An Exploratory Approach to Measuring Collaborative Engagement in Child Robot Interaction

Yanghee Kim[*]
ykim9@niu.edu
CREATE Center[†]
Northern Illinois University
DeKalb, IL

Sachit Butail
sbutail@niu.edu
Mechanical Engineering
Northern Illinois University
DeKalb, IL

Michael Tscholl
mtscholl@niu.edu
CREATE Center
Northern Illinois University
DeKalb, IL

Lichuan Liu
liu@niu.edu
Electrical Engineering
Northern Illinois University
DeKalb, IL

Yunlong Wang
ywang18@niu.edu
Electrical Engineering
Northern Illinois University
DeKalb, IL

Abstract

This study explored data analytic approaches to assessing young children's engagement in robot-mediated collaborative interaction. To develop our analytic models, we took a case-study approach and looked closely into four children's behaviors during three conversational sessions. Grounded in engagement theory, three sources of multimodal behavioral data (utterances, kinesics, and vocie) were coded through human annotation and automatic speech recognition and analysis. Then, information-theoretic methods were used to uncover nonlinear dependencies (called mutual information) among the multimodal behaviors of each child. From this, we derived a model to compute a compound variable of engagement. This computation produced engagement trends of each child, the engagement relationship between two children in a pair, and the engagement relationship with the robot over time. The computed trends corresponded well with the data from human observations. This approach has implications for quantifying engagement from rich and natural multimodal behaviors.

[*]Corresponding author

[†]Cross-disciplinary Research on Engaging Advanced Technology for Education http://CreateCenter.net.

LAK '20, March 23-27, 2020, Frankfurt, D

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

CCS Concepts

• **Applied computing** → *Education*; • **Social and professional topics** → *Children*; • **Mathematics of computing** → *Information theory*; • **Theory of computation** → Models of computation; Theory and algorithms for application domains.

Keywords

Learning analytics, multimodal data analytics, child robot interaction, engagement, collaborative problem solving, automatic speech recognition, mutual information, information theory, social robotics, human computer interaction

ACM Reference Format:
Yanghee Kim, Sachit Butail, Michael Tscholl, Lichuan Liu, and Yunlong Wang. 2020. An Exploratory Approach to Measuring Collaborative Engagement in Child Robot Interaction. In *Proceedings of The 10th International Learning Analytics & Knowledge Conference (LAK '20).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In recent years, embodied humanoid robots are increasingly experimented in various walks of life from health through retail business to education. This new generation of service robots are distinguished from more conventional industrial and manufacturing robotic systems. The robots are designed as partners or companions of humans sharing everyday life space [1] and perceived by their human partners as lifelike having social and relational capacities [2].

In education, social robots are designed as assistive tools to interact and collaborate with students and teachers. While working with the robots, these users seem to develop interpersonal and affective relationships with their robots [3]. Research and development in the educational use of social robots have been prolific for a range of learner populations

from young children to older people in the last decade. A recent review has identified language learning, social skill development, and robotics education as most popular domains of educational robotic applications [4]. Initial evidence of such applications has been quite promising especially in learners' engagement in robot-assisted tasks.

This study examined if a social robot could help develop collaboration skills of young children who just started schooling. Being able to collaborate with peers are deemed as foundational for both social and intellectual development of children, leading to short and long-term academic success. We instantiated an interaction triad of a robot and two children, where the children engaged in conversations and co-creation of digital artifacts on a shared tablet to solve problems (Figure 1). In this process, the robot prompted the children to the task, soliciting collaborative behaviors, such as turn-taking, shared decision-making, negotiating, and reaching agreement.

Figure 1: A Robot-Mediated Collaboration Triad

A significant challenge we faced was the valid and reliable evaluation of the interaction triad in terms of children's engagement in collaborative interaction. Young children's interactions with each other and with the robot were richly multimodal in nature. In line with some previous work in the LAK community (e.g., [5], [6]), the authors found that conventional psychological measures seemed very limited in capturing such rich behaviors of children.

Alternatively, we explored a multimodal data analytics (MDA) approach where we used both manual and computational coding and processing to assess children's engagement in triadic collaboration. Acknowledging MDA is in its infant stage, we had four preliminary research questions to guide our analytic processes: i) *In what way does a child's engagement in the collaborative task progress over time?* ii) *To what degree do three types of multimodal data (utterances, kinetic, and vocal cues) conform with each other for the child?* iii) *To what degree does the robot's mediation relate to the child's engagement?* iv) *To what degree does the engagement relationship of two children in a pair evolve over time?*

2 BACKGROUND

2.1 Learner Engagement

The study of engagement is increasingly becoming a focal issue in education research because of its observed correlations with learning outcomes and potential predictability of the outcomes [7]. Broadly defined, engagement refers to a student's participation in the learning process, and is considered an expression of internal states, such as commitment, motivation, or interest [8]. It is a multi-faceted phenomenon that is distinguished into cognitive, behavioral, and emotional engagement, each of which is identified through specific behavioral indicators [9]. Cognitive engagement, for example, is identified through students' ongoing effort on task and task performance, behavioral engagement through bodily actions, and emotional engagement through students' positive and negative emotional states during a lesson or class [10].

Nonetheless, these aspects of engagement are often studied in isolation, and holistic accounts of engagement are rare. In fact, learners' internal states can be expressed in several different behavioral forms (e.g., talk, emotion, action). Likewise, learners' engagement in collaboration can be identified through their talks, emotions, posture, and gestures. Each of these can be an indicator of collaborative engagement.

Another shortcoming of the research on engagement to date relates to capturing variations of engagement over a period of interest, such as a lesson or task. Engagement can vary highly within even short sessions as learners react quickly to changes in their environment (e.g., changes in materials or in social relations). This seems to be the case particularly for young children who are receptive even to minute stimuli. In our study, we also have noticed that children's collaborative behaviors developed gradually over time. A classical pre/posttest approach in a controlled setting would not provide an authentic assessment of collaborative engagement. Rather, it must be assessed as development in progress [11].

Further, young children's language and literacy competencies are still developing with substantial individual differences in their development. Frequently used methods, such as self-report surveys and interviews, are less likely to provide valid and reliable measures of engagement.

There is a great need to study engagement through non-intrusive methods as it is demonstrated in various behaviors and as it changes over time. This is especially important for technology-rich environments which present learners new learning material and artifacts to which learners react quickly [12].

In this study, grounded in the literature on engagement [10], we collected three sources of multimodal behavior of children while they participated in the robotic triad. First we recorded the utterances (i.e., linguistic alignment) of the children and the robot as an indicator of cognitive engagement. Next we captured kinetic behaviors (i.e., gaze, posture, and movement) as an indicator of behavioral engagement. Lastly, we recorded children's voices (vocal acoustics) as an indicator of emotional engagement. We analyzed these data sets using both manual annotations and computational techniques, such as speech recognition and signal processing and information theoretic measures.

2.2 Acoustic Signal Processing and Analysis

Over the decades, researchers in psychology and computer science examined vocal behavior (e.g., acoustics) to assess emotional states of interlocutors in social interaction [13, 14]. In this study, we used this emotional vocalization of children as a marker for emotional engagement. Automatic speech recognition (ASR) technology has been widely used in many areas in recent years, such as industry, communication, consumer electronic products, and in medical care [15]. Speech recognition is a procedure of signal processing, changing speech signal waveform in the spatial domain into a series of coefficients which can be recognized or understood by machine. There are four primary auditory features associated with sound: Intensity (loudness), pitch, timbre, and the source of the sound [16]. Intensity (loudness) is a quantitative measure of the amplitude of the sound compared to a reference level. Pitch is a quantitative measure of the actual fundamental frequency of a sound. Timbre is a qualitative measure of a sound that can differentiate between two sounds of equal loudness and pitch through the tonal quality of the sound. The source of the sound has some effect on the perception of the three other features.

Essentially whenever a sound is heard, our brain will actively consider those four analog auditory qualities and make a decision regarding it. We use similar techniques in ASR to analyze children's audio signal in three steps.

(1) Speech signal detection: Short-time energy is used to detect speech segment and silent segments, which will be used for pattern extraction later. Signal detection processes instances of voiced activity without wasting computational time during silent periods. To accurately detect voiced activity, short-term energy, short-term magnitude and short-term zero-crossing are the general methods one can use [15].
(2) Feature extraction: Commonly used features in the ASR community (e.g., LPC, MFCC) are not designed for measuring emotional states. In the current study, we used the intensity and pitch as features. In our previous work for infant cry signal classification, we have found that the abnormal cry occurring when the infant was in great pain had much higher intensity and pitch compared to the ordinary cries occurring when the infant was hungry or its diaper needed to be changed [17, 18].
(3) Feature analysis: We used short-time energy for each segment for intensity, so we can track the time-varying characteristics of the speech signal and compare them with the engagement levels. For pitch, we analyzed the power of the signal in terms of each frequency component, so we can connect the power spectrum density of each segment with the speaker's engagement.

2.3 Information Theoretic Measures of Mutual Information

Methods in time-series analysis that capture agreement (synchrony) between two time series differ in the assumptions about the expected mathematical relationship. Pearson correlation, for example, assumes a linear relationship. It also assumes homoscedasticity, meaning that the data from multiple sources vary within the same range [19]. Such methods are also ill-equipped to study temporal correlations of a data set that draws on measurements obtained using varying sampling rates.

Nonlinear correlation methods overcome these limitations by capturing the statistical similarities between pairs of time series [20]. Among these, the information theoretic measure of mutual information captures the dependence between two time series in terms of the information they share, without assuming linear relationship between the time series. Mutual information has been used in a wide variety of fields ranging from the cognitive sciences to highlight the role of executive functions [21], to the neurosciences to measure connections between behavior and stimuli [22], to weather forecasting for making weather predictions [23].

In this study, we represent a behavior (e.g. voice, kinesics, and linguistic alignment) as a random variable X. The amount of information, or uncertainty, contained within the random variable is defined as entropy $H(X) = -\sum_i p_i \log p_i$ where p_i denotes the probability of the variable taking a discrete value i among all the possible values that X can take [24]. Mutual information, captures the amount of information shared by two random variables X and Y, and is defined as $I(X, Y) = H(X) + H(Y) - H(X, Y)$, where $H(X, Y)$ is the joint entropy between X and Y [24]. A high mutual information implies more agreement (dependence) between two variables, whereas a mutual information of 0 implies complete independence. Mutual information is symmetric so that $I(X, Y) = I(Y, X)$.

Estimating mutual information depends on the reliability of the estimate of the entropy of each process. Since entropy is a probabilistic representation of a process, sufficient data

are required to guarantee reliability of an estimate [25]. If not enough data are available, permutation entropy provides an alternate representation of the data [26]. In this representation, a time series is mapped to a unique sequence of symbols of length m in terms of the order in which the data appear. For example, with $m = 2$, a time series of 0s and 1s can be rewritten in terms of $(0, 0), (0, 1), (1, 0), (1, 1)$. If we denote each of these pairs of values with a new sequence, so that $(0, 0) = 1, (0, 1) = 2$, and so on, $X(t) = \{1, 0, 0, 1, 1, 0\}$ can be rewritten as $\hat{X}(t) = \{3, 1, 2, 4, 3\}$. This form allows us to write a binary time series with higher resolution while capturing transitions in behavior. The value of symbol length m is a design parameter that must be selected to satisfy robustness in results. We use mutual information to first quantify the temporal agreement between each of the three multimodal datasets, and then to capture how interaction with the robot and between children varies throughout multiple sessions.

3 METHODS

3.1 Robot-Mediated Collaboration Context

For the collaborative triad, we designed the robot Skusie as a new friend from another planet who liked to learn about life on earth. It asked children to work together to help him learn about animals, birthdays, school, and family. Each of these topics was covered in a series of two sessions (conversational and tablet-based digital making as in Figure 2), each session taking fifteen to twenty minutes. These triadic interaction sessions were implemented naturally in a school library during the regular school hours. Children participated in a total of six sessions (three conversation sessions and three tablet sessions) on a daily basis over two weeks excluding school events days. For Skusie's behavior, we adopted a wizard-of-oz method, where a researcher controlled Skusie remotely hidden behind the scene.

The sessions were recorded in real time using two HD video cameras located on the right and left sides in front of the children. Two directional microphones (AT897) were located on each side of the children to capture a child's speech separately. Also, since the interaction sessions were run in a natural setting, there were constant noises surrounding the setting while children talked to each other and with the robot. We placed two ambient microphones (Shure SM94) at two corners of the walls (one in front and the other behind the children) to capture the background noises. Lastly, two researchers took descriptive observation notes.

3.2 Participants and Data Selection

A total of eighty English-speaking children (aged five to six) participated in the robotic sessions. We paired two children to form a robotic triad in consideration of mixed gender and mixed ethnicity, also avoiding extant close friendships.

Figure 2: Sample Sessions for Conversational (a) and Tablet-Based Digital Making (b)

For this phase of model development, we randomly selected two triads based on the quality of audio data. Per triad, we had a total of 100 to 120-minute-long audio and video data. Vocal cues were processed and analyzed at a 3 second interval by ASR; matching video data was annotated manually to code for kinesics and linguistic alignment (see Section 3.4 for more information).

3.3 Automatic Extraction of Acoustic Pitch and Intensity of Children

We extracted acoustic pitch and intensity of children from audio streams using short time processing technique. In order to obtain the intensity information, we calculated the short-time energy (STE) defined as the average of the square of the sample values in a suitable window. It can be mathematically described as follows [27].

$$E(n) = \frac{1}{N} \sum_{m=0}^{N-1} [W(m)x(n-m)^2] \tag{1}$$

where w(m) are the coefficients of a suitable window function of length N. The Hamming window has been chosen as it minimizes the maximum side lobe in the frequency domain. The intensity of a speech signal is related to the speaker's condition, the microphone and its placement, the pre-amplifier and the recorder as well. We calculate the intensity level based on a threshold which can avoid the effects from the hardware setup and placement. Then the intensity level is defined by comparing it with the threshold t.

$$\text{intensity} = \begin{cases} 1 \text{ (high)} & \text{if } E(n) > t, \\ 0 \text{ (low)} & \text{if } E(n) < t. \end{cases} \tag{2}$$

where the threshold t can be obtained by evaluating all the speech segments in the audio data file, meaning that we select all speech segments from the recorded file, ignoring the silent ones, and then calculate the average energy of all the samples in these speech segments. $t = \frac{1}{L} \sum_{i=1}^{L} [x(i)^2]$ with $x_i(n)$ is the nth samples in ith speech segments. Figures 3 and 4 show the waveforms and short time energy of different speech segments. Figure 3 shows one segment with high intensity and Figure 4 shows a low intensity segment.

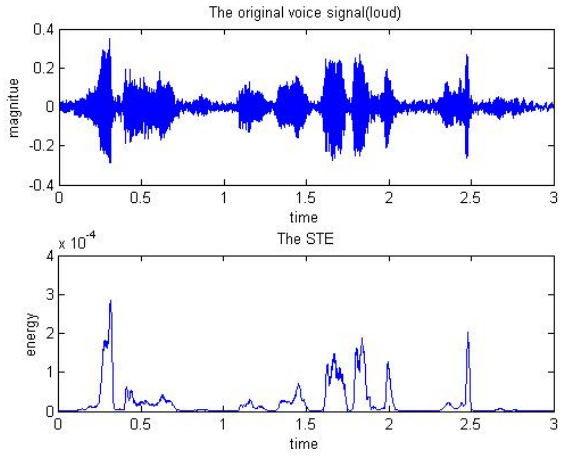

Figure 3: The Segment with High Intensity

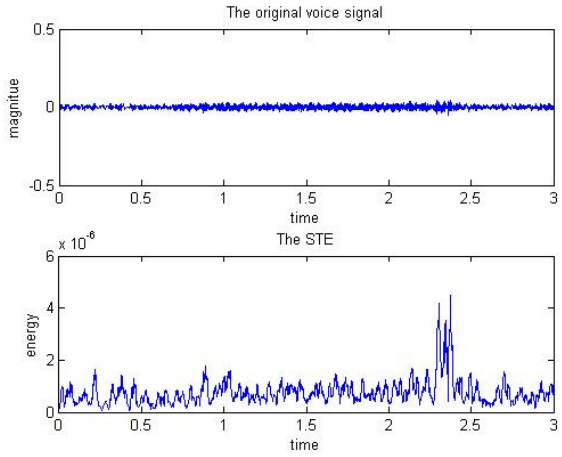

Figure 4: The Segment with Low Intensity

For pitch, we used time frequency analysis to analyze the kindergartners' speech signals. It is well known that Discrete Fourier Transform (DFT) of a long sequence is an estimate of the power spectrum density (PSD), also called periodogram [16]. Different speech signals from different children would produce similar gross PSD. Therefore, we use short-time Fourier Transform (STFT) to obtain the time-varying properties of speech signals. STFT is defined as

$$X_n(e^{jw}) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-jwn} \quad (3)$$

Fourier transform (FT) of the sequence of input signal is convolved with the FT of the shifted window. To represent $X(e^{jw})$ by using STFT $X_n(e^{jw})$, we choose a window function with spectral highly concentrate around origin. In this paper, the Hamming window was used to conduct STFT. Once $X_n(e^{jw})$ was obtained, we compared it with pitch threshold. If the majority of power was in the low frequency band (<440 Hz), we defined it as low pitch (coded as 0), otherwise, we defined it as high pitch (coded as 1). The threshold value of 440 Hz was set referring to literature [28] and also consulting disciplinary experts.

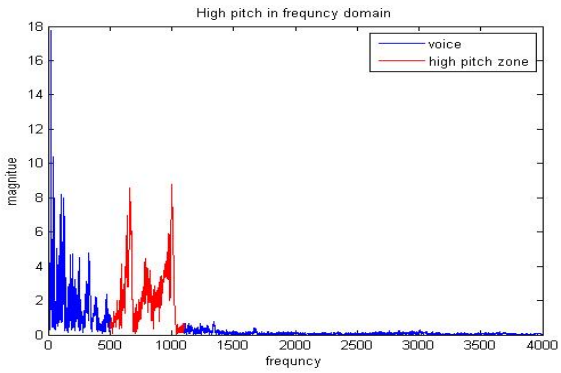

Figure 5: The Segment with High Pitch Component

In Figure 5, we can find that the speech segment has frequency components from 100Hz to 1000Hz, the high pitch parts are shown in red color. Figure 6 shows a segment has majority power located in the lower frequency band.

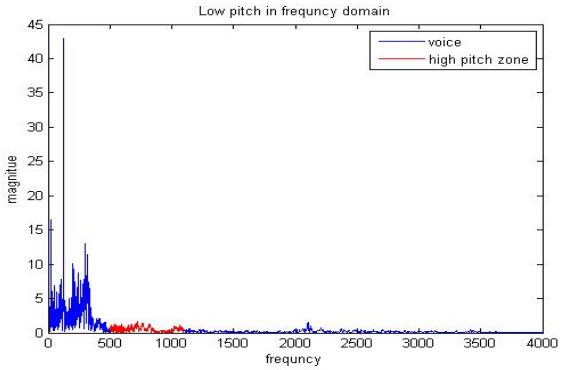

Figure 6: The Segment with Low Pitch Component

3.4 Annotation of Kinesics and Utterances

To assess learning behaviors, observational approaches are increasingly being used in research, overcoming some shortcomings of traditional methods such as self-report surveys or interviews [29]. In the observational approach, trained observers annotate behavior using predefined categories and their behavioral indications (BIs). The presence or absence of a BI within prespecified time segments is systematically recorded. Thus, changes in engagement can be traced over time and subsequently linked to other BIs or events occurring concurrently in that time segment (e.g. learning material being introduced or a question being asked). For our study, we developed three coding categories (Kinesics, Linguistic

alignment, and Robot talk) to capture behavioral indicators of collaborative engagement. Kinesics captured behavioral engagement evident in children's bodily movements and posture. Linguistic alignment captured whether a child's speech was a response to the speech by either the robot or the peer. For these categories, we developed the following behavioral indicators:

(1) *Kinesics*: eye contact, gaze orientation (looking at the robot), body orientation (facing peer or robot), posture (e.g. leaning forward), gestures or enactments of ideas (e.g. representing a concept), and facial expressions (e.g. smile).
(2) *Alignment*: responding to the robot's question, extending or elaborating talks by the peer or the robot, initiating a talk related to the current topic. Self-talk and talk or mumbling unrelated to prior robot/peer utterances were not coded as linguistically aligned.
(3) *Robot talk*: whether the robot makes utterances to mediate children's conversation.

Video recordings were segmented into 3-second periods. These were synchronized with the segments of vocal signals, aligning the times of occurrence of that event in the video and audio files.aligning the times of occurrence of that event in the video and audio files. We coded video recordings of pairs of children using the commercial software INTERACT [1]. Coders viewed a 3-second-long segment and recorded whether a behavioral indicator was present in the segment (1) or not (0). Initially, two student researchers and one senior researcher coded 5-minute segments of randomly selected videos independently using the behavioral indicators; inter-rater reliability (using Cohen's kappa) was computed. We repeated this process until we obtained kappa values of above .85, which are considered excellent agreements [30]. During this coding phase, we found more behavioral indicators which were added to the category descriptions in a code book. The student researchers then coded the remaining video data; any coding discrepancies between these coders were resolved in a series of consolidation meetings that included the student coders and two senior researchers.

3.5 Calculating Mutual Information among Data Sources

Four types of data (intensity, pitch, kinesics, and alignment) for each child were preprocessed as follows. First, because manual annotation is more reliable to identify a child, both intensity and pitch data were reassigned to each child using on the annotated data. Next, intensity and pitch for each child was set to 0 (silence) when the other child was speaking. Finally, for instances where both children were speaking (average 25% of the time for two datasets), we randomly re-assigned voice data to one or the other child.

To determine the extent to which each of these data sources agreed with each other for each child, we calculated mutual information between each pair of sources. Specifically, mutual information $I(X(t), Y(t))$ was computed for each pair of data sources (e.g. intensity × alignment) over a moving window of size w of the time series with a symbolic representation of length m. Computing mutual information over a moving window captures temporal agreement between the time series.

The size of the moving window wand the symbol length m were determined by computing mutual information over a range of values $w = \{30, 60, 120, 180, 240, 300\}$ seconds and $m = \{2, 3, 4\}$ and calculating the coefficient of variation (standard deviation/mean) [31] for each child in a sample dataset. We obtained the smallest coefficient of variation with a value of $m = 3, w = 30$ seconds, and this was selected for the subsequent analyses. Finally, because mutual information as calculated between two temporal segments was a relative quantity, we normalized it with the maximum mutual information. The maximum mutual information for a temporal segment is simply the mutual information between the temporal segment with itself. Therefore, for a given window mutual information varied between 0 and 1, with 0 denoting no dependence, and 1 denoting identical time series.

3.6 Synthesizing Engagement

Grounded in the theory of engagement, we observed engagement in terms of the different modalities of linguistic alignment as an indicator of cognitive engagement, kinesics as an indicator of bodily engagement, and vocal cues as an indicator of emotional engagement. Referring to the mutual information values, we proposed the following formula to compute engagement. The high mutual information between Intensity and Pitch (i.e., high correlation I >.75) led us to use an average of the two. Kinesics and linguistic alignment showed little correlation so they were considered independent. Engagement therefore is defined as the sum of bodily engagement (Kinesics), cognitive engagement (Alignment), and emotional engagement (the average of vocal Intensity and Pitch). Denoting these values at a time t, by K(t), A(t), I(t),and P(t), the engagement, E(t), is calculated as

$$E(t) = K(t) + A(t) + \frac{I(t) + P(t)}{2} \tag{4}$$

We used this formula to compute engagement for each child by sessions to answer research question 1, and used the resulting engagement values to answer research questions 2, 3, and 4.

[1] http:\\www.mangold.de

4 RESULTS

4.1 R 1 on Child Engagement

Research question 1 asked about the development of each child's engagement as the sessions progress. Figure 7 shows the regression lines for each child for all three sessions. Regression was fitted to frequency data (blue dots in graph) of the engagement value computed with the formula for 1-minute intervals. Figure 7 shows clearly that engagement has changed by children (A, B, C, D).

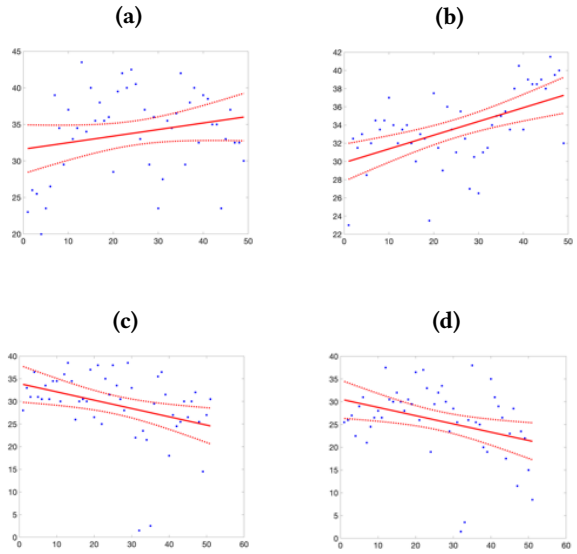

Figure 7: Regression Lines for Children A, B, C, and D for Three Sessions (Y-axis: Engagement, X-axis: Time, and Dotted Curve: Confidence Bounds)

Table 1 reports slope and fit of the Engagement regression lines by triadic interaction sessions (S1, S2 and S3) and for all sessions (S1-3). Fit varies for a child across sessions, clearly presenting dynamic trends in children's engagement session by session over time.

Table 1: Slope (β) and Fit (R^2) of Engagement Regression Lines by Session and Child

Child	S1		S2		S3		S1-3	
	β	R^2	β	R^2	β	R^2	β	R^2
A	1	0.6	0	0	0	0	0	0
B	0.2	0.3	0	0	0.4	0.4	0.2	0.3
C	0	0	-0.3	0	0.5	0.6	-0.2	0.1
D	0.3	0.2	-0.4	0	0	0	-0.2	0.1

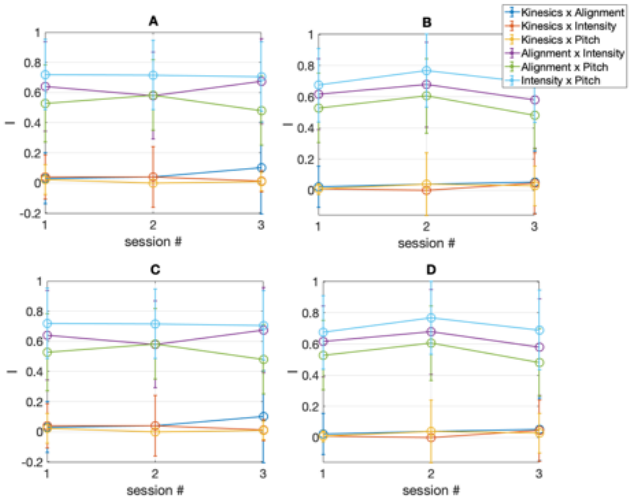

Figure 8: Mean and Standard Deviation of Normalized Mutual Information Between Data Sources for Each Child over Multiple Sessions

4.2 RQ 2 on Relationship Among Multiple Data Sources

Research question 2 asked, To what degree do three types of multimodal data (alignment, kinesics, and vocal cues) conform with each other for the child? Figure 10 presents the interactions among the data sources per child per session. We noted that some datasets such as voice Intensity and Pitch conformed highly with each other (I > .75 for all sessions) than others such as Kinesics and Alignment (I < 0.1, for all sessions). We also noted that Alignment conformed moderately with both Pitch and Intensity (I > 0.6, for almost all sessions), meaning that most of children's talk was collaborative, i.e. they responded to the robot or a peer, or elaborated on what had been said by the other. We did not find large variation between these values across sessions.

4.3 RQ 3 on Relationship Between Child Engagement and Robot Mediation

Research question 3 asked, To what degree is the robot's mediation related to the child's engagement? Frequencies of Engagement values and robot talk per 1-minute interval are shown for the children A and B in Figure 9, and for the children C and D in Figure 10. Trends in a child's Engagement and the robot's mediating talk over time moved along with each other in the same direction, showing that their interactions were reciprocal.

Table 2 presents the means and standard deviations of normalized information-theoretic mutual information between a child and the robot for each child in each session and all three sessions. Correlations between child engagement and robot talk ranged between 0.57 and 0.76, with little variation

between sessions. There were no statistical differences in this range.

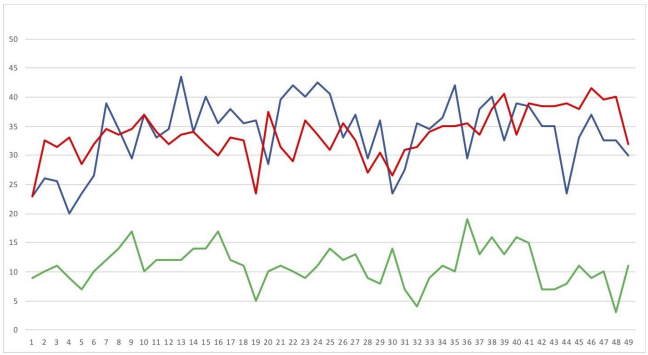

Figure 9: Engagement of Child A (Blue) and Child B (Red), and Robot Talk (Green)

Figure 10: Engagement of Child C (Blue) and Child D (Red), and Robot Talk (Green)

Table 2: Means and SDs of Mutual Information between Child Engagement and Robot Talk

Child	Session			All
	1	2	3	
A	0.7(0.23)	0.76(0.1)	0.62(0.3)	0.68(0.24)
B	0.68(0.22)	0.74(0.2)	0.57(0.26)	0.66(0.24)
C	0.67(0.22)	0.72(0.21)	0.63(0.28)	0.67(0.24)
D	0.63(0.25)	0.66(0.17)	0.57(0.3)	0.62(0.26)

4.4 RQ 4 on Engagement Relationship Between Two Children Within a Pair

Research question 4 asked, To what degree does the engagement relationship of two children within a pair evolve over time? Figures 9 and 10 present trends of the engagement relationship of two children within a group over time. Table 3 shows the means and standard deviations of normalized mutual information between two children within a group. Comparing engagement relationship by group, we found that the group A & B showed a higher mutual engagement relationship than the group C & D.

Table 3: Means and SDs of Mutual Information of Two Children Within a Group

Child	Session 1	Session 2	Session 3	All
A & B	0.67(0.12)	0.6(0.2)	0.62(0.2)	0.63(0.2)
C & D	0.54(0.2)	0.51(0.2)	0.48(0.28)	0.51(0.23)

5 DISCUSSION

5.1 The Compound Variable Engagement

In this study, we developed a preliminary model to compute a compound variable of engagement as a function of three multimodal indicators (utterances, kinesics, and vocal cues) which represent cognitive, behavioral, and emotional engagement. The resulting engagement values for each child in a time series, enabled us to calculate the slope of fitted regression lines, which provided quantitative evidence for the child's engagement in the robot-mediated collaborative activities. The computed engagement variable also allowed us to show overall trends of not only individual children, but also groups, which enabled us to track the progression of collaborative engagement over time. Applying our formula to the data from four children revealed marked differences in the progression of a child's engagement overall.

Importantly, the resulting dynamic trends of collaborative engagement were confirmed by qualitative observations. The progression in collaborative engagement of child C and child D (Figure 7), for example, showed that their engagement was very high at the beginning compared to children A and B. This means that children C and D started with high levels of collaboration, and then their engagement settled at a level similar to the other group. This can be explained by a ceiling effect initially. From on-site observations, in fact, we noticed that the interest of children C and D moved from the task at hand to the robot itself. For example, as the sessions progressed, children started asking the robot personal off-task questions such as "are you a boy or a girl?" Their speech digressed from the prescripted topics, reducing the linguistic alignment value. In parallel to this digression, our qualitative observations noted an increase in their competition with each other for the robot's attention, with a consequent reduction in their collaboration. Utterances like "The robot was talking to me!" increased in frequency. In sum, the quantitative engagement values closely mirrored our on-site observations.

The flexibility afforded by mutual information allowed comparisons between behaviors captured by different data sources within and between individuals. The choice of symbol length and window size was determined based on the robustness of the measure on complete sessions. However, it is also possible that mutual information may evolve during a session and therefore an alternative measure of robustness, one for example that agrees with human coding, should be developed. We interpreted values of mutual information by normalizing it with maximum expected values. An alternative strategy that can be explored in the future would involve statistical comparisons with mutual information between data sources which are known to be independent [32]. We would, for example, expect that the mutual information between engagements of children from different sessions will be significantly less than mutual information between engagements of children in the same session.

5.2 Issues and Recommendations

Overall, the current work presents a volume of issues to be pursued in subsequent research. First, our subsequent work will refine the calculation of the compound Engagement. The current formulation of the compound variable captures that children express their engagement in one or another modality (e.g. with their bodies or their talk), which is consistent with what is known in the literature. However, we also expect the correlations among the observed modalities, which the current formula does not take into consideration. It is likely that these modalities exist in complex relationships that are not yet explained. The on-going analysis of our full datasets will improve the formula by identifying such relationships, which will be integrated into the refined calculation of the compound variable.

Related, the compound formula is currently based on the assumption that engagement is a linear function of all the modalities. We suspect that the relationship of the modalities could be nonlinear or linear with time-varying weights. For example, our qualitative observations revealed that children tended to demonstrate their engagement through body movements and posture more than speech. There were many instances where children would orient themselves towards the robot in an effort to listen to its questions (even when its voice was sufficiently loud). Or they would respond with gaze and leaning towards the robot before verbally answering a question and sometimes not speaking because they were shy. Then, kinesics may be assigned a higher weight; this weight may itself vary when the expression of engagement shifts to another modality (e.g. a child increasingly talk more after a while). Identifying the precise form of the function relating the modalities to the compound variable Engagement requires high-volume behavioral data.

Once the function is determined, the external validation of Engagement will be warranted, e.g., the qualitative equivalents of compound engagement. In future work, the on-site qualitative observations should be coded in a more detailed and systematic manner, providing an independent measure of engagement. Their coding categories could match the components in the computation of compound engagement.

In the ASR community, researchers generally focus on time and frequency features to recognize the speech signal since they are directly related with the content of speech. Little attention has been paid to timing, duration, pitch, and intensity. These features however are more likely to be related to speaker's engagement level. Also, vocal features may vary greatly by individuals, but in our formulation, children's vocal intensity and pitch did not take this individual difference into account. In the on-going analysis, we will use the baseline vocal information of each child to calculate the intensity and pitch of the child.

Lastly, continued developments in technology and analytical techniques could take over some of the current analytic work on kinesics by human annotators. Recently developed applications can detect and precisely describe body movements and posture, including orientation n [33].

To conclude, collaborative engagement in learning, and more broadly in any life experience, is increasingly being recognized in social scientific literature. Being able to engage with others in professional and social life may lead to successful functioning of individuals and their sense of well-being. Yet, there is a lack of understanding of robust ways to examine this phenomenon due to its complex and multimodal nature. The authors admit that the current work is preliminary and presents many issues to be resolved in subsequent research. Nonetheless, the work has potential for broader impact beyond children's engagement, extending to collaboration in the workplace and other social and professional arenas.

ACKNOWLEDGEMENTS

This research was funded by the U.S. National Science Foundation (NSF-IIS 1839194).

References

[1] J. Martínez-Miranda, H. Perez-Espinosa, I. Espinosa-Curiel, H. Avila-George, and J. Rodríguez-Jacobo. 2018. Age-based differences in preferences and affective reactions towards a robot's personality during interaction. *Computers in Human Behavior.* 84, 245-257.

[2] C. Breazeal. 2003. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies,* 59(1-2), 119-155.

[3] N. Picarra, and J. C. Giger. 2018. Predicting intention to work with social robots at anticipation stage: Assessing the role of behavioral desire and anticipated emotions. *Computers in Human Behavior,* 86, 129-146.

[4] Y. W. Cheng, P. C. Sun, and N. S. Chen. 2018. The essential applications of educational robot: Requirement analysis from the perspectives of experts, researchers and instructors. *Computers Education,* 126, 399-416.

[5] P. Blikstein. 2011. Using learning analytics to assess students' behavior in open-ended programming tasks. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge LAK '11*, Banff, Alberta, Canada, Feb 27, 2011.

[6] M. Worsley and P. Blikstein. 2015. Using learning analytics to study cognitive disequilibrium in a complex learning environment. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge LAK '15*, Poughkeepsie, NY, USA, March 16-20, 2015.

[7] M. R. Reyes, M. A. Bracket, S. E. Rivers, M. White and P. Salovey. 2012. Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology,* 104(3), 700-712.

[8] E. A. Skinner and J. R. Pitzer. 2012. Developmental dynamics of student engagement, coping, and everyday resilience. In S. L. Christenson, A. L. Reschly, C. Wylie (Eds.), *Handbook of Research on Student Engagement* (21-44). Boston, MA: Springer, US.

[9] A. L. Reschly and S. L. Christenson. 2012. Jingle, jangle, and conceptual haziness: evolution and future directions of the engagement construct. In S. L. Christenson, A. L. Reschly and C. Wylie (Eds.), *Handbook of Research on Student Engagement* (3-20). Boston, MA: Springer, US.

[10] J. A. Fredricks and W. McColskey. 2012. The measurement of student engagement: a comparative analysis of various methods and student self-report instruments. In S. L. Christenson, A. L. Reschly, C. Wylie (Eds.), *Handbook of Research on Student Engagement* (763-782). Boston, MA: Springer, US.

[11] S. Grover, M. Bienkowski, A. Tamrakar, B. Siddiquie, D. Salter and A. Divakaran. 2016. Multimodal analytics to study collaborative problem solving in pair programming. *Proceedings of the 6th International Conference on Learning Analytics and Knowledge LAK '16*, Edinburgh, United Kingdom, April 25-29, 2016.

[12] C. R. Henrie, L. R. Halverson and C. R. Graham. 2015. Measuring student engagement in technology-mediated learning: A review. *Computers & Education,* 90, 36-53.

[13] B. W. Schuller. 2018. Speech emotional recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM,* 61(5), 90-99.

[14] P. Juslin and K. Scherer. 2008. Speech emotion analysis. *Scholarpedia.org,* 3(10), 4240. [Online]. Available: http://www.scholarpedia.org/article/Speech_emotion_analysis

[15] A. M. Kondoz. 2004. *Digital Speech*, John Wiley & Sons Ltd, West Sussex, England.

[16] M. H. Hayes. 1996. *Statistical Digital Signal Processing and Modeling*, John Wiley & Sons Ltd, West Sussex, England.

[17] L. Liu, Y. Li and K. Kuo. March 2018. Feature Extraction and Recognition of Infant Cry Signals, *IEEE International Conference on Information and Computer Technologies*, 159-163.

[18] L. Liu, W. Li, X. Wu and B. Zhou. 2019. Infant Cry Language Analysis and Recognition: An Experimental Approach, *IEEE/CAA Journal of Automatic Sinica*, 6(3), 778-788.

[19] R. Steuer, J. Kurths, C. O. Daub, J. Weise and J. Selbig. 2002. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18, 231-240.

[20] H. Kantz, J. Kurths and G. Mayer-Kress. 2012. Nonlinear analysis of physiological data. *Springer Science Business Media*, Secaucus, NJ.

[21] E. Koechlin and C. Summerfield. 2007. Handbook of Biological Statistics. *Trends in Cognitive Sciences*, 11(6), 229-235.

[22] G. Jensen, R. D. Ward and P. D. Balsam. 2013. Information: theory, brain, and behavior. *Journal of the Experimental Analysis of Behavior*, 100(3), 408-431.

[23] T. DelSole and M. K. Tippett. 2007. Predictability: Recent insights from information theory. *Reviews of Geophysics*, 45(4).

[24] T. M. Cover and A. J. Thomas. 2004. *Elements of Information Theory*. John Wiley Sons Ltd, West Sussex, England.

[25] A. Kraskov, H. Stögbauer and P. Grassberger. 2004. Estimating mutual information. *Physical Review E*, 69(6).

[26] C. Bandt and B. Pompe. 2002. Permutation entropy: a natural complexity measure for time series. *Physical Review Letters*, 88(17).

[27] A. M. Kondoz. 2004. *Digital Speech* John Wiley Sons Ltd, West Sussex, England.

[28] A. J. Oxenham. 2012. 'Pitch Perception', *Journal of Neuroscience*, 32(39), 13335-13338, September 2012; DOI: https://doi.org/10.1523/JNEUROSCI.3815-12.

[29] A. M. Briesch, E. M. Hemphill, R. J. Volpe and B. Daniels. 2015. An evaluation of observational methods for measuring response to classwide intervention, *School Psychology Quarterly*, 30(1), 37-49.

[30] J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data, *Biometrics*, 33, 159-174.

[31] J. H. McDonald. 2009. An information theoretical approach to prefrontal executive function. *Sparky House Publishing*, Baltimore, MD.

[32] S. Butail, F. Ladu, D. Spinello and M. Porfiri. 2014. Information flow in animal-robot interactions. *Entropy*, 16(3), 1315-1330.

[33] G. Nagymáté and R. M. Kiss. 2018. Application of OptiTrack motion capture systems in human movement analysis: A systematic literature review. *Recent Innovations in Mechatronic*, 5(1).