Student B"

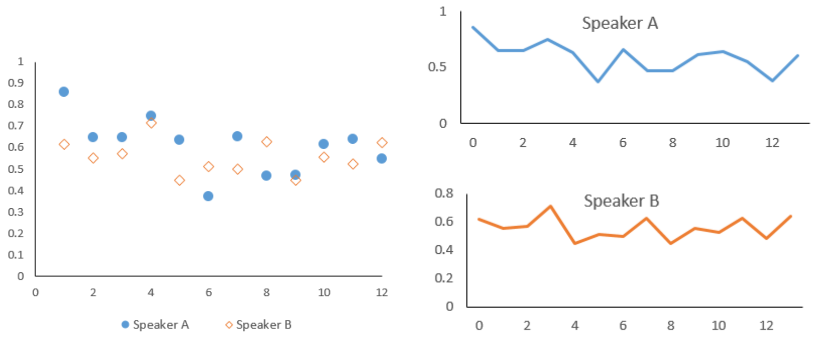
This question was adopted from the rapport scale statements developed by Gratch et al. [13]. We check for inter-rater agreement using percent agreement and Cohen's Kappa; the average pairwise percent agreement across all segments is 63.5% while the average pairwise Cohen's Kappa is 0.41. This is lower than we would like but it is not entirely unexpected [1, 22]. However, given that the level of agreement between the annotators is lower, we validate the perceptual observations against the measures of self-reported rapport we collect from five of the eight dyads.

To obtain a measure of *self-reported rapport*, we pose to each of the participants two questions which have a similar connotation. We do this at the end of the session. The participants respond to the following statements, again using a three-point Likert scale (Agree, Neutral, Disagree):

> "My partner created a sense of closeness between us"

> "I tried to create a sense of closeness between us"

There are two primary differences between the question we pose to the annotators and those we pose to the partic-



(a) Proximity & Convergence (b) Synchrony

Figure 3: *Representations of the entrainment between two speakers from our corpus, highlighting the patterns which are relevant for the three measures of entrainment. In both figures, the x-axis represents time and the y-axis gives the raw feature values for intensity mean.*

ipants. The first difference is the participants' self-reported responses are based upon the entire 30-40 minute session rather than a two-minute segment. The second difference is that while the annotators respond to a single, consolidated statement, *each* participant answers both of the above questions.

### 3.5.1 Validating Perceptual Rapport

We validate the measures of perceptual rapport against the measures of self-reported rapport. We convert the responses we collect for the perceptual and self-reported rapport into numerical quantities we can evaluate by representing each response as either $0, 0.5$, or $1$, for "Disagree","Neutral", "Agree", respectively. To aggregate the perceptual observers' responses for each segment, we find the average of the three observers' ratings giving us a single value between 0 and 1 for each segment. For the self-reported rapport, we also take the average of the responses, this time from both of the participants, to obtain a single value between 0 and 1.

We compare the results from the perceptual annotators to the self-reported rapport scores for the five dyads. As we divided the dialogue of each dyad into four two-minute segments and the annotators provided perceptual observations for each segment, we find the true difference between the perceptual score of each segment and the overall self-reported score for that dyad. The segments are aligned temporally, in the order in which they occurred in the dialogue. Figure 4 depicts the results of this comparison for all five dyads.

Examining Figure 4, we find that for the first two segments the perceptual observers are not very aligned with the views of the participants but when we look at the last two segments, the perceptual rapport scores begin to reflect the self-reported rapport from the participants with increasing accuracy. Based on this observation, we choose to only compare the latter two segments from each dyad when examining the relationship between entrainment and rapport.

## 4. RESULTS

We first investigate whether acoustic-prosodic entrainment exists within collaborative learning dialogues, what type of entrainment appears to be the most common, and which acoustic-prosodic features are the most prominent. We then find whether there is a relationship between acoustic-prosodic entrainment and perceptual rapport, reporting on features of acoustic-prosodic entrainment which correlate significantly with our rapport scores.

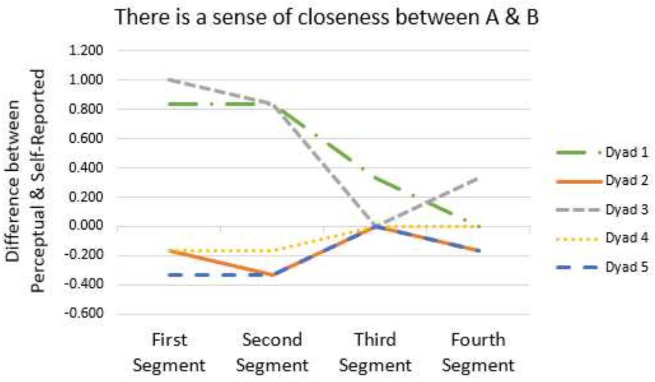


Figure 4: *Difference in perceived vs. self-reported rapport for the five dyads.*

### 4.1 Entrainment in the Corpus

We first examine entrainment across the entire set of dialogues (i.e. all 32 segments), and we find evidence of significant entrainment in the collaborative learning dialogues. Looking at all three measures of entrainment, we see evidence for all three forms; however, participants appear to entrain more in the form of proximity than convergence or synchrony. This means that speakers are matching each other at adjacent turns. We find that synchrony and convergence are present across the corpus but the correlations are smaller.

Looking at specific features for **Proximity** as shown in Table 2, our results are consistent with Levitan and Hirschberg's findings. We find that speaker's are matching each other most significantly in terms of intensity, providing more evidence that the speakers may be changing their normal behavior in intensity in order to conform to that of their partner. This also aligns with the findings from Coulston et al. which found that the majority of children actively accommodated the amplitude of their software partner [7].

In contrast to Levitan and Hirschberg, we find significance in only a subset of the features we examined for synchrony

Table 2: **Finding Entrainment in the Corpus** — *Proximity is measured using a paired t-test while both Convergence and Synchrony are measured using Pearson's correlation. Values shown are significant at $p < 0.05$; values marked with an $^*$ are significant at $p < 0.01$.*

| | **Feature** | **Functional** | **Paired t-test $t$** |
|---|---|---|---|
| Proximity | Intensity | position max | 2.29 |
| | | std dev | $-2.84^*$ |
| | | max | $-2.83^*$ |
| | Pitch - F0 | mean | $-1.98$ |

| | **Feature** | **Functional** | **Pearson's Corr. $r$** |
|---|---|---|---|
| Synchrony | Intensity | mean | $0.12^*$ |
| | | std dev | $0.11^*$ |
| | | max | 0.09 |
| | Pitch - F0 | mean | 0.08 |

| | **Feature** | **Functional** | **Pearson's Corr. $r$** |
|---|---|---|---|
| Convergence | Local Jitter | position max | $-0.09$ |
| | | max | $-0.08$ |

and convergence, where as they found significance in every feature. Speakers exhibit **Synchrony** when they adjust their speech in tandem with that of their partners. In our corpus, speakers are entraining in this manner on intensity; however while the correlations we find are significant, they are also weak, as seen in Table 2. That intensity is the strongest feature for synchrony again makes sense given previous findings. For **Convergence**, we find that only local jitter is significant when examining it at the turn-level across the whole corpus.

While we do find significant acoustic-prosodic features for all three aspects of entrainment, we do not find the same level of entrainment as found by Levitan and Hirschberg. This could be due to several factors. One may be the differences in domain. Niederhoffer and Pennebaker found that entrainment is associated with the degree of engagement [18]. The Columbia Games corpus used by Levitan and Hirschberg makes use of the gaming domain and is more likely to have higher levels of engagement.

### *4.1.1 Entrainment within Each Dyad*

While proximity may be the most significant form of entrainment when looking across the entire set of dialogues, it may not be the most significant form of entrainment for each dyad. So in addition to looking at entrainment across the whole corpus, we also explore entrainment within each individual dyad. As shown in Table 3, we find that not every dyad entrains in all three ways. Synchrony is the most common form of entrainment; every dyad does entrain synchronously and for five out of the eight dyads, this is the most significant form of entrainment. Proximity is a close second; seven out of eight of the dyads also entrain with proximity, matching each other on a turn-by-turn basis. The least common form of entrainment within each dyad is convergence, with only five of the eight dyads showing any signs of convergence, and it is also the least significant form of entrainment.

Table 3: *Evidence of entrainment within each of the eight dyads. A check mark means that there was significant evidence ($p < 0.05$) of entrainment for that acoustic-prosodic feature and entrainment measure for that column's dyad*

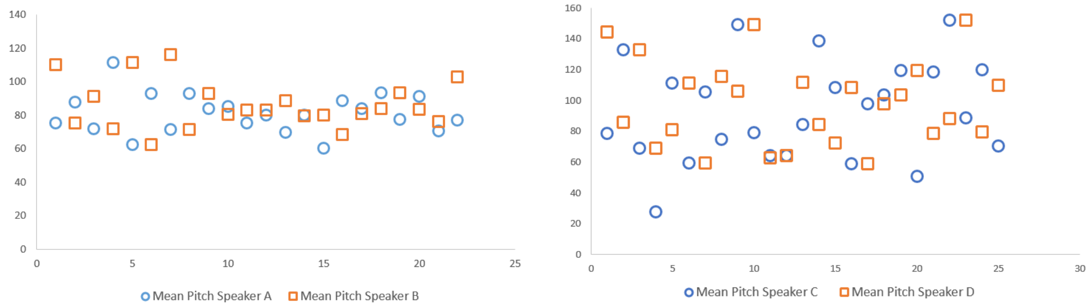
| | **Dyads** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| **Proximity** | | | | | | | | |
| Intensity | ✓ | ✓ | | | | | | ✓ |
| Pitch | | | ✓ | | ✓ | ✓ | | ✓ |
| Voice Quality | ✓ | ✓ | | | ✓ | | ✓ | |
| Speaking Rate | | | | | | ✓ | | |
| **Synchrony** | | | | | | | | |
| Intensity | ✓ | ✓ | ✓ | | | ✓ | | ✓ |
| Pitch | ✓ | | ✓ | ✓ | ✓ | | | |
| Voice Quality | ✓ | | ✓ | ✓ | ✓ | | ✓ | |
| Speaking Rate | | ✓ | | | | | | |
| **Convergence** | | | | | | | | |
| Intensity | | ✓ | | ✓ | | | ✓ | |
| Pitch | ✓ | | | | | | | ✓ |
| Voice Quality | ✓ | ✓ | ✓ | | | | ✓ | ✓ |
| Speaking Rate | | ✓ | | | | | | |

The acoustic-prosodic features which are important for each measure differ depending on the dyad and the measure. Intensity, pitch, and voice quality are distributed across the eight dyads. Speaking rate is entrained on the least. This could be due to the approach we took, which looks at IPUs as the unit of analysis, and the nature of the dialogues, where the IPUs were often shorter in duration.

Looking at the acoustic-prosodic entrainment measures at the dyad level, it is clear that entrainment is still very evident in the collaborative dialogues. It also highlights the complex nature of entrainment and an area for future work. While entrainment is an observable, global phenomenon in collaborative learning dialogues, we now also know that different pairs of people entrain in varying ways as shown in Table 3. There are a number of reasons for why individual pairs might entrain differently; investigating and understanding the elements which are contributing to the entrainment differences among dyads in future work will help to refine measures of entrainment and rapport. In the next section, we pursue the question of whether entrainment at the corpus-level can detect qualities like rapport, but future work will need to incorporate the abundant information which is available at the dyad level, including accounting for the differences in acoustic-prosodic entrainment which appear within and across dyads.

## 4.2 Entrainment and Rapport

To determine whether acoustic-prosodic features of entrainment can be used to detect rapport, we identify whether there is a relationship between entrainment and rapport by comparing the entrainment scores from the whole set of dialogues with the perceptual rapport scores for the latter two segments of each dyad. We consider the latter two segments to reflect a more accurate picture of the perceptual rapport, having validated it against the self-reported rapport we ob-



(a) High Rapport: Score of 0.84

(b) Low Rapport: Score of 0

Figure 5: *Comparison of proximity entrainment on pitch mean for two dyads, one with high rapport and one with low rapport, shows a distinct difference in how similar the two speakers' of each dyad are to each other at each turn.*

tained from five of the eight dyads. We compare entrainment and rapport by finding Pearson's correlation coefficient with a two-tailed t-test.

Table 4: **Correlating Rapport and Entrainment** — *We find the below acoustic-prosodic features for each of the three entrainment measures are the most indicative of rapport. Values shown are significant at $p < 0.05$; values marked with an * are significant at $p < 0.01$.*

| | **Feature** | **Functional** | **Pearson's Corr.** $r$ |
|---|---|---|---|
| Proximity | Pitch - F0 | max | 0.842* |
| | | mean | 0.804 |
| | | std dev | 0.510 |
| | Jitter - DDP | max | 0.644* |
| | | std dev | 0.512 |
| Synchrony | Pitch - F0 | std dev | 0.568 |
| | Jitter - Local | std dev | 0.741* |
| Convergence | Pitch - F0 | std dev | 0.586 |
| | Jitter - Local | std dev | 0.634* |

We find that **Proximity** has the most acoustic-prosodic indicators of rapport. An interesting observation is that despite being the most significant feature of proximal entrainment for the corpus as a whole, intensity does not appear as a feature that may be indicative of rapport. While people entrain more on intensity in general in our corpus, it appears that entraining on pitch may be more pertinent when looking for indicators of rapport in collaborative learning dialogues. Figure 5 illustrates how speakers entrain differently when there is high rapport versus low rapport. It depicts the proximity entrainment for two dyads from our corpus on the acoustic-prosodic feature pitch mean.

**Synchrony**, where speakers change their behavior in sync, is positively correlated to rapport for two acoustic-prosodic features: the standard deviation of the pitch (F0) and the standard deviation of the local frame-to-frame jitter. Looking at **Convergence**, we see these same two features. All of these features are strongly correlated, indicating that for rapport, synchrony and convergence also play a role, and that pitch is once again a pertinent feature.

We find that all three measures of entrainment correlate to rapport, but of the four features of acoustic-prosodic entrainment (i.e. intensity, pitch, voice quality, and speaking rate), we find only two are statistically significant. Both pitch and voice quality (in the form of jitter) are positively correlated with rapport for all three forms of entrainment. Intensity and speaking rate do not appear as significant with any entrainment measure in the presence of rapport.

# 5. DISCUSSION AND FUTURE WORK

The two goals of this paper are to investigate acoustic-prosodic entrainment in collaborative learning dialogues and to discover if acoustic-prosodic entrainment can be used to detect complex qualities of interaction like rapport. We investigate three measures of entrainment, **Proximity**, **Synchrony**, and **Convergence** using four acoustic-prosodic features (intensity, pitch, voice quality, and speaking rate). We find that all three measures of entrainment do exist in collaborative learning dialogues though to a lesser extent when compared to previous works such as on the Columbia Games corpus [16]. We also find that people entrain the most by *intensity*. We then collect perceptual ratings of rapport, validate these against self-reported ratings of rapport, and finally, compare these to our acoustic-prosodic features of entrainment. We find that all three forms of entrainment correlate with rapport. People appear to entrain the most by **Proximity**, matching the acoustic-prosodic features of their speech on a turn-by-turn basis, and this form of entrainment has the most significant relationship to rapport. *Pitch* and *voice quality* appear to be the most significant acoustic-prosodic features people entrain on when rapport is present.

One of the primary motivations behind this work is to identify whether we can detect rapport using entrainment. Automatically detecting rapport in human-to-human and human-to-computer interactions can have real-world implications. In the classroom, automatically detecting rapport can serve as a guide for teachers when students are engaged in collaborative activity. In tutorial dialogue systems, detecting rapport has implications for improving dialogue success and quality. With the knowledge that pitch and voice quality appear to be the most significant acoustic-prosodic features when rapport is present, we can build systems which can support and provide interventions when we detect entrainment or a lack of entrainment on these features. However, while our focus was on detection of rapport, our findings can also have interesting implications for designing the output of intelligent agents by informing the design of agents which build a rapport with their users. Past works on intelligent agents, rapport, and prosody have focused on manipulating specific features such as speaking rate [23] or intensity [7] or by looking at a broad number of features, such as Acosta and Ward's Gracie, which builds rapport with a user by analyzing the user's prosody on 37 features and generating an emotionally colored response [1]. The results here suggest that increased focus on the prosodic features of pitch and voice quality may present an opportunity for further understanding and building an emotional connection with a virtual agent. This is a challenging direction for future work as producing prosodic features which accurately reflect pitch is a difficult problem [27].

A difficulty posed by this work was if, how, and when to consider multimodal data. We ultimately decided to only utilize the speech data in our examination of entrainment and rapport despite collecting multi-modal data in the form of video of the overall interaction, video close-ups of the individuals facial expressions, screen-monitoring of the tablets, and log data from the application. We chose to focus only on the audio because entrainment and rapport in speech has received less attention when it comes to collaborative learning scenarios. In future work, we will consider incorporating the multimodal data in our investigations of entrainment and rapport. Students also completed pre-tests and post-tests, but we are still refining these measures of learning. We leave correlations of learning to entrainment and rapport to future work as well.

## 6. ACKNOWLEDGMENTS

This material is based upon work supported by the Ira A. Fulton Schools of Engineering through the Dean's Fellowship. We would like to thank Kurt VanLehn, Sree Aurovindh Viswanathan, Sujith Gowda and the rest of the FACT team, for their help and support in the data collection process. We would also like to thank Arun Reddy Nelakurthi for guidance in the dialogue setup, and Erin Walker for providing excellent insight and expertise in the writing process.

## 7. REFERENCES

[1] J. C. Acosta and N. G. Ward. Achieving rapport with turn-by-turn, user-responsive emotional coloring. *Speech Communication*, 53(9):1137–1148, 2011.

[2] F. J. Bernieri, J. S. Reznick, and R. Rosenthal. Synchrony, pseudosynchrony, and dissynchrony: Measuring the entrainment process in mother-infant interactions. *Journal of personality and social psychology*, 54(2):243, 1988.

[3] J. Cassell, A. J. Gill, and P. A. Tepper. Coordination in conversation and rapport. In *Proceedings of the workshop on Embodied Language Processing*, pages 41–50. Association for Computational Linguistics, 2007.

[4] C. Chapman, L. Ramondt, and G. Smiley. Strong community, deep learning: Exploring the link. *Innovations in education and teaching international*, 42(3):217–230, 2005.

[5] T. L. Chartrand and J. A. Bargh. The chameleon effect: The perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893, 1999.

[6] M. T. Chi. Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, 1(1):73–105, 2009.

[7] R. Coulston, S. Oviatt, and C. Darves. Amplitude convergence in children's conversational speech with animated personas. In *Proceedings of the 7th International Conference on Spoken Language Processing*, volume 4, pages 2689–2692, 2002.

[8] N. H. de Jong and T. Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390, 2009.

[9] A. L. Drolet and M. W. Morris. Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. *Journal of Experimental Social Psychology*, 36(1):26–50, 2000.

[10] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM, 2010.

[11] H. Friedberg, D. Litman, and S. B. Paletz. Lexical entrainment and success in student engineering groups. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 404–409. IEEE, 2012.

[12] A. L. Gonzales, J. T. Hancock, and J. W. Pennebaker. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 2009.

[13] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy. Creating rapport with virtual agents. In *Intelligent Virtual Agents*, pages 125–138. Springer, 2007.

[14] J. L. Lakin, V. E. Jefferis, C. M. Cheng, and T. L. Chartrand. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of nonverbal behavior*, 27(3):145–162, 2003.

[15] R. Levitan, A. Gravano, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova. Acoustic-prosodic entrainment and social behavior. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–19. Association for Computational Linguistics, 2012.

[16] R. Levitan and J. B. Hirschberg. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proceedings of Interspeech*, 2011.

[17] A. Nenkova, A. Gravano, and J. Hirschberg. High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 169–172. Association for Computational Linguistics, 2008.

[18] K. G. Niederhoffer and J. W. Pennebaker. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360, 2002.

[19] A. Ogan, S. Finkelstein, E. Walker, R. Carlson, and J. Cassell. Rudeness and rapport: Insults and learning gains in peer tutoring. In *Intelligent Tutoring Systems*, pages 11–21. Springer, 2012.

[20] S. Oviatt, A. Cohen, and N. Weibel. Multimodal learning analytics: Description of math data corpus

for icmi grand challenge workshop. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 563–568. ACM, 2013.

[21] D. Reitter and J. D. Moore. Predicting success in dialogue. In *Annual Meeting-Association for Computational Linguistics*, volume 45, page 808, 2007.

[22] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth. Acoustic emotion recognition: A benchmark comparison of performances. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 552–557. IEEE, 2009.

[23] D. Schulman and T. Bickmore. Modeling behavioral manifestations of coordination and rapport over multiple conversations. In *Intelligent Virtual Agents*, pages 132–138. Springer, 2010.

[24] J. Thomason, H. V. Nguyen, and D. Litman. Prosodic entrainment and tutoring dialogue success. In *Artificial Intelligence in Education*, pages 750–753. Springer, 2013.

[25] L. Tickle-Degnen and R. Rosenthal. The nature of rapport and its nonverbal correlates. *Psychological inquiry*, 1(4):285–293, 1990.

[26] D. R. Traum and P. A. Heeman. Utterance units in spoken dialogue. In *Dialogue processing in spoken language systems*, pages 125–140. Springer, 1997.

[27] M. Wagner and D. G. Watson. Experimental and theoretical advances in prosody: A review. *Language and cognitive processes*, 25(7-9):905–945, 2010.

[28] A. Ward and D. J. Litman. Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. In *SLaTE*, pages 57–60, 2007.