Using the Tablet Gestures and Speech of Pairs of Students to Classify Their Collaboration

Sree Aurovindh Viswanathan and Kurt VanLehn

Abstract—Effective collaboration between student peers is not spontaneous. A system that can measure collaboration in real-time may be useful, as it could alert an instructor to pairs that need help in collaborating effectively. We tested whether superficial measures of speech and user interface actions would suffice for measuring collaboration. Pairs of students solved complex math problems while data were collected in the form of verbal interaction and user action logs from the students' tablets. We distinguished four classifications of interactivity: collaboration, cooperation, high asymmetric contribution and low asymmetric contribution. Human coders used richer data (several video streams) to choose one of these codes for each episode. Thousands of features were extracted computationally from the log and audio data. Machine learning was used to induce a detector that also assigned a code to each episode as a function of these features. Detectors for combinations of codes were induced as well. The best detector's overall accuracy was 96 percent (kappa = 0.92) compared to human coding. This high level of agreement suggests that superficial features of speech and log data do suffice for measuring collaboration. However, these results should be viewed as preliminary because the particular task may have made it relatively easy to distinguish collaboration from cooperation.

Index Terms—Collaborative learning, machine learning, educational data mining, learning analytics

1 Introduction

Effective collaboration between student peers is not spontaneous even in the presence of computer support. Therefore, it is a worthwhile endeavor to develop technologies that are able to automatically assess the quality of collaboration and in this way provide the teacher with information to improve their ability to guide, assist and scaffold student groups. This paper presents a method of measuring collaboration based on data collected from tablet computers used by small groups of students who are solving problems together in a lab setting. Two sources of data were used for measuring collaboration: log data (i.e., recordings of the user interface actions done on the tablets) and the audio recordings from headset microphones worn by students. For analysis of the speech data, only low-level acoustic and prosodic features were used, such as pitch, jitter, shimmer and linear spectral features. No attempt was made to convert the speech to text or to understand it. The log data were also analyzed in terms of low-level features; no attempt was made to recognize the plans and goals of the students. The features were extracted by algorithms, and no human coders were involved in feature extraction.

In order to construct and evaluate the collaboration measure, the judgements of human coders were used as a "gold standard" classification of the groups' processes. The human judges had much more data about the group, including videos from several cameras, than the data available to the machine learner. Moreover, the judges could understand most of the students' speech, plans and goals. Although human judgment of collaboration is not perfect, the human judges had enough data that their judgments were probably as good as or better than the judgments of instructors in the classroom.

Collaboration measures were induced from the human judgment data using standard machine learning algorithms: random forests and additive logistic regression. Accuracy was measured by 10-fold cross-validation. The results were quite promising.

The paper is organized as follows. The remainder of this section reviews the literature. Section 2 establishes the context and the domain in which this study was conducted. Section 3 and 4 describes the task and software that was used. Section 5 describes the study and data collection. Section 6 describes how human judges coded the data and the operational definitions of collaboration they used. Section 7 describes the machine learning methods that were used to induce the collaboration measures. Section 8 and 9 describes the results, interprets them and their limitations, and suggests future research directions.

1.1 Definitions of Collaboration

When students work in a group, their behavior only sometimes qualifies as collaborative. Dillenbourg, et al. [1] make a clear distinction between collaborative and cooperative learning. Following Roschelle and Teasley [2], they defined collaboration as "mutual engagement of participants in a coordinated effort to solve the problem together." The phrase "coordinated effort" implies that collaborating students' actions have the same immediate goal in mind. In contrast, cooperation refers to a situation in which students

• *The authors are with the School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287. E-mail: {sviswa10, kurt.vanlehn}@asu.edu.*

Manuscript received 17 Dec. 2016; revised 1 May 2017; accepted 2 May 2017. Date of publication 15 May 2017; date of current version 18 June 2018. (Corresponding author: Sree Aurovindh Viswanathan.)
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TLT.2017.2704099

divide their task into subtasks and each subtask is solved with substantial amount of individual effort. In other words, cooperating students work with different immediate goals in mind.

A metaphor from Clark and Brennan [3] illustrates the difference between collaboration and cooperation. If two people are at each end of a table, carrying it into their house, then they are collaborating. On the other hand, if they are each carrying a chair into the house, they are cooperating. In both cases, they share the same top-level goal of moving furniture into the house. However, when collaborating, they share the same immediate goal.

1.2 Why Is It Important to Measure Collaboration?

Collaboration has been embraced by the education and training communities as a 21st century skill [4]. Workplace problem solving is often done in small groups, so effective skill in collaboration should make the groups more effective. Collaborative learning now appears in national pedagogical standards, such as the Common Core State Standards for Mathematics [5, practice #3] and the Next Generation Science Standards [6]. A device that measured collaboration could be useful for both assessment and for adapting instruction.

Collaboration has also been studied as a method for increasing learning, motivation and other valued outcomes. However, such studies have evolved over time. According to Dillenbourg, et al. [1], collaboration was originally compared to working alone or to cooperation in studies that asked which method of learning was more effective. However, it became clear that there were too many forms of collaboration, so asking if "collaborative learning" was effective was a bit like asking whether "taking medicine" was effective. Thus, researchers began asking which instructional conditions caused collaborative learning to be effective.

However, as a complex mixture of results accumulated, two simpler questions about the nature of the results emerged. As stated by Dillenbourg, et al. [1, pg. 200], "The questions 'under which conditions is collaborative learning efficient?' was split into two (hopefully simpler) subquestions: which interactions occur under which conditions, and what effects do these interactions have." This paradigm still seems an apt description of current work on collaborative learning. A device that provided objective, real-time measures of the degree of collaboration would probably be quite useful in such studies.

2 Related Work

Many projects have worked on the challenge of automating the analysis of interaction among group memebers. There are a large number of systems of this kind so they are briefly reviewed based on a two-dimensional classification. Only a few projects fall into the same cell of this classification as our project. They are reviewed in detail.

2.1 Classifying Prior Work by Its Purpose

Some projects have studied students working together over several weeks or a whole semester as they used widely available collaboration tools, such as forums, wikis, email or source code repositories [7], [8], [9], [10], [11], [12], [13]. This kind of project is rather different than our project in both the duration of the activity and the way group interactions are monitored, so such projects will not be reviewed here.

The other major class of projects, of which this project is a member, has students working together on a shared editor or some other shared workspace for at most a few hours. Collaboration is measured as a function of their activity on the shared workspace and/or the communication among group members as they work.

There are a large number of systems of this kind, so they will be briefly reviewed by defining two dimensions, purpose and input, then describing the few systems whose position along these two dimension match the position of the project reported here. The two dimensions are excerpted from several similar multi-dimensional reviews [14], [15], [16], [17], [18].

The first dimension concerns the purpose or function of the collaboration measure. That is, what does the system do with the output data? Extending Soller, et al. [19], this dimension has the following categories:

- *Clustering*: Some projects used unsupervised machine learning methods such as clustering or sequence mining to find common patterns of group behavior and display them to researchers [10], [20], [21].
- *Classification*: These projects used human judges to code group interactions into a variety of collaboration categories, then used supervised machine learning methods to induce classifiers (also called detectors) whose accuracy was measured and reported to researchers [8], [9], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39].
- *Mirroring*: Unlike the preceding two categories of systems, which display their results only to researchers, the remaining categories refer to systems that display their results to students and/or teachers. The first of these categories, traditionally called "mirroring," includes systems that display results to students without making any judgement about whether the performances are good or bad [23], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51]
- *Metacognitive*: These systems are like mirroring systems, except that they also display an assessment of the group's interactions. That is, they visually compare the group's current interactions to what the group processes ideally should be at that time [46], [52], [53], [54], [55], [56], [57], [58].
- *Guiding*: These systems include a real-time classifier, but when it detects problems with the group's interaction (e.g., one student is dominating the interaction; off-task conversation), then the system may offer advice directly to the students. Such systems typically include logic that prevents them from overwhelming the students with advice or giving advice at inopportune times [28], [48], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81]
- *Orchestration*: These projects are like mirroring and guiding, except that the assessments of group processes and advice are displayed on a dashboard held by the teacher instead of being sent to the student [7], [62], [77], [82], [83], [84]. In principle, analyses of

collaboration could be shown to both teachers and students using a large display visible to everyone [e.g., 85] or desk-mounted lanterns [86], [87], but so far, such displays have only been used with task information, such as the group's progress.

- *Restructuring*: Some systems use analyses of a group's interactions to dynamically restructure the activity by, for example, requiring students to take turns [43]. Analyses of collaboration could be used in principle to dynamically assign roles to students, change the script or phases that students are required to follow or even changing the group's membership.

Our project fits into the *classification* category, in that the output of the collaboration detector is compared to the judgements of human coders, and the accuracy is reported to researchers.

2.2 Classifying Prior Work by Type of Input

As mentioned earlier, this project is a member of a large group of project which have students working together on a shared editor or some other shared workspace for at most a few hours. Such projects often give students an individual, private workspace. Their editing actions, both to their individual workspace and the group workspace, may be the only input to the collaboration analyzer [22], [23], [36], [37], [65], [71], [72], [73], [79], [80], [88]. However, in many other projects, a second type of input to the collaboration analyzer is some form of communication among the group members. The communication methods fall into several classes:

- Some systems require users to communicate in a formal language [89].
- Some systems have participants use a small set of buttons, e.g., OK, NOT or ?, to express agreement/disagreement with the most recent edit done on the group editor [90] or other dialogue acts [91].
- Although many systems have students communicate by typing natural language and classifying their contribution using a sentence-opener or a menu of speech acts, some systems ignore the text and use only the students' classifications of their text [24], [26], [29], [52], [61], [62], [63], [68], [70]
- Among systems that allow students to communicate via typing (chat) either with or without sentence openers or other self-classifications of the contributions, some have human "wizards" who select codes in real time [75] while others explored automated analysis of the text using either keywords [28], [32], [59], [66], [71], [83] or machine-learned text classifiers [27], [28], [31], [38], [67], [74], [77], [81], [84]
- Some systems allow participants to converse in unconstrained speech, recorded by individual microphones. Some projects divided the audio signal into periods of talk and silence, then used machine-learned classifiers on the resulting patterns [21], [33], [35], [51], [58], [83], [92]. Other projects used machine-learned detectors operating on a wide variety of acoustic features [34], [39]. Another approach was to use a human wizard to code the speech in real time [64].

Our project falls into the last of the categories listed above. Our students entered actions on a common workspace while communicating via unconstrained speech that is analyzed with a machine-learned detectors.

2.3 Collaboration Detection Projects That Use Speech Input for Classificaiton

Only two projects fall into the same pair of classifications as ours: the projects' purpose is to develop a collaboration detector and measure its accuracy, and the communication among participants is spoken. The two antecedent projects will be reviewed here.

One project was done by Gahgene Gweon and her colleagues. Gweon et al. [39] developed two detectors. One differentiated spoken utterances that contained domain reasoning from those that did not. The other detector split the utterances that contained reasoning into those that built upon earlier reasoning (transacts) and those that did not (externalizations). Both detectors were trained and then evaluated against human-segmented and coded speech. The speech detectors used acoustic, prosodic and other features. The reasoning and transact detectors had moderate accuracy ($F = 0.56$ and $F = 0.35$) compared to a majority-class baseline ($F = 0.20$ and $F = 0.12$). Interestingly, the speech feature that contributed the most to both detectors was the length of the segment. Longer segments of speech tended to have reasoning and to be more likely to follow from earlier segments. Gweon et al. [34] developed a more complicated detector that first assessed the degree of prosodic similarity of the speakers (entrainment; speech style accommodation) and then showed that this machine-learned measure's output was moderately correlated with transactivity ($R = 0.36$) as judged by human coders.

Our project is similar to Gweon's projects in that it used machine-learned classifiers based on low-level features of speech, but it differs in several ways. First, whereas Gweon's classifiers used only the students' speech, ours used their actions as well. This allowed us to compare the accuracy attained from actions alone, speech alone and both actions and speech together. Second, whereas collaboration was the focal code in both Gweon's projects and ours, the two project chose different non-collaboration codes. For example, Gweon's codes did not include cooperation (defined below), perhaps because there was none in her corpus. This choice may impact accuracy, so we measured the accuracy of classifiers trained with different non-collaboration codes.

Our project is similar to one done by Martinez-Maldonado et al. [83] in using both speech and actions. Their analysis of the participants' speech used a silence detector to convert the speech into a binary feature (present versus absent). Similarly, they abstracted participant's user interface actions to present versus absent. Recordings of the sessions were divided into 30 second segments, and features were assigned based on the number of participants acting, dispersion of participants' activity (Gini coefficient), the number of participants talking, the total duration of their talk, and the dispersion of their talking. Machine-learned detectors induced from these features were only moderately accurate [33], [35]. Increasing the length of the segments to 60 seconds or 90 seconds did not have much impact on accuracy [33]. The group next used differential sequence mining to find sequences of speech, silence and action that would reliably split groups into high and low collaboration [21], [35], [36].

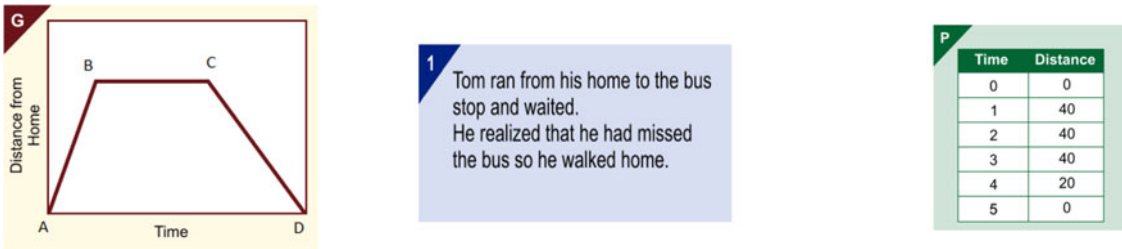


Fig. 1. A row in the card matching problem describing a story.

Hundreds of such patterns were found, and some made sense with respect to common views of collaboration. However, the researchers did not convert any of these patterns to classifiers in order to measure their accuracy, perhaps because they were concerned that the patterns overfit the data.

Generalizing across these two projects, it appears that a simple binary analysis of the speech signal into "talk" versus "silence" suffices for achieving moderate accuracy in collaboration detection. This was a Martinez-Maldonado finding in several studies, and it is consistent with the strength of such features in Gweon's detectors.

Building on these results, our research questions are: (1) What accuracy could be achieved with the addition of other acoustic or prosodic features? (2) What is the relative accuracy of detectors using speech alone, actions alone or the combination of speech and action? Our project has addressed these questions, as well as studying classification accuracy in yet another task domain.

3 Collaborative Card Sorting

This section describes the mathematical problem that the study subjects solved. As will be seen, its characteristics simplify detection of collaboration. This problem was developed by Mathematics Assessment Project (map.mathshell.org), whose problems have been used by hundreds of teachers over several years. They are rich and complex, and are intended to be used for both instruction and formative assessment. This particular problem asks students to interpret distance-time graphs abstractly and quantitatively. Students must be able to interpret slopes of these graphs, to make arguments about their hypothesis and to critically reason about the arguments made by their peers.

3.1 Cards

Students are given a table with three columns and nine rows and a set of 27 cards to put into the 27 cells of the table. A few of the cards are blank, and students must fill them out in a way that makes the blank cards fit properly in the table. There are three types of cards: Graph cards, table cards and story cards. The graph cards are already positioned in the left column of the table. The students should position the story cards in the middle column and the table cards in the right column. All the cards in a row should describe the same process, which involves Tom making a short journey. Fig. 1 shows one such row. Its graph card has four inflection points: A, B, C and D. Students should make the following inferences about them: a) Point A: Tom is at his home b) Line segment A-B: Tom moved away from home at a fast pace. Hence the segment has a steep slope c) Point B: Tom stopped d) Line segment B-C: Tom waited in the same location. e) Point C: Tom started moving again f) Line segment C-D: Tom returned back home at slow pace g) Tom reached home.

The cards are designed around commonly observed misconceptions. For instance, students often view the graphs as a cross-section, so they often match the card in Fig. 1 to the following story "Opposite to Tom's home is a hill. Tom climbed slowly up the hill, walked across the top and rank quickly towards the down the other side". Our subjects found this task rather difficult, but all were able to solve it in less than 90 minutes.

This task consists of 18 subtasks, one for each card to be placed into the table. We expected to see collaborating student working together on placing a card, whereas cooperating students would work simultaneously on placing different cards. We selected this task because we believed that human coders could easily distinguish collaboration from cooperation, and we wanted to see if log data and speech data would allow the machine to do so as well.

4 System Setup

This section describes the hardware and software setup that was used. Although students worked in pairs and sat beside each other at a table, they each had their own tablet, which was a Samsung Galaxy Note 10.1. The tablet had active digitizer technology and a stylus that allowed students to write easily and legibly on the 10 inch touchscreen. The tablets were connected via a wireless network to a laptop computer that acted as a server. Although not important for this study, the reader may be interested to know that when the system is used in a classroom, it consists of 36 tablets, a laptop, a wireless access point and a wireless adapter that plugs into the classroom digital display projector. The whole rig is portable, weighs about 50 pounds, and packs into 4 toolboxes. It has been used in approximately 30 classroom trials so far. However, because no speech was collected during the classroom trials, the classroom log data are not analyzed here.

The software used by participants is called FACT, an acronym for Formative Assessment using Computational Technology (fact.engineering.asu.edu). The FACT user interface is intended to mimic a large poster upon which students can write and place small cards. They can also write on the cards and move them. The distance-time problem was originally developed using real posters (about 24" by 36") and cards (about 1" by 2").

Of course, students cannot see much of the poster using a tablet with a small screen, so FACT provides the usual swiping and pinching gestures for scrolling and zooming the poster. Users must use two fingers for scrolling, because the single-finger drag is used for moving cards. All writing is done with the stylus. Although the user interface has

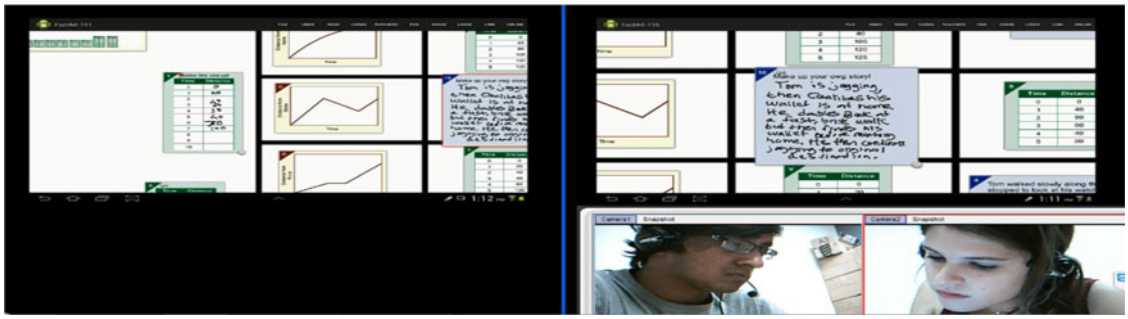


Fig. 2. Snapshot of desktop recording student's action along with gestures.

other features (e.g., for resizing cards), the only ones necessary for solving the time-distance problem were scrolling, zooming, moving cards and writing.

FACT operates in two different modes: solo and group. In solo mode, students work on their own poster with their own cards. In group mode, students work on the same poster and set of cards. They can scroll and zoom independently, so they can be working on different parts of the poster at the same time. If they happen to be viewing the same part of the poster, then card movements and digital ink are visible more-or-less simultaneously on both tablets.

All user actions are recorded as data structures called events, which are sent to the server as soon as possible and stored as log data. When FACT is in group mode, events from one tablet are transmitted quickly to the other tablet, thus allowing both users to see the same poster. The events contain the start and end time of the user action as well as parameters that describe the event precisely enough to replicate it on the other tablet.

5 Methodology

5.1 Participants and Duration

The study was conducted in the lab setting. A total of 28 participants were enrolled in the study, and were compensated $10 per hour for their participation. They were a mix of graduate and undergraduate students from Arizona State University. None of the students found the problem simple, which is a bit surprising given that the problem is intended for middle school students.

There were no strict time deadlines enforced for this study. However, the overall session lasted an average of 60-110 minutes. Initially participants were briefed about the activities that they were expected to perform during this study and voluntary consent was obtained. A pretest gauged their mathematical ability before the actual session. No students were eliminated due to high scores. On completion of the pretest, students were given a paper list of all the user interface features that they would need. The tablets were then put in solo mode by the experimenter, and students were asked to discover each feature on the list. Pilot studies suggested that this short (e.g., 10 minutes) activity sufficed as user interface training.

The problem solving session started with the experimenter briefly describing the distance time interpretation problem to the student. The FACT system was put in group mode so that both students shared the same poster, and then subjects began to solve the distance time problem. During the problem solving session, they talked with each other and solved the problem together and/or the worked individually. They mostly worked on their own tablets, but occasionally would huddle over one of them. They worked until the entire 9x3 table was filled up with cards. The overall duration of this phase was between 30-40 minutes.

Some students completed the distance time interpretation problems relatively quickly, so they were given additional problems to work, because we are considering using those problems in subsequent experiments. Once the problem solving session was complete, all students were given a posttest, which was identical to the pretest, and a survey. We intend to use the pre-test, post-test and survey data to study learning, but those data have not been analyzed yet.

5.2 Raw Data Collection

The recording setup combined several different input streams:

1. Unidirectional headset microphones were used to capture each user's speech. Each audio stream was recorded on the corresponding tablet for later analysis. The audio streams were also sent to a desktop computer.
2. The tablet screen content was streamed to a desktop computer using an HDMI cable. The video stream was captured at Full HD resolution
3. Log data were collected at the server.
4. Web cameras, one per student, recorded the student's head and shoulders.

The desktop computer showed all four videos on its screen: two tablet screen videos and two head and shoulder videos. It also received the two audio streams. Fig. 2 shows a snapshot of the desktop computer's screen. In order to synch all 6 streams, the desktop screen was saved as a single video with a mono (not stereo, unfortunately) audio track. Thus, all the data sources except the log data were synched as they were recorded.

6 Coding Categories

This section describes the coding categories that were used to create the "gold standard" against which machine-learned detectors would be judged. When the collaboration monitoring system is used the classroom, it should help a teacher make a binary decision—whether to visit a group or not. Thus, only two categories are essential: successful collaboration or not. Although earlier work has often used a coding scheme developed by Meier, Spada and Rumel [93], it produces scores for each episode along nine dimensions which must somehow be aggregated in order to decide whether participants are collaborating during that episode. For instance, Martinez-Maldonado added the 9 scores and

then used a threshold on the sum to divide groups into collaborating versus non-collaborating. Because we did not need the detail of nine dimensions, we simplified the coding task to just discriminating among a handful of categories. Hence we defined four classifications:

- *Collaboration*: The interaction between the pair was considered collaboration when they both worked on placing the same card (i.e., they had the same immediate goal) and each person often built on other's reasoning. In a few other situations, students engaged in argumentative co-construction process by which they resolved their issues and converged on an understanding of the mathematical content. This definition of collaboration includes various characteristics or attributes of joint problem solving such as common ground, knowledge convergence, and co-construction, transactivity, scaffolding contributions and making a shared conception of the problem. The following is an example

  Student A: It is S [a table card] because all other tables are ending at zero.

  Student B: No. This cannot be right. The distance can never be decreasing. In that [card] the distance is decreasing with time. In S. . .

  Student A: 40. . .80. . .60. . .40. . .80. . . [reading the table card] He is never going back.

  Student B: This one is for [the story card where] he has forgotten his watch

  Student A: Oh! Okay hmm. . .
- *Asymmetric Contribution*: The interaction between a pair was considered asymmetric contribution when they were working on placing the same card but one student did most of the work. That is, one person led the conversation and the other person added at most a few reasoning statements. We define two different levels of asymmetric contribution
  - *Asymmetric Contribution (Low)*: The interaction between a dyad was characterized as "low" when no reasoning statements or exchanges occurred, but the human coder could tell from the videos that both students were attending to the same card. The following is an example:

    Student A: For this card. . .

    Student B: Yes Tom is. . . (The person moves the card to the solution grid) Yeah done.

    Student A: (Head Nod and both moves to the next problem)
  - *Asymmetric Contribution (High)*: The interaction between a dyad was characterized as "high" when one person expressed his hypothesis with related reasoning statements and the other person generally accepts the reasoning made by him and they continue to solve the sub problems. The following is an example:

    Student A: Probably, the first one will be 20 40 40 [Reading a table card] and it goes to zero. So that table. . .Because the slope while going up is little longer than the slope while coming down. So T [a table card] goes with E [a graph card]

    Student B: Yeah. . . Yeah. . ..

    This definition of asymmetric contribution shares a few characteristics of joint problem solving sessions such as common ground and establishing a shared conception of a problem. However, it lacks other properties such as transactivity, scaffolding contributions or argumentative co-construction.
- *Cooperation*: The interaction between a dyad was considered cooperation when subjects have different immediate goals, that is, they were working on placing different cards. Although there was usually little or no conversation between the pair, sometimes one student idly chattered about the problem they were trying to solve and the other student did not respond back. Since students worked on different immediate goals, cooperation episodes do not have any characteristic attributes of joint problem solving.

This coding scheme is similar to ones used by other projects. In addition to a Collaboration code, almost all have noted that students sometimes work independently (our Cooperation code) and that sometimes one student is passive while the other does most of the work (our Asymmetric Contribution code). The distinction between High and Low Asymmetric Contributions is included because episodes where one person is explaining their actions while the other appears to listen (coded as High Asymmetric Contribution) may be consider a form of collaboration. In several studies where asymmetric verbal collaboration was pointed out to participants, the passive participants rejected it as a meaningful measure of their participation [51], [58], which suggests that they considered listening intently to be a form of collaboration

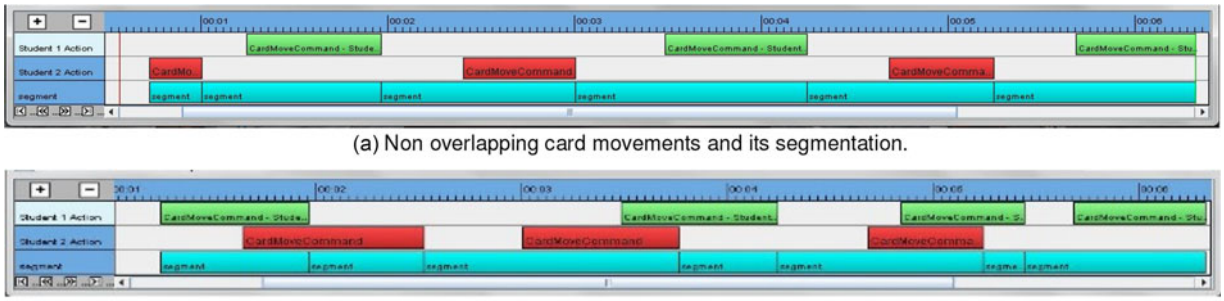
7 Analysis Methods

This section describes the rest of our approach, after collecting the audio data and the logs of student interaction.

7.1 Audio Processing

The audio streams that were collected on the tablets each corresponded to a single speaker's voice. They were used for developing the detectors. The merged audio stream, which was recorded on the desktop, was only used by the human coder.

The first step involved in the audio preprocessing was noise removal. The quality of the audio is important for the extraction of features, and in particular for detecting silence. Because we did not use a microphone array [94], the microphone caught some of the other participants voice. However, the amplitude of the audio signal was louder for the person who was wearing the microphone when compared to the other person's voice. Hence we used an amplitude filter and reduced the impact of this artifact. Background noise was reduced by the following steps. When noise was present, a constant linear band could be seen in the spectrum display above the base of the waveform. For segments where noise was present, the noise profile was extracted by selecting the region of speech data that contains only noise and no audio data. The noise was removed by subtracting its signal from signal of the affected segment. These steps were repeated until no noise could be heard. Both of the above steps were carried out using Audacity [95] software.



(a) Non overlapping card movements and its segmentation.

(b) Overlapping card movements and its segmentation.

Fig 3. (a) Non overlapping card movements and its segmentation. (b) Overlapping card movements and its segmentation.

7.2 Synchronizing and Cleaning the Data

We were surprised to find out how difficult it is to get all the data synchronized. Since this "trick of the trade" isn't covered in the literature we reviewed, we will add a few comments here about synchronization. Log events were created by each tablet separately and sent to the server. In order to detect simultaneity at the millisecond level, it was necessary to insure that both tablets' clocks showed the same time down to the millisecond. FACT was designed to run independently of the web so a universal time source could not be used. Thus, both tablets were synched to the server's time. Although the tablets communicated with the server wirelessly and the wireless transmission delays are not constant, all the computers were in the same small room so our solution seems to have provided sufficiently precise synchronization. Hence we utilized Network time protocol (NTP) to synchronize the clocks together

On the other hand, the audio-video recording, which was taken from the desktop computer as it merged the 4 video streams and 2 audio streams, was time-stamped with the desktop's clock which could be a little bit off from the server's clock. In order to measure the offset, when a session started, the experimenter ostentatiously clicked on a Time Sync menu item on one of the tablets. The current time of the tablet was logged and later compared to the desktop's time for the Time Sync menu event as seen on the videos. Subtracting the two times provided the offset, which was used to synch the two streams.

7.3 Segmentation

Segmentation refers to the process of dividing chronological data into segments or episodes [96]. Segmentation is necessary whenever assigning a code (e.g., "Collaboration" versus "Cooperation") to a whole session would be unreliable and perhaps even nonsensical. Although some projects have used overlapping segments [97], [98], most projects define segments to partition the whole session. Some projects use constant duration segments, such as 30 seconds [99], [100]. However, mechanical segmentation can make it difficult to assign codes reliably, so many projects do manual segmentation. That is, human coders, perhaps aided by a set of written rules (rubrics), decides where to place segment boundaries. As it turns out, we were able to use a novel segmentation method.

The goal of our segmentation process was to divide the solution of the time-distance problem into separate subproblems. Recall that the overall task of the participants was to place 18 cards into 18 cells in a table such that the cards in a row all described the same journey. A sub-problem is considered "done", and a segment boundary is placed, whenever a story or table card was placed in a table cell and then the participant(s) moved on to another card. If the participants come back later and move that card to a different cell, this new placement is considered a new segment.

A simple piece of software inserted a segment boundary whenever a card was placed and the person who placed it (or both people, if they are collaborating) had ceased to move it. That is, the next card event in the log file involved a different card. Figs. 3a. and 3b. shows two cases illustrating this definition. Three tiers are shown in each figure. Student 1 Action tier and Student 2 action tier captures the card-move events of the two participants. The third tier shows the segments and their boundaries. Fig. 3a. shows non-overlapping card-moving events. Fig. 3b. shows overlapping ones.

7.4 Human Coding

Once the segmentation was performed, human annotators classified each segment as either cooperative (P), Low asymmetric contribution (L-A), High asymmetric contribution (H-A) or collaboration (C). The annotators used all the data available, including the audio-video stream and the log data, which were separated into two tiers according to participant, as shown in figus. 3a and 3b. Two human coders tagged a sample of 35 percent of the overall segments. Interrater agreement was considered acceptable with Cohen's kappa $\mathrm{K} = 0.78$. For consistency across the whole dataset, the classifications of one annotator (the first author) were used in subsequent analyses.

7.5 Feature Extraction from Log Data

The goal of the feature extraction process was to obtain superficial features from the students' work that could potentially differentiate between collaboration, cooperation and asymmetric contribution. Features were extracted computationally from audio files and log files; video data were ignored. The rest of this section describes some of the key ideas. In inventing these features, we borrowed liberally from the sequence mining results of Martinez-Maldonado et al. [35], [101].

Features were assigned to segments based on the distribution of specific events that occurred inside the segment. For instance, one type of event was a student moving a card. Each card movement had a duration. Sometimes a student would move a card, pause, move it a bit more, pause, etc. When multiple events of the same type occur in a segment, then it is unclear how best to aggregate them. For instance, how should the durations of all the card

movements in the segment be aggregated? In such cases, 7 segmental features were assigned with values equal to the mean, median, standard deviation, inter-quartile range, minimum, maximum and count of the events

When students collaborate with each other, they tended to look at the same card. In order to do so, one student decided on which sub-problem to solve and prompted the other person to look at the same sub-problem. Typically, the first person held his tablet's view of the poster and the other person scrolled and zoomed until his tablet displayed the same part of the poster. For this process to be captured as features, each zoom event was categorized as "read" or "search" depending on the total amount of time elapsed between each scroll/zoom command. Searching was a sequence of scrolls or zooms with little time in between them. In order to determine if the two tablets were viewing the same portion of the poster, the distance between the centers of participants' viewports and the area of overlap were calculated.

In addition to the features that applied to a single segment, some features were defined that captured the past behavior of the students up to and including this segment. Some examples are:

- A card is moved twice by the same person to two different table cells
- A card is moved twice, by two different people, to two different table cells
- Total number of sub-problems solved so far
- Type of card the student worked on during consecutive move operations
- The person who moved the card to a position in the solution grid later changed it to another position after some talk.

Also, many of the features that applied to a single segment (e.g., duration of searching events) were also applied to the time from the beginning of the problem to the current segment

As the discussion above suggests, the log data features were specific to the time-distance problem. In future work, we hope to generalize some of them (e.g., whether the two participants are looking at the same part of the poster). The discussion section considers the challenges of generalization

7.6 Feature Extraction from Audio Data

The duration of the time when students talk with each other (speech time) and the duration of time when students did not talk with each other (silence time) were measured using the "Sound Finder" feature of Audacity. The audio levels below -26 dB were treated as silence and the minimum duration between two audio signals to be considered for silence is taken to be 1 second.

The remaining audio features were extracted using the OpenSMILE audio feature set, which represents the "state of the art for affect and paralinguistic recognition" [102]. The features used are shown in Table 1. Note that none of the features are specific to the time-distance problem, so perhaps a collaboration detector developed using these audio features would work well on other problems.

7.7 Feature Selection

Feature selection was performed because the number of features was greater than the number of observations. Pairwise correlations were performed on features likely to be redundant. Sets of highly correlated features (coefficient > 0.9) were reduced to a single feature chosen arbitrarily from the set. Next, we applied resampling of the attributes in order have uniform distribution across class labels. Finally, we applied an attribute selection algorithm using best first search in Weka in order to reduce the feature set further. This reduced the set of features were used to train the model. The total number of features obtained is summarized later in Table 5.

TABLE 1
Audio Features from Open SMILE Toolkit

Source of Feature	Total Number of feature
MFCC (short term power spectrum)	630 (15 Fe * 15 D * 21 F)
Mel Frequency	336 (8 Fe * 8 D * 21 F)
Linear spectral coefficient	336 (8 Fe * 8 D * 21 F)
Loudness	42 (1 Fe * 1 D* 21 F)
Voicing	42 (1 Fe * 1 D * 21 F)
Fundamental Frequency Envelope	42 (1 Fe * 1 D * 21 F)
Pitch	38 (1 Fe * 1 D * 19 F)
Jitter	38 (1 Fe * 1 D * 19 F)
Jitter (DP)	38 (1 Fe * 1 D * 19 F)
Shimmer	38 (1 Fe * 1 D * 19 F)
Pitch Onsets	1 Fe
Duration	1 Fe

Fe = Features, D = Deltas; F = Functionals.

8 Results

The overall goal of the study is to induce a classifier that can distinguish collaboration from cooperation. However, there is some ambiguity about how to treat the asymmetric contribution category, so three levels of granularity were defined and a classifier was induced for each:

- *Quaternary*: This classifier was trained to distinguish all four categories coded by the human annotator. That is, its output was drawn from the set: Collaboration, Asymmetric contribution high, Asymmetric contribution low, and Cooperation
- *Ternary:* This classifier lumped together the two Asymmetric contribution categories, so its output was drawn from the set: Collaboration, Asymmetric contribution and Cooperation.
- *Binary:* This classifier lumped Collaboration with Asymmetric contribution, so its output was drawn from the set: Non-cooperation and Cooperation

8.1 Results from the Binary Classifier

This section reports on the binary classifier, which was trained to discriminate only two categories: Cooperation versus Non-cooperation, where the latter category includes Collaboration and Asymmetric contribution codes. We built classifiers for both the audio data alone, the log data alone and both sources of data combined

Random forest yielded the best result for all three data sets when compared to other algorithms such as J48 graft, bagging, additive logistic regression and boosting. Models were evaluated using the tenfold cross validation

The confusion matrix for the binary classifier is shown in Table 2. The classifiers correctly classified approximately 93 percent of the 325 episodes, and their accuracies were

TABLE 2
Confusion Matrices and Accuracies for the Binary Classifiers

		Predicted Class (Audio) 93% ($\kappa = 0.85$; F = 0.95)		Predicted Class (Log) 92%($\kappa = 0.83$; F = 0.94)		Predicted Class (combined) 96% ($\kappa = 0.92$; F = 0.97)	
		NP	P	NP	P	NP	P
True Class	NP	203	6	199	10	207	2
	P	16	100	15	101	10	106

P = Cooperation, NP = Non - Cooperation.

TABLE 3
Confusion Matrices and Accuracies for the Ternary Classifiers

		Predicted Class (Audio) 88% ($\kappa = 0.82$)			Predicted Class (Log) 85% ($\kappa = 0.78$)			Predicted Class (Combined) 86% ($\kappa = 0.79$)		
		C	A	P	C	A	P	C	A	P
True Class	C	88	10	1	80	11	8	82	12	5
	A	12	87	8	7	87	13	12	89	6
	P	1	8	110	4	5	110	1	9	109

P = Cooperation, C = Collaboration, A = Asymmetric Contribution.

TABLE 4
Confusion Matrices and Accuracies for the Quaternary Classifier

		Predicted Class (Audio) 85% ($\kappa = 0.79$)				Predicted Class (Log) 77% ($\kappa = 0.66$)				Pred. Class (Combined) 87% ($\kappa = 0.81$)			
		Cc	H-A	L-A	P	C	H-A	L-A	P	C	H-A	L-A	P
True Class	C	64	10	5	3	58	1	3	20	74	7	1	0
	H-A	11	27	4	2	3	19	1	21	9	29	3	3
	L-A	5	2	57	3	2	0	45	20	5	1	55	6
	P	0	2	2	128	1	1	1	129	0	3	4	125

P = Cooperation, C = Collaboration, H-A = High Asymmetric Contribution, L-A = High Asymmetric Contribution.

close to each other. Kappas for all three classifiers are discussed later.

8.2 Results from the Ternary Classifier

The ternary classifier distinguished between Collaboration, Cooperation and Asymmetric contribution, where low and high asymmetric contribution was lumped together into one category. Additive logistic regression performed the best for audio and combined feature sets while random forests yielded the best result for features extracted from logs. Models were evaluated using the tenfold cross validation

The confusion matrix for the ternary classifier is shown in Table 3. The overall performance of logs, audio and combined feature sets were similar to each other, approximately 87 percent.

8.3 Results from the Quaternary Classifier

This classifier used the same codes as the human annotators. Random forests performed the best for audio and combined feature sets while additive logistic regression performed the best for log-based feature sets.

The confusion matrix for the quaternary classifier is shown in Table 4. Accuracies were high for the audio and combined classifiers, but not quite as high for the classifier that used log data only. This makes sense, because the difference between high and low asymmetric collaboration depends on the amount of conversation by the more passive participant, and hence it would have been difficult for the log based features to detect this.

8.4 Summary

Table 5 compares the 9 classifiers by showing their accuracy, kappa and the total number of features included in the detector. The kappa and accuracy of the binary and ternary classifier were comparable across categories of input, whereas the log classifier with the quaternary output performed poorly when compared to the audio and combined categories, as explained earlier. As one would expect, the simplest classification scheme, binary, enjoyed the highest accuracy.

9 Discussion and Conclusion

When this project began, we did not think the induced detectors would be accurate because they used only low-level features that do not understand what the participants are saying nor what plans and goals the task involves. Against these low expectations, the results were surprisingly good, with accuracies between 87 and 96 percent.

Besides exploring collaboration detection in a new task domain, this project addressed two research questions: (1) what accuracy could be achieved with the addition of low-level acoustic or prosodic features beyond simple silence detection? (2) What is the relative accuracy of detectors

TABLE 5
Summary of All Nine Classifiers

	Binary Classifier			Ternary Classifier			Four way Classifier		
Input	Kappa	Accuracy	Total # Features	Kappa	Accuracy	Total # Features	Kappa	Accuracy	Total # Features
Audio	0.85	93.23	11	0.82	87.69	14	0.79	84.92	15
Log	0.83	92.30	10	0.78	85.23	9	0.66	77.23	15
Combined	0.92	96.30	15	0.79	86.15	17	0.81	87.06	22

using speech alone, actions alone or the combination of speech and action?

The answer to the second question is addressed in Table 5, which shows that log data alone provide the least accuracy, whereas audio alone was often not much different than the combination of audio and log data. The answer to the first question is more complex, as it involves an uncontrolled comparison of our accuracy results to those of our predecessors. Although our accuracies are high, it now appears to us that this is at least partly due to the differences in task.

First, the task used here was much like moving furniture or assembling a jigsaw puzzle. When people are physically collaborating, they are moving the same physical object or at least, one person is moving it and the other is watching and offering comments. On the other hand, when people are cooperating, they are simultaneously moving different objects. In fact, a feature used in all the log data detectors was whether the two participants moved different cards simultaneoulsy.

In all these task domains, subtasks correspond to moving an object into a location. We are not sure how well our method of using low level features will work when the task does not align subproblems with object movements. Thus, our next study (which is in progress) uses a task with no moving objects: Two people are sharing responsibility for providing written answers to questions.

A second limitation is that the use of an object-moving task allowed automation of segmentation. Usually, segmentation into subtasks has to be done by a human annotator who can understand the participants' speech, plans and goals. We are not sure if subtask-based segmentation can be automated with problems where the subtask boundaries are less salient.

The third limitation was that the audio collection and cleaning used here would not be robust enough for use in classrooms. We are currently working on a better audio collection method that makes it unnecessary to clean the audio data before processing. Synchronization also needs to be improved, as it currently requires too much human attention.

A fourth limitation is that we used machine learning to induce the detectors, which means that a gold standard must be available to train the detectors. The gold standard was provided by human coders here, but that may not be sustainable in practice. However, we are encouraged by the results from the audio-only detector. This detector does not "know" anything about the task being done by the participants. Although more studies are needed in order to see if this finding occurs again, we are hopeful that the following, more sustainable method would work:

- A classroom of students wears headsets with high-quality, close-talk, noise cancelling microphones.
- The audio is processed automatically, with no human help, to code episodes as collaborative versus cooperative.
- The audio-generated coding is used to train a log-data detector using the log data from the class that wore microphones.
- Now the log-data detector can be used in subsequent class; the microphones no longer need to be used.

If this method works, then it may become practical to equip a variety of software intended to be used collaborative with detectors which can be used in real world learning environments. We envision the collaboration signal being sent to a handheld teacher dashboard, like the one use by Martinez-Maladondo [83] and the FACT project. However, it could also be sent to the students themselves.

Acknowledgments

This research was funded by The Diane and Gary Tooker Chair for Effective Education in Science, Technology, Engineering and Math, and by the Bill and Melinda Gates Foundation under grant OPP106128.

References

[1] P. Dillenbourg, M. Baker, A. Blaye, and C. O'Malley, "The evolution of research on collaborative learning," in *Learning in Humans and Machine: Towards an Interdisciplinary Learning Science*, H. Spada and P. Reimann, Eds. Amsterdam, Netherlands: Elsevier, 1996.

[2] J. Roschelle and S. D. Teasley, "The construction of shared knowledge in collaborative problem solving," in *Computer-Supported Collaborative Learning*, C. O'Malley, Ed. Heidelberg, Germany: Springer-Verlag, 1995.

[3] H. H. Clark and S. E. Brennan, "Grounding in communication," *Perspectives on Socially Shared Cognition*. Washington, DC, US: American Psychological Association, Jan. 01, 1991.

[4] F. W. Hesse, E. Care, J. Buder, K. Sassenberg, and P. Griffin, "A framework for teachable collaborative probelm solving skills," in *Assessment and Teaching of 21st Century Skills*, P. Griffin and E. Care, Eds. Berlin, Germany: Springer, 2015, pp. 37–56.

[5] CCSSO, "The common core state standards for mathematics," Oct. 31, 2011. [Online]. Available: www.corestandards.org

[6] N. L. States, Next Generation Science Standards: For States, By States, Spi edition. Washington, D.C: National Academies Press, 2013.

[7] M. J. Rodriguez-Triana, A. Martinez-Mones, J. I. Asensio-Perez, and Y. Dimitriadis, "Scripting and monitoring meet each other: Aligning learning analytics and learning design to support teachers in orchestrating CSCL situations," *British J. Educ. Technol.*, vol. 46, pp. 330–343, 2015.

[8] I.-A. Chounta and N. Avouris, "Time series analysis of collaborative activities," in *Collaboration and Technology*, V. Herskovic, U. Hoppe, M. Jansen, and J. Ziegler, Eds. Berlin, Germany: Springer, 2012, pp. 145–152.

[9] A. R. Anaya and J. g. Boticario, "Application of machine learning techniques to analyze student interactions and improve the collaboration process," *Expert Syst. Appl.*, vol. 38, pp. 1171–1181, 2011.

[10] D. Perera, J. Kay, I. Koprinska, K. Yacef, and O. R. Zaiane, "Clustering and sequential pattern mining online collaborative learning data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 6, pp. 759–772, Jun. 2009.

[11] A. R. Anaya and J. G. Boticario, "A data mining approach to reveal representative collaboration indicators in open collaboration frameworks," presented at the Edu. Data Mining, Cordoba, Spain 2009.

[12] L. Talavera and E. Gaudioso, "Mining student data to characterize similar behavior groups in unstructured collaboration spaces," presented at the Workshop Artif. Intell. CSCL: 16th Eur. Conf. AI, Chicago, IL, USA, 2004.

[13] S. J. Simoff, "Monitoring and evaluation in collaborative learning environments," presented at the Comput. Support Collaborative Learning, Mahwah, NJ, USA, 1999.

[14] D. Diziol and N. Rummel, "How to design support for collaborative e-learning: A framework of relevant dimensions," in E-Collaborative Knowledge Construction: Learning from Computer-Supported and Virtual Environments, B. Ertl, Ed. Hershey, PA, USA: IGI Global, 2010, pp. 162–179.

[15] N. Rummel, E. Walker, and V. Aleven, "Different futures of adaptive collaborative learning suport," *Int. J. Artif. Intell. Edu.*, vol. 26, pp. 784–795, 2016.

[16] A. Soller, A. Martinez, P. Jermann, and M. Muehlenbrock, "From mirroring to guiding: A review of state of the art technology for supporting collaborative learning," *Int. J. Artif. Intell. Edu.*, vol. 15, pp. 261–290, 2005.

[17] I. Magnisalis, S. Demetriadis, and A. Karakostas, "Adaptive and intelligent systems for collaborative and learning support: A review of the field," *IEEE Trans. Learning Technologies*, vol. 4, no. 1, pp. 5–20, Jan.-Mar. 2011.

[18] K. VanLehn, "Regulative loops, step loops and task loops," *Int. J. Artif. Intell. Edu.*, vol. 26, pp. 107–112, 2016.

[19] A. Soller, A. Martínez, P. Jermann, and M. Muehlenbrock, "From mirroring to guiding: A review of state of the art technology for supporting collaborative learning," *Int. J. Artif. Intell. Edu.*, vol. 15, pp. 261–290, 2005.

[20] R. Martinez-Maldonado, K. Yacef, J. Kay, A. Al-Qaraghuli, and A. Kharrufa, "Analyzing frequent sequential patterns of collaborative learnign activity around an interactive tabletop," presented at the 4th Int. Conf. Edu. Data Mining, Eindhoven, the Netherlands, 2011.

[21] R. Martinez-Maldonado, J. Kay, and K. Yacef, "An automatic approach for mining patterns of collaboration around an interactive tabletop," in *Artificial Intelligence in Education*, K. Yacef, Ed. Berlin, Germany: Springer-Verlag, 2013, pp. 101–110.

[22] M. Muehlenbrock and U. Hoppe, "Computer supported interaction analysis of group problem solving," presented at the Conf. Comput. Support Collaborative Learning, Palo Alto, CA, USA, 1999.

[23] M. Muhlenbrock and U. Hoppe, "A collaboration monitor for shared workspaces," presented at the Int. Conf. Artif. Intell. Edu., Berlin, Germany, 2001.

[24] A. Soller, J. Wiebe, and A. Lesgold, "A machine learning approach ot assessing knowledge sharing during collaborative learning activities," in *Proc. Int. Conf. Comput. Supported Collaborative Learning*, 2002, pp. 128–137.

[25] M. A. Redondo, C. Bravo, J. Bravo, and M. Ortega, "Applying fuzzy logic to analyze collabortive learning experiences in an e-learning environment," *J. United States Distance Learning Assoc.*, vol. 17, pp. 19–28, 2003.

[26] A. Soller, "Computational modeling and analysis of knowledge sharing in collaborative distance learning," *User Model. User-Adapted Interaction*, vol. 14, Jan. 01, 2004, Art. no. 351.

[27] P. Donmez, C. P. Rose, K. Stegmann, A. Weinberger, and F. Fischer, "Supporting CSCL with automatic corpus analysis technology," in *Proc. Int. Conf. Comput. Supported Collaborative Learning*, 2005, pp. 125–134.

[28] B. Goodman, F. Linton, R. Gaimarl, J. M. Hitzeman, H. J. Ross, and G. Zarrella, "Using dialogue features to predict trouble during collaborative learning," *User Model. User-Adapted Interaction*, vol. 15, pp. 85–134, 2005.

[29] R. Duque and C. Bravo, "A method to classify collaboration in CSCL systems," in *Proc. 8th Int. Conf. Adaptive Neural Comput.*, 2007, pp. 649–656.

[30] S. Ravi and J. Kim, "Profiling student interactions in threaded discussions with speech act classifiers," in *Proc. Artif. Intell. Educ. Conf.*, 2007, pp. 357–364.

[31] C. P. Rose, et al., "Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning," *Int. J. Computer-Supported Collaborative Learning*, vol. 3, pp. 237–271, 2008.

[32] T. Dragon, M. Floryan, B. P. Woolf, and T. Murray, "Recognizing dialogue content in student collaborative conversation," presented at the Intell. Tutoring Syst., Berlin, Germany, 2010.

[33] R. Martinez-Maldonado, J. Wallace, J. Kay, and K. Yacef, "Modelling and identifying collaborative situations in a collocated multi-display groupware setting," in *Proc. Int. Conf. Artif. Intell. Edu.*, 2011, pp. 196–204.

[34] G. Gweon, M. Jain, J. McDonough, B. Raj, and C. P. Rose, "Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation," *Int. J. Comput.-Supported Collaborative Learning*, vol. 8, pp. 245–265, 2013.

[35] R. Martinez-Maldonado, Y. Dimitriadis, A. Martinez-Mones, J. Kay, and K. Yacef, "Capturing and analyzing verbal and physical collaborative learning interactions at an enriched interactive tabletop," *Computer-Supported Collaborative Learning*, vol. 8, pp. 455–485, 2013.

[36] R. Martinez-Maldonado, K. Yacef, and J. Kay, "Data mining in the classroom: Discovering groups' strategies at a multi-tableop environment," presented at the Int. Conf. Edu. Data Mining, Memphis, TN, USA, 2013.

[37] I.-A. Chounta and N. Avouris, "It's all about time: Towards the real-time evaluation of collaborative activities," in *Proc. IEEE 14th Int. Conf. Advanced Learning Technol.*, 2014, pp. 383–285.

[38] M. Joshi and C. P. Rose, "Using transactivity in converstation for summarization of educational dialogue," in *Proc. Annu. SLATE Conf.*, 2007, pp. 53–56.

[39] G. Gweon, P. Agarawal, B. Raj, and C. P. Rose, "The automatic assessment of knowledge integration processes in project teams," presented at the Int. Conf. Comput. Supported Collaborative Learning, Hong Kong, China, 2011.

[40] C. Gutwin, G. Stark, and S. Greenberg, "Support for workspace awareness in educational groupware," presented at the ACM Conf. Comput. Supported Collaborative Learning, Indiana University, Bloomington, IN, USA, 1995.

[41] J. Donath, K. Karahalios, and F. Viegas, "Visualizing conversations," *J. Comput.-Mediated Commun.*, vol. 4, 1999, http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.1999.tb00107.x/full

[42] C. Plaisant, A. Rose, G. Rubloff, R. Salter, and B. Shneiderman, "The design of history mechanisms and their use in collaborative educational simulations," presented at the Conf. Comput. Support Collaborative Learning, Palo Alto, CA, USA, 1999.

[43] A. Vizcaino, J. Contreras, J. Favela, and M. Prieto, "An adaptive, collaborative environment to develop good habits in programming," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2000, pp. 262–271.

[44] B. Wasson, F. Guribye, A. I. Morch, and E. Andreassen, "Project DoCTA: Design and use of collaborative telelearning artefacts," Forsknings-og Kompetansenettverk for IT i Utdannning, Universitetet i Oslo, Oslo, Norway, 2000.

[45] B. Goodman, M. Geier, L. Haverty, F. Linton, and R. McCready, "A fremework for asynchronous collaborative learning and problem solving," presented at the 10th Int. Conf. Artif. Intell. Edu., San Antonio, TX, USA, 2001.

[46] P. Jermann, "Computer support for interaction regulation in collaborative problem-solving," PhD dissertation, Psychology and Science Education, University of Geneve, Geneva, Switzerland, 2004.

[47] W. Chen, "Supporting teachers' intervention in collaborative knowledge building," *J. Netw. Comput. Appl.*, vol. 29, pp. 200–215, 2006.

[48] J. Zumbach, P. Reimann, and S. C. Koch, "Monitoring students' collaboration in computer-mediated collaborative problem-solving: Applied feedback approaches," *J. Edu. Comput.*, vol. 35, pp. 399–424, 2006.

[49] J. Janssen, G. Erkens, G. Kanselaar, and J. Jaspers, "Visualization of participation: Does it contribute to successful computer-supported collaborative learning?" *Comput. Edu.*, vol. 49, pp. 1037–1065, 2007.

[50] P. Jermann and P. Dillenbourg, "Group mirrors to support interaciton regulation in collaborative problem solving," *Comput. Edu.*, vol. 51, pp. 279–296, 2008.

[51] F. Roman, S. Mastrogiacomo, D. Mlotkowski, F. Kaplan, and P. Dillenbourg, "Can a table regulate participation in top level managers' meetings," presented at the 17th ACM Int. Conf. Supporting Group Work, Sanibel Island, FL, USA, 2012.

[52] C. Bravo, M. A. Redondo, M. F. Verdejo, and M. Ortega, "A framework for process-solution analysis in collaborative learning environments," *Int. J. Human-Comput. Studies*, vol. 66, pp. 812–832, 2008.

[53] A. Martinez, Y. Dimitriadis, B. Rubia, E. Gomez, and P. De La Fuente, "Combining qualitative evaluation and social network analysis for the study of classroom social interactions," *Comput. Edu.*, vol. 41, pp. 353–368, 2003.

[54] M. Muhlenbrock, *Action-Based Collaborative Analysis for Group Learning*. Amsterdam, Netherlands: IOS Press, 2001.

[55] K. Nurmela, E. Lehtinen, and T. Palonen, "Evaluating CSCL log files by social network analysis," presented at the Conf. Comput. Support Collaborative Learning, Palo Alto, CA, USA, 1999.

[56] H. Ogata, K. Matsuura, and Y. Yano, "Active knowledge awareness map: Visualizing learners activities in a web-based CSCL environment," presented at the Int. Workshop New Technol. Collaborative Learning, Kanazawa, Japan, 2000.

[57] A. Soller and A. Lesgold, "A computational approach to analyzing online knonwledge sharing interaction," presented at the Artif. Intell. Edu., Sydney, Australia, 2003.

[58] J. M. DiMicco, A. Pandolfo, and W. Bender, "Influencing group participation with a shared display," presented at the Comput. Supported Collaborative Work, Chicago, IL, USA, 2004.

[59] D. Adamson, G. Dyke, H. Jang, and C. P. Rose, "Toward an agile approach to adapting dynamic collaboration support to student needs," *Int. J. Artif. Intell. Educ.*, vol. 24, pp. 92–124, 2014.

[60] G. Ayala and Y. Yano, "A collaborative learning environment based on intelligent agents," *Expert Syst. Appl.*, vol. 14, pp. 129–137, 1998.

[61] N. Baghaei, A. Mitrovic, and W. Irwin, "Supporting collaborative learning and problem-solving in a constraint-based CSCL environment for UML class diagrams," *Comput.-Supported Collaborative Learning*, vol. 2, pp. 159–190, 2007.

[62] B. Barros and M. F. Verdejo, "Analyzing student interaction processes in order to improve collaboration. The DEGREE approach," *Int. J. Appl. Artif. Intell.*, vol. 11, pp. 221–241, 2000.

[63] M. de los Angeles Constantino-Gonzalez, D. Suthers, and J. G. E. de los Santos, "Coaching web-based collaborative learning based on problem solution differences and participation," *Int. J. Artif. Intell. Edu.*, vol. 13, pp. 263–299, 2003.

[64] A. Deiglmayr and H. Spada, "Developing adaptive collaboration support: The example of an effective training for collaborative inferences," *Edu. Psychology Rev.*, vol. 22, pp. 103–113, 2010.

[65] D. Diziol, E. Walker, N. Rummel, and K. R. Koedinger, "Using intelligent tutor technology to implement adaptive support for student collaboration," *Edu. Psychological Rev.*, vol. 22, pp. 89–102, 2010.

[66] J. Israel and R. Aiken, "Supporting collaborative learning with an intelligent web-based system," *Int. J. Artif. Intell. Edu.*, vol. 17, pp. 3–40, 2007.

[67] R. Kumar, C. P. Rose, Y.-C. Wang, M. Joshi, and A. Robinson, "Tutorial dialogue as adaptive collaborative learning support," in *Artificial Intelligence in Education*, R. Luckin, K. R. Koedinger, and J. Greer, Eds. Amsterdam, Netherlands: IOS Press, 2007, pp. 383–390.

[68] M. M. McManus and R. M. Aiken, "Teaching collaborative skills with a group leader computer tutor," *Edu. Inf. Technol.*, vol. 1, pp. 75–96, 1996.

[69] T. Okamoto and A. Inaba, "The intelligent discussion supporting system over the computer network," in *Inf. Technol. Edu. Manage. Schools Future*, A. W. Fung, A. Visscher, B.-Z. Barta, and D. B. Teather, Eds. Berlin, Germany: Springer, 1997, pp. 138–145.

[70] J. Robertson, J. Good, and H. Pain, "BetterBlether: The design and evaluation of a discussion tool for education," *Int. J. Artif. Intell. Edu.*, vol. 9, pp. 219–236, 1998.

[71] M. C. Rosatelli and J. A. Self, "A collaborative case study system for distance learning," *Int. J. Artif. Intell. Edu.*, vol. 14, pp. 97–125, 2004.

[72] S. Suebnukam and P. Haddawy, "Modeling individual and collaborative problem solving in medical problem-based learning," *User Model. User-Adapted Interaction*, vol. 16, pp. 211–248, 2006.

[73] S. Suebnukarn and P. Haddawy, "COMET: A collaborative tutoring system for medical problem-based learning," *IEEE Intell. Syst.*, vol. 22, no. 4, pp. 70–77, Jul.-Aug. 2007.

[74] P. A. Tedesco, "MArCo: Building an artificial conflict mediator to support group planning interactions," *Int. J. Artif. Intell. Edu.*, vol. 13, pp. 117–155, 2003.

[75] D. Tsovaltzi, et al., "Extending a virtual chemistry laboratory with a collaboration script to promote conceptual learning," *Int. J. Technol. Enhanced Learning*, vol. 2, pp. 91–110, 2010.

[76] D. Tsovaltzi, et al., "CoChemEx: Supporting conceptual chemistry learning via computer-mediated collaboration scripts," presented at the 3rd Eur. Conf. Technol. Enhanced Learning, Berlin, Germany, 2008.

[77] A. C. Vieira, L. Teixeira, A. Timoteo, P. A. Tedesco, and F. Barros, "Analyzing on-line collaborative dialogues: The OXEnTCHÊ–Chat," presented at the Int. Conf. Intell. Tutoring Syst., Alagoas, Brazil, 2004.

[78] A. Vizcaino, "Enhancing collaborative learning using a simulated student agent," PhD dissertation, Departamento de Informatica, Universidad de Castilla-La Manch, Ciudad Real, Spain, 2001.

[79] E. Walker, N. Rummel, and K. R. Koedinger, "To tutor the tutor: Adaptive domain support for peer tutoring," in *Proc. 9th Int. Conf. Intell. Tutoring Syst.*, 2008, pp. 626–635.

[80] E. Walker, N. Rummel, and K. R. Koedinger, "CTRL: A research framework for providing adaptive collaborative learning support," *User Model. User-Adapted Interaction*, vol. 19, pp. 387–431, 2009.

[81] E. Walker, N. Rummel, and K. R. Koedinger, "Adaptive intelligent support to improve peer tutoring in algebra," *Int. J. Artif. Intell. Educ.*, vol. 24, pp. 33–61, 2014.

[82] R. Martinez-Maldonado, A. Clayphan, K. Yacef, and J. Kay, "MTFeedback: Providing notifications to enhance teacher awareness of small group work in the classroom," *IEEE Trans. Learning Technol.*, vol. 8, no. 2, pp. 187–200, Apr.-Jun. 2015.

[83] R. Martinez-Maldonado, K. Yacef, and J. Kay, "TSCL: A conceptual model to inform understanding of collaborative learning processes at interactive tabletops," *Int. J. Human-Comput. Studies*, vol. 83, pp. 62–82, 2015.

[84] B. McLaren, O. Scheuer, and J. Miksatko, "Supporting collaborative learning and e-discussions using Artificial Intelligence techniques," *Int. J. Artif. Intell. Educ.*, vol. 20, pp. 1–46, 2010.

[85] S. Do-Lenh, "Supporting reflection and classroom orchestration with tangible tabletops," PhD dissertation, Information and communications, Ecole Polytechiuqe Federale de Lausanne, Lausanne, Switzerland, 2012.

[86] P. Dillenbourg and P. Jermann, "Technology for classroom orchestration," in *New Science of Learning: Cognition, Computers and Collaboration in Education*, M. S. Khine and I. M. Saleh, Eds. New York, NY, USA: Springer, 2010.

[87] P. Dillenbourg, et al., "Classroom orchestration: The third circle of usability," in *Proc. Int. Conf. Comput. Supported Collaborative Learning*, 2011, pp. 510–517.

[88] A. Harrer, R. Hever, and S. Ziebarth, "Empowering researchers to detect interaction patterns in e-collaboration," in *Proc. 13th Conf. Artif. Intell. Edu.*, 2007, pp. 503–510.

[89] P. A. Tedesco, "MArCo: Building an artificial conflict mediator to support group planning interactions," *Int. J. Artif. Intell. Edu.*, vol. 13, pp. 117–155, 2003.

[90] M. de los Angeles Constantino-Gonzalez, D. D. Suthers, and J. G. E. de los Santos, "Coaching web-based collaborative learning based on problem solution differences and participation," *Int. J. Artif. Intell. Edu.*, vol. 13, pp. 263–299, 2003.

[91] G. Ayala and Y. Yano, "A collaborative learning environment based on intelligent agents," *Expert Syst. Appl.*, vol. 14, Jan. 01, 1998, Art. no. 129.

[92] K. Bachour, F. Kaplan, and P. Dillenbourg, "An interactive table for supporting participation balance in face-to-face collaborative learning.," *IEEE Trans. Learning Technol.*, vol. 3, no. 3, pp. 203–213, Jul.-Sep. 2010.

[93] A. Meier, H. Spada, and N. Rummel, "A rating scheme for assessing the quality of computer-supported collaboration processes," *Comput.-Supported Collaborative Learning*, vol. 2, pp. 63–86, 2007.

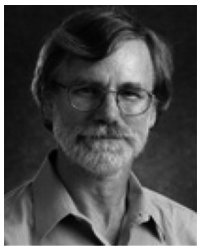
[94] M. Brandstein and D. Ward, Eds., Microphone Arrays: Signal Processing Techniques and Applications. Berlin, Germany: Springer, 2013.

[95] Audacity Team (2014). Audacity(R): Free Audio Editor and Recorder [Computer program]. Version 2.1.0 retrieved 20 Apr., 2014, http://audacity.sourceforge.net/

[96] M. T. Chi, "Quantifying qualitative analyses of verbal data: A practical guide," *J. Learning Sc.*, vol. 6, pp. 271–315, 1997.

[97] C. Rosé, et al., "Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning," *Int. J. Comput.-Supported Collaborative Learning*, vol. 3, Jan. 01, 2008, Art. no. 237.
[98] G. Gweon, M. Jain, J. McDonogh, B. Raj, and C. P. Rose, "Predicting idea co-construction in speech data using insights from sociolinguistics," presented at the Int. Conf. Learning Sci., Sydney, Australia, 2012.
[99] R. Martinez, J. Kay, J. Wallace, and K. Yacef, "Modelling symmetry of activity as an indicator of collocated group collaboration," in *User Modeling, Adaption and Personalization*, J. Konstan, R. Conejo, J. Marzo, and N. Oliver, Eds. Berlin, Germany: Springer, 2011, pp. 207–218.
[100] R. Martinez-Maldonado, J. Kay, and K. Yacef, "An automatic approach for mining patterns of collaboration around an interactive tabletop," in *Proc. Int. Conf. Artif. Intell. Edu.*, 2013, pp. 101–110.
[101] R. Martinez-Maldonado, J. Kay, J. Wallace, and K. Yacef, "Modelling symmetry of activity as an indicator of collocated group collaboration," presented at the Int. Conf. User Model. Adaptation Personalization, Girona, Spain 2011.
[102] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," presented at the Int. Conf. Multimedia, Firenze, Italy, 2010.

Sree Aurovindh Viswanathan received the master's degree in computer science from Arizona State University and is currently working toward the doctoral degree. Previously, he was with Software Engineering and Technology Labs, Infosys Technologies Ltd., India. His current research interests include application of machine learning and data mining techniques to understand student collaborative behavior and intelligent tutoring systems.



Kurt VanLehn is the Diane and Gary Tooker chair of Effective Education in Science, Technology, Engineering and Math in the Ira. Fulton Schools of Engineering, Arizona State University. He has published more than 125 peer-reviewed publications, is a fellow in the Cognitive Science Society, and is on the editorial boards of Cognition and Instruction and the *International Journal of Artificial Intelligence in Education*. His research focuses on intelligent tutoring systems, classroom orchestration systems, and other intelligent interactive instructional technology.