Multimodal, Multiparty Modeling of Collaborative Problem Solving Performance

Shree Krishna Subburaj
Department of Computer Science
University of Colorado Boulder
Boulder, CO, USA
shree.subburaj@colorado.edu

Angela E.B. Stewart
Institute of Cognitive Science &
Department of Computer Science
University of Colorado Boulder
Boulder, CO, USA
angela.stewart@colorado.edu

Arjun Ramesh Rao
Department of Computer Science
University of Colorado Boulder
Boulder, CO, USA
arjun.rao@colorado.edu

Sidney K. D'Mello
Institute of Cognitive Science &
Department of Computer Science
University of Colorado Boulder
Boulder, CO, USA
sidney.dmello@colorado.edu

ABSTRACT

Modeling team phenomena from multiparty interactions inherently requires combining signals from multiple teammates, often by weighting strategies. Here, we explored the hypothesis that *strategic weighting* signals from individual teammates would outperform an *equal weighting* baseline. Accordingly, we explored role-, trait-, and behavior-based weighting of behavioral signals across team members. We analyzed data from 101 triads engaged in computer-mediated collaborative problem solving (CPS) in an educational physics game. We investigated the accuracy of machine-learned models trained on facial expressions, acoustic-prosodics, eye gaze, and task context information, computed one-minute prior to the end of a game level, at predicting success at solving that level. AUROCs for unimodal models that equally weighted features from the three teammates ranged from .54 to .67, whereas a combination of gaze, face, and task context features, achieved an AUROC of .73. The various multiparty weighting strategies did not outperform an equal-weighting baseline. However, our best nonverbal model (AUROC = .73) outperformed a language-based model (AUROC = .67), and there were some advantages to combining the two (AUROC = .75). Finally, models aimed at prospectively predicting performance on a minute-by-minute basis from the start of the level achieved a lower, but still above-chance, AUROC of .60. We discuss implications for multiparty modeling of team performance and other team constructs.

ICMI'20, October 25–29, 2020, Virtual Event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10...$15.00.
DOI: https://doi.org/10.1145/3382507.3418877

CCS CONCEPTS

Human-centered computing → Collaborative and social computing → Empirical studies in collaborative and social computing

KEYWORDS

Collaborative Problem Solving; Multimodal Multiparty Modeling

ACM Reference format:

Shree Krishna Subburaj, Angela E.B. Stewart, Arjun Ramesh Rao and Sidney K. D'Mello. 2020. Multimodal, Multiparty Modeling of Collaborative Problem Solving Performance. In *Proceedings of 2020 ACM International Conference on Multimodal Interaction (ICMI'20), October 25–29, Virtual Event, Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3382507.3418877

1 Introduction

Multiparty interactions are common in everyday life. For example, small group work is routine in classrooms, team meetings are commonplace in business, and virtual happy hours are increasingly prevalent in the age of social distancing. Although these activities might seem dissimilar, they all involve coordinated behaviors among multiple parties to achieve a desired goal. Here, we consider whether automated methods can be trained to predict collaborative outcomes (e.g., team performance or rapport), with applications to research, assessment, and intervention.

Just as *multimodal modeling* entails combining signals from various modalities, *multiparty* modeling of *team-level* outcomes

requires a mechanism to combine signals of individual teammates. One way to do this is to concatenate feature sets of each teammate by assigning them slots (e.g., features 1 to 4 for participant A, 5 to 8 for B, etc), but this raises the question of how to assign teammates to slots when roles are not pre-defined. Another approach is to develop predictive models for each teammate and integrate their predictions via decision-level fusion. However, this entails using individual features to model a group-level outcome.

An alternate approach is to weight the features of individual teammates, which requires making assumptions about team interaction. For example, signals from multiple teammates might be considered to be equal, which would support averaging signals across the team (e.g. [17]). However, this assumes that all team members are in fact equal in terms of their goals, behaviors, and contributions, which is often not the case [29,31,32]. Another option is to pool occurrences of behaviors regardless of team member. For example, [41] combined language products across the team independent of speaker. However, this approach implicitly gives more weight to the more verbose teammate, which may or may not be warranted depending on the collaboration context (e.g., collaborative or competitive).

Consequently, we hypothesize that team-level outcomes are influenced by complex group interactions, where the behaviors of all teammates do not contribute equally toward the outcome. Thus, we explored the hypothesis that *strategic weighting* signals from individual teammates would outperform an *equal weighting* baseline.

We conduct our research in the context of collaborative problem solving (CPS), using mixed-media videoconferencing (Zoom). CPS occurs when two or more people engage in a coordinated attempt to construct a solution to a problem [43,47]. We specifically model CPS task performance, which is a common objective CPS outcome [40,41,61], though our methods could be applied to other outcomes as well, such as shared knowledge building [11] or rapport [53].

We leverage multimodal data of 101 triads collaboratively playing Physics Playground, an educational game [58]. We focus on nonverbal signals, which have been shown to index constructs important for communication and team functioning [2,9,33]. Further, we focus on nonverbal signals, as they will likely generalize better than linguistic information, which encodes information specific to the task at hand [57]. In particular, we include facial expressions and acoustic-prosodic information to index emotional states [16,18], eye gaze as a measure of social visual attention [46], and high-level task context information as it provides insight into the unfolding problem solving process [56]. Our *central goal* is to investigate the effectiveness of various multiparty weighting schemes in predicting team-level performance from nonverbal signals of individual team members to adjudicate the strategic- vs. equal- weighting hypotheses.

1.1 Related Work

The existing research on group behavior and outcomes is vast [39,49,51]. Group dynamics and collaborative constructs, such as task performance [61], focus of attention [23], agreeableness [37], and so on, have been modeled from behavioral cues like head pose [44], eye gaze [60], acoustics and prosody [40,41], and language [21]. Several studies have focused on analyzing behavior of teams in online game environments [4,34] and how team behavior relates to task performance [34,36]. We scope our review to methods for combining behavioral features in multiparty scenarios as this is most relevant to our research goals.

Multimodal signals of teammates have been combined using feature-level fusion in convolutional neural networks [44] or by calculating aggregate statistics of behaviors across the group (mean, range, standard deviation) [40,54]. For example, Miura and Okada [40] used utterance counts, speaking length, head movement, acoustic-prosodic cues, and language features to predict expert-rated metrics of quality of group discussions as a measure of task performance. They obtained team-level features by first averaging features across each teammate's utterances and then averaging across the team, effectively weighting the teammates' contributions equally. The researchers also computed descriptives (maximum, standard deviation) of each teammate's features and used these as additional features to capture teammate-level influences better than the equally-weighted features alone. Their best-performing model achieved a Spearman correlation of .76 between predicted and expert coded group performance scores.

Related, Murray and Oertel [41] predicted CPS task performance from acoustic-prosodic and language features. Data was combined across the team by calculating features from pooled utterances regardless of speaker, effectively weighting more verbose speakers higher. Their best performing models achieved a mean squared error of 64.4, compared to a baseline of 79.3.

Task context features are commonly used in multiparty modeling as they are inherently at the team-level. Vrzakova et. al., [59] extracted change (and lack thereof) in areas of interest on the screen as high-order measures of a CPS programming task (e.g., idling, generating code, executing code). They also considered face/body movements and speech rate and combined these across teammates at the feature-level. They found that unimodal patterns of screen activity change were correlated with task scores, and that, while some combinations of modalities improved the correlation, others reduced or even eliminated it.

Finally, an individual's role on the team has long been considered in multiparty modeling, specifically in emergent leader identification tasks, which aim to predict dominant teammates [5,27,50]. Avici and Aran [3] leveraged teammate dominance when predicting group performance with SVMs and coupled Hidden Markov Chain models. Their model used a dominance score along with audio, video, and gaze features to achieve an accuracy of .91, beating a baseline of .60.

1.2 Contribution, Novelty, & Research Questions

Multiparty models of team-level outcomes require a mechanism to combine signals from individual teammates. Existing research in modeling multiparty outcomes makes inherent assumptions about each teammate's contribution to the task by either equally weighting or weighting dominant members more [17,26,40,41,56]. Here, we investigated the strategic weighting hypothesis by

exploring three strategic weighting strategies, two static and one behavior-based. The first strategy *statically* weights multimodal signals based on a team member's assigned role due to inherent differences in affordances offered by *assigned role* (contributor vs. collaborator). Additionally, we investigate static weighting based on *individual differences*, including prior experience, personality and attitudes towards teamwork, which have been theoretically and empirically linked to group outcomes [35]. For example, personality predicts collaboration quality [55] and prior domain experience predicts task performance [42]. We also investigate methods weighting signals based on behaviors that occur during team interaction, such as verbosity, eye gaze, facial expressions, which index emotional states, social visual attention and other constructs important for communication and team functioning (e.g. a team member's dominance in terms of verbal contributions or level of joint attention with teammates). To our knowledge, this is the first study to jointly investigate methods for combining multimodal signals from multiple participants.

We address four research questions.

RQ1. *How accurately do unimodal and multimodal non-verbal behavioral signals predict task performance*? Given that verbal contributions are principal in the context of CPS, it is an open question as to whether nonverbal signals can yield models of sufficient accuracy in predicting task performance. This is an essential first step before we can examine multiparty methodologies for combining signals. In the present study, we examine whether nonverbal signals 60s prior to the end of a game level predict success (i.e., whether or not they solved the level).

RQ2: *How do role-based, individual difference-based, and behavior-based approaches to combining multiparty features compare to an equal-weighting baseline?* Here, we explore different approaches for combining signals across the team and compare them to an equally weighting baseline.

RQ3: *How do nonverbal and language models compare and does combining the two improve performance?* Language models utilize utterances that are specific to the task, which may result in improved prediction accuracies at the cost of generalizability.

RQ4. *How accurately can we prospectively predict task performance*? Here we examine how our models can be used in real-time to support effective team performance, for example, by providing dynamic interventions. Whereas the models used to address the first three RQs focused on multimodal signals 60s before the end of a trial, for RQ4, we tested models that analyze the data in 60s intervals from the start of the trial. These models are expected to have lower accuracy due to the greater temporal distance between the signals and outcomes, but the pertinent question is whether accuracy is above chance.

2 Data Collection

The data were collected as part of a larger study on CPS [58]; only details pertinent to the present study are reported here.

2.1 Participants

Participants were 303 students (56% female, average age = 22 years) from two large public universities (38.5% from University 1). Based on self-reported demographics, participants were 47% Caucasian, 28% Hispanic/Latino, 18% Asian, 2% Black or African American, 1% American Indian or Alaska Native, and 4% reported "other". Students were assigned to 101 teams of three based on scheduling constraints. Thirty students from 18 teams (26%) indicated they knew at least one person from their team prior to participation. Participants were compensated monetarily with a $50 Amazon gift card (95.8%) or with course credit (4.2%).

2.2 Physics Playground

We used Physics Playground (Figure 1) as our problem-solving environment. Physics Playground is a two-dimensional educational game that aims to teach students basic Newtonian physics concepts (e.g., Newton's laws, energy transfer, and properties of torque) through gameplay [1,62]. Students complete levels by using mouse input to draw simple machines (ramps, levers, pendulums, and springboards) that guide a green ball to a red balloon. All objects in the game (including ones that students draw) obey the laws of physics. A trophy is earned upon completion of a level, which involves navigating a ball to a red balloon via the simple machines and other objects (e.g., weights). Students may choose to restart, exit, or change levels at any time during gameplay. The game is organized into multiple playgrounds (a playground has several levels) and students are free to navigate the levels as they please. No hints or other support mechanisms were provided to students, with the exception of a tutorial on game mechanics, which could be viewed at any time.

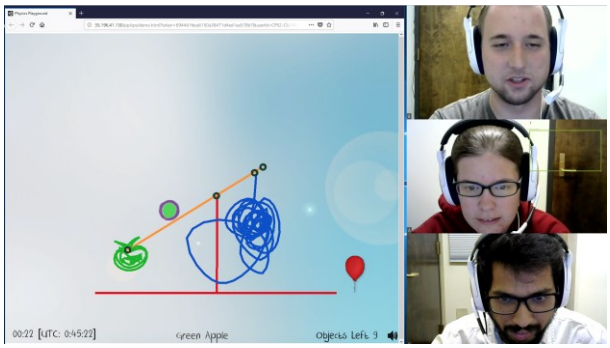

Figure 1: A triad using a lever and weight to navigate the green ball to the red balloon in the Physics Playground environment.

2.3 At-home Surveys

At least 24 hours prior to their scheduled lab session, students were emailed a survey that assessed a variety of individual difference measures, such as demographics, academic background, and personality. Students self-reported their formal *physics coursework* (e.g. none, high school, introductory college courses, or multiple college courses). We also used a validated measure of *physics self-efficacy* [65] to assess personal beliefs in their ability to succeed in physics. We used the validated short version of the Big Five inventory [22] to assess *personality dimensions* of extraversion, agreeableness, conscientiousness,

emotional stability, and openness to experience. We assessed *leadership self-efficacy* (self-belief of leadership capability), with the Leadership Domain Identification Measure [25] and *collectivism* (willingness to work in teams) and *teamwork self-efficacy* (personal perception of ability to work in teams) with the Individual Satisfaction with the Team Scale [35]. Leadership and teamwork self-efficacy were correlated (Pearson's r = .57), so we combined them by z-scoring each, and averaging the z-scores. Finally, students completed an expert-created ten-item physics pretest that assessed knowledge of two focal physics concepts used in gameplay: energy transfer and properties of torque.

After completing the survey measures, students completed a short tutorial on how to use Physics Playground, such as how to draw simple machines like ramps and springboards. After completing the tutorial, students were given 15 minutes to complete five easy levels to familiarize themselves with the game.

2.4 In-lab Procedure

The in-lab session took place after the at-home surveys were completed. Students were individually assigned to one of three computer-enabled workstations that were either partitioned in the same room or located in different rooms depending on the school where data was collected. There were no face-to-face interactions. All computers were equipped with a webcam and headset for video conferencing and screen sharing via Zoom (https://zoom.us). Each computer was also equipped with a separate webcam to record video of the participant's face and upper body at 10 frames per second. A separate audio stream (48000 Hz) was recorded for each participant using the same headset used for Zoom. Additionally, each computer was equipped with a Tobii 4C eye gaze tracker, which recorded eye gaze at a variable sampling rate, up to 90 Hz.

Teams collaborated in Physics Playground for three 15-minute blocks (total of 45-minutes of collaborative gameplay). For each block, one randomly assigned teammate was given the role of controller and the other two were assigned as observers. The controller was in charge of all mouse interaction with Physics Playground, while the observers viewed the controller's screen (through screen share) and were tasked with contributing to the solution and gameplay. A different teammate was the controller for each block, such that each student controlled the interaction for one block. An on-screen warning was displayed when there was ten and five minutes left in each block.

3 Data Processing and Machine Learning

We developed machine-learned models to predict whether or not a team successfully earned a trophy on a given level attempt (trial) from a combination of nonverbal signals.

3.1 Level Attempt Segmentation

Level attempts were segmented from the Physics Playground logs, which keep track of when a team enters a given level or earns a trophy. An attempt begins when the team enters a level and ends under three possible conditions: 1) the team earns a trophy; 2) the team begins a different level; or 3) time runs out in the block. We pooled across experimental blocks, resulting in 1220 total level attempts, of which 55.4% yielded a trophy (the positive class).

3.2 Feature Processing

We computed features for facial expression, gaze, acoustic-prosodic, and task context. Because sampling rate varied per modality, we aggregated features across non-overlapping 1s windows, which was deemed an appropriate unit of analysis since conversational turns, defined as spoken utterances, were in this range (median of 1.4s). The averaging also served as a smoothing filter. Given our relatively small number of instances, we strategically selected a small number of features per modality.

Facial Features. Facial expressions have been shown to index emotion in communication [16], and have been linked to task performance [40,41]. We used the videos of teammate's faces, sampled at 10 Hz, to extract face features for each frame in the video. We used Emotient [38], which produces a binary value for whether the face could be tracked in a given frame, estimates of face width and height in the frame, and likelihood estimates of the presence of 20 action units [16]. Width and height were converted to *face area*, which served as a proxy for how close a teammate's face was to the screen, a proxy for engagement [14]. We used the action unit estimates to compute *positive* and *negative valence* according to mappings in [13]. We also computed *expressivity*, a measure of the overall activity in a given frame as the mean value across the action units. Finally, we computed *face/upper body motion*, a measure of arousal [14], using a validated motion estimation algorithm [7]. We mean aggregated these frame-level features across 1s windows. All features were then z-scored per individual and then per block, to account for differences.

Gaze Features. Eye gaze provides information into social visual attention [52], an important aspect of CPS. We computed gaze features using raw data from the Tobii4C. First, we computed the proportion of samples in a given second where both eyes were successfully tracked (*validity*). We then computed *fixation dispersion* as the mean Euclidean distance of each raw gaze point in a 1s window from the centroid. We computed fixations, which are points where gaze is maintained on a location (maximum of 25 pixels apart) for at least 50ms [15]. Fixations longer than 1s were trimmed to 1s whereas fixations that overlapped second boundaries were assigned to the majority second. Fixations were then aggregated over 1s windows by computing the *number of fixations* and *mean fixation duration* in that second. We also computed *mean saccade amplitude*, which is the pixel distance of eye movement between fixations, across the 1s window. These measures provide a broad index into the spatial and temporal patterns of visual attention and index a variety of cognitive states, such as cognitive load, mind wandering, and distraction [20]. Finally, we computed the *mean distance between the centroid of each teammate's gaze and their two partners* (in a given second), which is a proxy for joint attention [60].

Acoustic-Prosodic features. Acoustic-prosodic information is important in modeling task performance as they index emotional states in communication and conversational dynamics [16]. We used the individual audio files to extract acoustic-

prosodic features over 10ms windows. We used OpenSmile [19] to compute *fundamental frequency* (pitch), *loudness* (energy), *center frequency and amplitude of the first through third formants*, *harmonics to noise ratio*, *jitter*, and *shimmer*. Using this small set of features, we accounted for the acoustic-prosodic information without representing spoken content. These features were mean aggregated over 1s windows. Similar to facial features, acoustic-prosodic features were z-scored per individual and by block to account for individual- and block-level differences.

Task-Context features. Task context is crucial in modeling task performance as they provide insights into the unfolding problem-solving process [56]. We extracted high-level task context features that represent overall patterns of interaction with the CPS environment, without encoding task-specific information (for generalizability). Specifically, we extracted (on a frame-by-frame basis) *changes on the Physics Playground (PP) area* of the screen (see Figure 1) using the same motion estimation algorithm used for face motion. This was then aggregated across 1s windows by taking the mean. This provides a measure of screen activity to distinguish moments of deliberation (or silence) from action. We computed *time spent on level attempt* as the difference between the start of the interval that is selected (as discussed in Section 3.3) and the start of the level attempt. Finally, we computed *visit number* (see Section 3.1) as the attempt number for a level.

3.3 Instance Creation

We created an instance for each level attempt. First, we adjusted the end time of the level-attempt by 10s in order to avoid peeking into the behaviors that occurred when the team wins the trophy or quits the level (e.g. celebration of level completion or clicking to exit the level). We deemed 10s to provide a sufficient buffer by analyzing level attempts from 20 teams at random. End times were not adjusted if time ran out in the block. We then selected data from 60s prior to the end time. We chose a 60s interval after comparing model performance (Section 4) on intervals of different sizes (15, 30, 45, 60, 90s), which all yielded similar performance. Level attempts under 60s (431 instances out of 1220) were not considered germane for modeling purposes since they reflect exceedingly easy levels or cases where the team entered a level and immediately exited. This resulted in 789 instances, of which 53.2% were successfully solved.

Aggregation & missing data treatment. In order to yield a single value for each feature and each teammate for a level attempt, we aggregated the feature values by taking the mean over the 60s interval (e.g. the mean of expressivity over the 60s interval for a given individual). If the data was missing for a given 60s interval, we applied mean-imputation by replacing the missing value with the mean value of that feature for the block. This strategy has been previously used for up to 10% of missing data [10]. In our data, we mean-imputed 1.4% and 3.2% of the total data for the facial expression, and gaze modalities, respectively (mean-imputed instances for the other modalities were negligible). We applied zero-imputation (replacing a feature value with zero) for the fewer than 1% of cases when the data for all three teammates was missing over the block. Task context features (excluding PP screen motion) were computed at the interval-level and therefore were not aggregated.

3.4 Multiparty weighting strategies

We explored a variety of weighting strategies to aggregate multimodal features across the triad. Each strategy is a weighted mean with different methods for weighing individual teammates. The strategies do not apply to task context features as they are inherently at the team-level so no aggregation is needed. If the data for weighting each teammate was missing, we discarded that instance. This resulted in a dataset ranging from 718 instances (for individual difference weighting since some teammates did not complete the at-home survey) to 789 instances (complete dataset).

Role-based weighting (static). The different role-based weighting strategies were as follows: *Equally weighted* - Each teammate is weighted equally; *Controller only* - Consider only the controller's signals (observers weights are set to 0); *Observers only* - Consider only the observer's signals and weight them equally; *Controller + Observers* - Include controller's features and separately include the mean of the observer's features (doubles the number of features). *Controller weighted twice* - Weight the controller's signals twice as much as the observers.

Individual difference-based weighting (static). We weighted teammates based on the following individual difference measures: *attitude towards teammates* (collectivism, leadership and teamwork self-efficacy), *prior physics experience* (prior physics courses, pretest score, physics self-efficacy), and the *Big Five Inventory personality measures* (extraversion, agreeableness, conscientiousness, emotional stability and openness). Specifically, for a given individual difference measure value m, and teammates A, B and C, the weight w_A, w_B, w_C applied to their feature values would be as follows ($w_A = m_A/(m_A + m_B + m_C)$). For example, if A, B, and C have an extraversion value of 1, 2, and 3, respectively, then $m_A = 1, m_B = 2, m_C = 3$ and all features are weighted (w_A, w_B, w_C) as $\frac{1}{6}, \frac{1}{3}, \frac{1}{2}$ for A, B, and C, respectively.

Behavior-based weighting. We weighted each teammate's features based on how they displayed certain behaviors during the collaboration as opposed to the previous static weights. The weights are computed by first averaging the signal over the 60s interval for each teammate and then averaging across teammates. We weighted based on the following features: *verbosity* - total words spoken over the 60s interval; mean *loudness* for the 60s interval; mean *expressiveness* for the 60s interval; *partner's distance* - how far a teammate's gaze is from the other two teammates. Unlike the other three measures, lower gaze distance suggests coordination, so we inversely weighted the distance as $1 - w$. Thus, higher verbosity, louder voices, greater facial expressivity, and more gaze coordination are weighted higher.

3.5 Supervised Classification and Validation

We considered Random Forest classifiers (from sci-kit learn [45]) as our primary machine learning model to predict successful or unsuccessful level attempts. We used team-level nested five-fold cross validation, where all level attempts of a team were either in the training set or testing set, but never both. Within each of the

five iterations, the training set was split into three folds for hyperparameter tuning, where we used grid search to tune the number of trees in the forest (150, 200, 250, 350, 400 or 500) and maximum depth of the trees (no maximum depth, 10, 20 or 50).

4 Results

We used AUROC as the performance metric for the main results and hyperparameter tuning as it assesses true positive and false positive tradeoff across prediction thresholds [24].

4.1 (RQ1) Accuracy of non-verbal models

Our first research question investigates feasibility of using nonverbal signals to predict task performance. These analyses used an equal weighting scheme where the mean of each feature computed across the three teammates was used for modeling. We compared the accuracy of our models trained on different unimodal and multimodal behavioral signals with chance (AUROC = .50), as well as a shuffled baseline, where we randomly shuffled outcomes within a team. This essentially eliminated temporal dependencies between the behavioral signals and outcomes, while preserving concurrent behavioral signals and outcome base rates within a team.

As indicated in Figure 2, the face, gaze, and task context models consistently outperformed the shuffled baseline (AUROC = .56), whereas the acoustic-prosodic model did not. Importantly, a feature-level fusion of these four modalities yielded an additive effect in that the combined model (AUROC = .71) outperformed the best unimodal model (task context with an AUROC of .67).

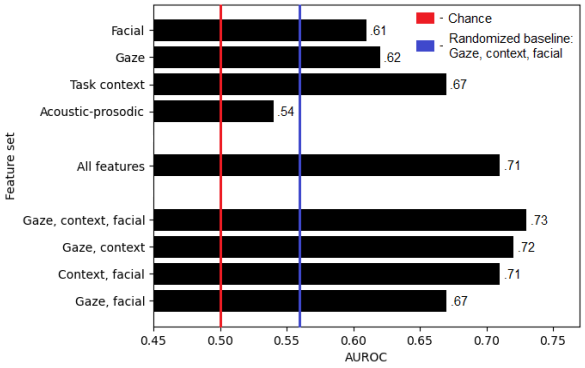

Figure 2: Performance comparison of model trained on different features sets and their multimodal combinations.

In order to identify the most informative modalities, we trained additional models that predicted task performance from various combinations of the three most predictive feature sets: gaze, task context and facial expressions. The combination of all three modalities achieved the best performance (AUROC = .73) which was higher than that achieved by including acoustic-prosodic features (AUROC = .71). Whereas the gaze-context and context-face models achieved similar accuracy (AUROCs of .72 and .71), accuracy of the gaze and facial expression model was lower (AUROC = .67), suggesting the importance of considering top-down context information with bottom-up gaze features.

We also experimented with decision-level fusion to combine modalities. Compared to feature-level fusion, when all four modalities were combined, decision-level fusion yielded a slight performance improvement (AUROC .72 vs. 71). However, the performance was the same (AUROC .73) for the best performing modalities (gaze, task context and facial expression).

Finally, in an attempt to understand which of the individual features in the modalities were most predictive of task performance, we investigated Random Forest importance scores [8] for each feature in the face + gaze+ task context model. The time spent on a level attempt and PP motion (both task context features) and overall body movement were most predictive (.11, .09, .10 importance). The other features excluding visit number (least predictive with .02 importance) had moderate (.06-.08) importance. Thus, a multimodal combination was essential.

4.2 (RQ2) Comparison of weighting schemes

To address this question, we explored approaches to multiparty weighting by using our best performing feature set: a feature-level fusion of gaze, task context and facial expressions. We compare our static and behavior-based multiparty weighting strategies to a baseline of equally weighting teammates (Section 4.1, AUROC = .73). As illustrated in Figure 3, all of the role-based weighting strategies performed slightly worse (average of -2.75%) than the baseline model. Interestingly, the most notable decrease in accuracy occurred when considering only the controller's or observers' signals (decrease of 4.1% from baseline), compared to other role-based models (average decrease of 1.4% from baseline). This suggests that task performance is best predicted when behaviors of teammates with multiple roles are considered. That said, all role-based models were still moderately accurate (AUROC > .70), even when omitting signals from one or more teammates (e.g. controller only or observers only).

We also examined weighting based on three individual difference measures (attitudes towards teamwork, prior physics experience, and personality, see Section 2.3). We found that none of the individual difference-based weightings outperformed the baseline (AUROCs of .69-.73), but there were also no notable decrements apart from weighting based on prior physics coursework (AUROC = .69). This result is consistent with findings from similar research analyzing task performance [48] in that difference measures did not influence task performance.

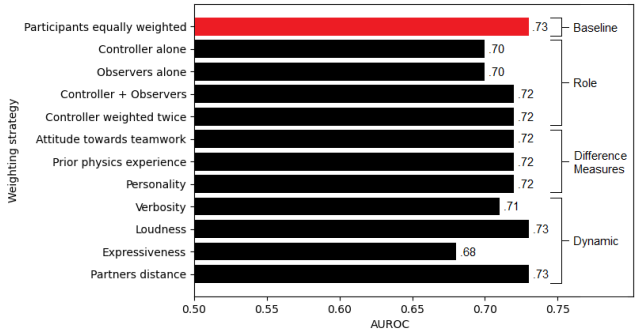

Figure 3: Comparison of multiparty weighting strategies on models trained on facial, gaze, context features.

Finally, we examined weighting based on behavior that occurred during the 60s interval. Weighting based on facial expressiveness yielded lower performance than the other behavior-based strategies (on average 6% lower). Weighting based on loudness, and inversely weighting based on gaze distance from partners performed slightly better (on average 3.4% better) than other behavior-based weighting strategies, but were similar to the baseline, suggesting no added value to behavior-based weighting.

4.3 (RQ3) Comparing nonverbal and language-based models

We compare our nonverbal models to language-based models, which directly index the content of the collaboration (i.e. the task was primarily verbal). We use the gaze, task context and facial expression models with teammates weighted equally as our comparison as it consistently outperformed the others (Section 4.2). To construct the language models, we first obtained automated transcriptions from IBM Watson [64]. Each word was assigned to the second in which it occurred, and words that overlapped seconds were assigned to the second in which they started. We concatenated utterances across all speakers in the 60s interval, from which we extracted n-gram [28] counts using NLTK's tokenizer [6]. In addition, we computed verbosity as the total number of words spoken and inter-turn duration as the mean duration between utterances. We also computed the proportion of utterances that overlapped across speakers.

We trained a Random Forest language model on these features and tuned two hyperparameters in addition to those listed in Section 3.5. First, we tuned whether to use unigrams, bigrams, tri-grams, or a combination (e.g., unigrams & bigrams). Second, we tuned pointwise mutual information (from 2 to 4) [12], which is used to filter and include only relevant n-grams. The language-based model yielded an AUROC of .67, which was lower than the AUROC of .73 obtained from the nonverbal model.

We also experimented with a transfer learning approach using state-of-the-art pre-trained Bidirectional Encoder Representations from Transformers (BERT) model [30] from [63]. This model was fine-tuned on utterances from our dataset over four epochs with a batch size of 32 and sequences were padded or truncated to have a fixed length of 300 words (hyper-parameters were chosen based on recommendations while fine-tuning BERT [30]). We found that the Random Forest model trained with language features alone outperformed the equivalent BERT model by 24% (AUROCs of .67 and .54 for Random Forest and BERT respectively).

Finally, we combined language along with gaze, task context and facial features using feature-level fusion to predict task performance. The multimodal feature-level combination of all four models yielded an AUROC of .75, reflecting a small improvement over the nonverbal models (AUROC = .73).

4.4 (RQ4) Prospectively predicting performance

In a real-time application, we do not have information on the timing of the end of a level attempt. Therefore, we cannot model task performance with the final 60s of data as in previous models. Therefore, we trained models analogous to a real-time application. We split level attempts that were at least 60s long (789 out of 1220) into 60s intervals from the start of the attempt. We considered each 60s interval as an instance (3156 instances) and aggregated gaze, task context and facial features across those 60s intervals using the equal weighting scheme. We labeled instance outcomes as the final outcome of the level attempt (i.e., whether or not a trophy was earned for that level attempt). This resulted in 44.2% of the instances being labelled as positive. The Random Forest model trained on these data using similar settings as above yielded an AUROC of .60, which was lower than the AUROC of .73 obtained when the final 60s interval were alone considered (Section 4.2). However, this could be expected since we are analyzing data early into the level to prospectively predict outcomes several minutes later (level attempts ranged from 1-15 mins with a mean of 4.6-min and a standard deviation of 3.4-min).

5 Discussion

We contrasted a strategic- with an equal- weighting alternative in combining behavioral features from three teammates to predict team-level performance on a CPS task. Specifically, we examined role-, trait-, and behavior-based strategies for combining facial expression, acoustic-prosodic eye gaze, and task context across the teammates, and also compared multimodal to unimodal and language-based approaches.

5.1 Main Findings

We found that a multimodal combination of gaze, task context and facial features best predicted task performance compared to other combinations of features including models that used all features and unimodal models. Task context features provide information on patterns of interaction in the collaboration environment, such as time spent on the level and high-level behaviors as defined by changes to the screen. It is important to note that we intentionally used a restricted set of three task context features that are likely applicable in other CPS environments. Further, gains can be expected in model accuracy as features more specific to the task (e.g., difficulty of individual levels, specific problem-solving strategies) are considered.

In addition to task context, eye gaze indexes attention to the task as well as coordination amongst teammates [51], which are presumably important for task performance. Facial expressions index emotional states [16] that are critical for communication and team functioning, and are therefore important cues into task performance. Thus, a combination of behavioral signals that index different constructs (attention, interaction patterns, emotion) work best for predicting task performance.

We compared role-, trait-, and behavior-based strategies for combining individual signals across the team. All aggregation strategies yielded similar performance to a baseline where all teammate's signals were weighted equally. It might be the case that teams form collaborative patterns independent of pre-existing traits or specified roles on the team. This is in line with previous research [58] that found that team makeup measures did not predict task performance (though they did predict subjective

perceptions of the collaboration). We also examined behavior-based weighting using verbosity, loudness, expressiveness and distance from partners' gaze, to account for changes in teammates' behaviors during the task. However, this approach also failed to outperform the baseline. Further research is needed to investigate the conditions when these (or alternate) weighting strategies might outperform equal-weighting.

We focused on nonverbal models because they likely generalize better to new contexts than language models which can focus on verbal cues specific to the task [57]. Surprisingly, our nonverbal models that combined facial expression, gaze, and task context performed better than language-only models (9% increase in AUROC). That said, a combination of language and nonverbal features yielded slightly better (2.7%) performance overall, suggesting some merit to combining the two within a single task.

We also found that the accuracy of models that predict task performance a minute prior to the end of the level was moderate (AUROC of .73 without language; .75 with language) and higher than AUROCs for models that predicted accuracy prospectively on a minute-by-minute basis from the start of the level (AUROCs of .60). The lower, but above-chance, accuracies for the latter models are expected since the models are essentially predicting further into the future (a challenging task no doubt).

It is also notable that the role-based weighting approaches yielded reasonable accuracies (AUROCs of .70) even when we considered signals for a subset of teammates (e.g., controller or observers only), suggesting that models can be robust even when entire signals are missing from someone on the team.

5.2 Applications

Our approach is applicable to other CPS contexts, though the models would need to be retrained based on data collected in new contexts. Further, although we focused on modeling task performance, our approach can be used to model other team-level constructs including team cohesion, rapport, or subjective perceptions of the collaboration. It is likely that precisely what multiparty combination strategy will be effective might depend on the construct being modeled and thus the weighting strategy should be adjusted accordingly.

In the particular case of modeling CPS performance, our model could be used to monitor remote virtual collaborations. A real-time system could intervene when the model predicts a low likelihood of success at the task. For example, an intelligent CPS interface could provide a hint or suggest an alternate strategy when the current approach seems unlikely to yield success. The CPS task performance model could also be combined with related models that monitor effective CPS processes, such as construction of shared knowledge, negotiation/coordination, and maintaining of team function [57], such that interventions could be sensitive to both CPS processes and outcomes.

5.3 Limitations and Future Work

Our study has limitations that must be addressed in the future. First, we chose to use nonverbal behavioral signals because they are ostensibly generalizable to different tasks (compared to language-based models). However, we did not test this hypothesis and only used data from a single task. Additionally, we focused on a single team-level construct (task performance), but did not investigate how the various multiparty weighting strategies might affect modeling of other constructs (e.g. turn-taking, team cohesion, subjective perceptions).

Second, we only included standard classifiers in the present work. We did conduct preliminary investigations with deep learning architectures (LSTMs in our case) using an equal weighting strategy. Whereas these models did not outperform the present Random Forest models (and are not reported here), this could be due to the fact that the deep learning models were not fully optimized and there was insufficient training data. It is likely that a comprehensive exploration of appropriate deep learning model architectures and inclusion of multiparty weighting in these architectures could result in improvements in model accuracy. Expanding the dataset by including additional and more diverse teams is also warranted in future work.

Third, our study was conducted in a controlled lab environment. We can expect patterns to change as models are deployed in real-world scenarios, where the signals are noisier and interaction dynamics might differ. We are in the process of analyzing an additional CPS dataset in real-world classrooms, which we will use to test the robustness of our models.

Fourth, we considered simple weighting strategies that only used singular metrics (e.g., expressiveness, extraversion). Perhaps a sophisticated weighting approach based on a combination of these metrics might improve model accuracy.

Fifth, although we explored a large set of nonverbal signals (facial expression, eye gaze, acoustics and prosody, task context), we did not include measures of physiological arousal. Therefore, future work can incorporate physiology signals, such as electrodermal response, which might improve model accuracy. Additionally, since coordination is an important aspect of collaboration, incorporating additional measures of coordination, for example expression mirroring, beyond gaze coordination considered here is an item for future work.

5.4 Concluding remarks

We investigated methods for combining nonverbal signals from multiple teammates during multiparty interactions in the context of triadic collaborative problem solving. We were moderately successful in predicting team performance using multimodal nonverbal signals and found that the simplest strategy of equally weighting signals of all teammates yielded the best performance. Our results have implications for the emergent field of multimodal, multiparty modeling and interaction.

ACKNOWLEDGMENTS

We thank our collaborators at ASU (Nick Duran and his team) and FSU (Valerie Shute and her team). This research was supported by the National Science Foundation (DUE 1745442, IIS 1921087, SES 1928612, SES 2030599) and the Institute of Educational Sciences (IES R305A170432). Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] Juan Miguel L. Andres, Ma Mercedes T. Rodrigo, Jessica O. Sugay, Ryan S. Baker, Luc Paquette, Valerie J. Shute, Matthew Ventura, and Matthew Small. 2014. An exploratory analysis of confusion among students using Newton's playground. In Proceedings of the 22nd International Conference on Computers in Education, ICCE 2014.

[2] Kathleen T Ashenfelter. 2008. Simultaneous analysis of verbal and nonverbal data during conversation: Symmetry and turn-taking. Dissertation Abstracts International, B: Sciences and Engineering.

[3] Umut Avci and Oya Aran. 2016. Predicting the Performance in Decision-Making Tasks: From Individual Cues to Group Interaction. IEEE Transactions on Multimedia 18, 4: 643–658. https://doi.org/10.1109/TMM.2016.2521348

[4] Aaron Bauer and Zoran Popović. 2017. Collaborative problem solving in an open-ended scientific discovery game. Proceedings of the ACM on Human-Computer Interaction 1, CSCW: 1–21. https://doi.org/10.1145/3134657

[5] Cigdem Beyan, Nicolò Carissimi, Francesca Capozzi, Sebastiano Vascon, Matteo Bustreo, Antonio Pierro, Cristina Becchio, and Vittorio Murino. 2016. Detecting emergent leader in a meeting environment using nonverbal visual features only. In ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction. https://doi.org/10.1145/2993148.2993175

[6] Steven Bird, Steven Bird, and Edward Loper. 2016. NLTK: The natural language toolkit NLTK: The Natural Language Toolkit. Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1. https://doi.org/10.3115/1225403.1225421

[7] Nigel Bosch, Sidney D'Mello, Ryan Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2015. Automatic detection of learning-centered affective states in the wild. In International Conference on Intelligent User Interfaces, Proceedings IUI. https://doi.org/10.1145/2678025.2701397

[8] Leo Breiman. 2001. Random forests. Machine Learning. https://doi.org/10.1023/A:1010933404324

[9] Emily A. Butler and Ashley K. Randall. 2013. Emotional coregulation in close relationships. Emotion Review. https://doi.org/10.1177/1754073912451630

[10] Jehanzeb R. Cheema. 2014. Some general guidelines for choosing missing data handling methods in educational research. Journal of Modern Applied Statistical Methods. https://doi.org/10.22237/jmasm/1414814520

[11] Prerna Chikersal, Maria Tomprou, Young Ji Kim, Anita Williams Woolley, and Laura Dabbish. 2017. Deep structures of collaboration: Physiological correlates of collective intelligence and group satisfaction. In Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW. https://doi.org/10.1145/2998181.2998250

[12] Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. https://doi.org/10.3115/981623.981633

[13] Jeffrey F. Cohn, Laszlo A. Jeni, Itir Onal Ertugrul, Donald Malone, Michael S. Okun, David Borton, and Wayne K. Goodman. 2018. Automated Affect Detection in Deep Brain Stimulation for Obsessive-Compulsive Disorder. https://doi.org/10.1145/3242969.3243023

[14] Sidney D'Mello and Art Graesser. 2009. Automatic detection of learner's affect from gross body language. Applied Artificial Intelligence. https://doi.org/10.1080/08839510802631745

[15] Edwin S. Dalmaijer, Sebastiaan Mathôt, and Stefan Van der Stigchel. 2014. PyGaze: an open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. Behavior research methods. https://doi.org/10.3758/s13428-013-0422-2

[16] Paul Ekman and Erika L. Rosenberg. 2012. What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). https://doi.org/10.1093/acprof:oso/9780195179644.001.0001

[17] Lucca Eloy, Angela E.B. Stewart, Mary J. Amon, Caroline Reindhardt, Amanda Michaels, Chen Sun, Valerie Shute, Nicholas D. Duran, and Sidney K. D'Mello. 2019. Modeling team-level multimodal dynamics during multiparty collaboration. In ICMI 2019 - Proceedings of the 2019 International Conference on Multimodal Interaction. https://doi.org/10.1145/3340555.3353748

[18] Florian Eyben, Klaus R. Scherer, Bjorn W. Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing. https://doi.org/10.1109/TAFFC.2015.2457417

[19] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In MM 2013 - Proceedings of the 2013 ACM Multimedia Conference. https://doi.org/10.1145/2502081.2502224

[20] Myrthe Faber, Kristina Krasich, Robert E. Bixler, James R. Brockmole, and Sidney K. D'Mello. 2020. The Eye-Mind Wandering Link: Identifying Gaze Indices of Mind Wandering Across Tasks. Journal of Experimental Psychology: Human Perception and Performance. https://doi.org/10.1037/xhp0000743

[21] Michael Flor, Su-Youn Yoon, Jiangang Hao, Lei Liu, and Alina von Davier. 2016. Automated classification of collaborative problem solving interactions in simulated science tasks. https://doi.org/10.18653/v1/w16-0504

[22] Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann. 2003. A very brief measure of the Big-Five personality domains. Journal of Research in Personality. https://doi.org/10.1016/S0092-6566(03)00046-1

[23] Carl Gutwin, Scott Bateman, Gaurav Arora, and Ashley Coveney. 2017. Looking away and catching up: Dealing with brief attentional disconnection in synchronous groupware. In Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW. https://doi.org/10.1145/2998181.2998226

[24] J. A. Hanley and B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. https://doi.org/10.1148/radiology.143.1.7063747

[25] Crystal L. Hoyt and Jim Blascovich. 2010. The role of leadership self-efficacy and stereotype activation on cardiovascular, behavioral and self-report responses in the leadership domain. Leadership Quarterly. https://doi.org/10.1016/j.leaqua.2009.10.007

[26] Hirofumi Inaguma, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2016. Prediction of ice-breaking between participants using prosodic features in the first meeting dialogue. In 2nd Workshop on Advancements in Social Signal Processing for Multimodal Interaction 2016, ASSP4MI 2016 - Held in conjunction with the 18th ACM International Conference on Multimodal Interaction 2016, ICMI 2016. https://doi.org/10.1145/3005467.3005472

[27] Dinesh Babu Jayagopi, Hayley Hung, Chuohao Yeo, and Daniel Gatica-Perez. 2009. Modeling dominance in group conversations using nonverbal activity cues. IEEE Transactions on Audio, Speech and Language Processing. https://doi.org/10.1109/TASL.2008.2008238

[28] Daniel Jurafsky and James Martin. 2014. Speech and Language Processing. In Speech and Language Processing. 83–120.

[29] Steven J. Karau and Kipling D. Williams. 1993. Social Loafing: A Meta-Analytic Review and Theoretical Integration. Journal of Personality and Social Psychology. https://doi.org/10.1037/0022-3514.65.4.681

[30] Ming-wei Chang Kenton, Lee Kristina, and Jacob Devlin. 2017. BERT paper. arXiv:1810.04805 [cs]. https://doi.org/arXiv:1811.03600v2

[31] Norbert L. Kerr. 1983. Motivation losses in small groups: A social dilemma analysis. Journal of Personality and Social Psychology. https://doi.org/10.1037/0022-3514.45.4.819

[32] Norbert L. Kerr and Steven E. Bruun. 1983. Dispensability of member effort and group motivation losses: Free-rider effects. Journal of Personality and Social Psychology. https://doi.org/10.1037/0022-3514.44.1.78

[33] Norbert L. Kerr and R. Scott Tindale. 2004. Group Performance and Decision Making. Annual Review of Psychology. https://doi.org/10.1146/annurev.psych.55.090902.142009

[34] Young Ji Kim, David Engel, Anita Williams Woolley, Jeffrey Yu Ting Lin, Naomi McArthur, and Thomas W. Malone. 2017. What makes a strong team? Using collective intelligence to predict team performance in League of Legends. In Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW. https://doi.org/10.1145/2998181.2998185

[35] José M. de la Torre-Ruiz, Vera Ferrón-Vílchez, and Natalia Ortiz-de-Mandojana. 2014. Team Decision Making and Individual Satisfaction With the Team. Small Group Research. https://doi.org/10.1177/1046496414525478

[36] Alex Leavitt, Brian C. Keegan, and Joshua Clark. 2016. Ping to win? Non-verbal communication and team performance in competitive online multiplayer games. Conference on Human Factors in Computing Systems - Proceedings: 4337–4350. https://doi.org/10.1145/2858036.2858132

[37] Rivka Levitan, Agustín Gravano, Laura Willson, Štefan Beňuš, Julia Hirschberg, and Ani Nenkova. 2012. Acoustic-prosodic entrainment and social behavior. In NAACL HLT 2012 - 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference.

[38] Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, and Marian Bartlett. 2011. The computer expression recognition toolbox (CERT). In 2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011. https://doi.org/10.1109/FG.2011.5771414

[39] Margaret M. McManus and Robert M. Aiken. 2016. Supporting Effective Collaboration: Using a Rearview Mirror to Look Forward. International Journal of Artificial Intelligence in Education. https://doi.org/10.1007/s40593-015-0068-6

[40] Go Miura and Shogo Okada. 2019. Task-independent multimodal prediction of group performance based on product dimensions. In ICMI 2019 - Proceedings of the 2019 International Conference on Multimodal Interaction. https://doi.org/10.1145/3340555.3353729

[41] Gabriel Murray and Catharine Oertel. 2018. Predicting group performance in task-based interaction. In ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction. https://doi.org/10.1145/3242969.3243027

[42] Timothy J. Nokes-Malach, Michelle L. Meade, and Daniel G. Morrow. 2012. The effect of expertise on collaborative problem solving. Thinking and Reasoning. https://doi.org/10.1080/13546783.2011.642206

[43] OECD. 2015. Pisa 2015 Draft Collaborative Problem Solving Framework March 2013. Oecd.

[44] Kazuhiro Otsuka, Keisuke Kasuga, and Martina Köhler. 2018. Estimating visual focus of attention in multiparty meetings using deep convolutional neural networks. In ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction. https://doi.org/10.1145/3242969.3242973

[45] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research.

[46] Daniel C. Richardson, Rick Dale, and Natasha Z. Kirkham. 2007. The Art of Conversation Is Coordination. Psychological Science. https://doi.org/10.1111/j.1467-9280.2007.01914.x

[47] Jeremy Roschelle and Stephanie D. Teasley. 1995. The Construction of Shared Knowledge in Collaborative Problem Solving. In Computer Supported Collaborative Learning. https://doi.org/10.1007/978-3-642-85098-1_5

[48] Yigal Rosen and Rikki Rimor. 2015. Teaching and assessing problem solving in online collaborative environment. In Professional Development and Workplace Learning: Concepts, Methodologies, Tools, and Applications. https://doi.org/10.4018/978-1-4666-8632-8.ch017

[49] Nikol Rummel and Hans Spada. 2005. Learning to collaborate: An instructional approach to promoting collaborative problem solving in computer-mediated settings. Journal of the Learning Sciences. https://doi.org/10.1207/s15327809jls1402_2

[50] Dairazalia Sanchez-Cortes, Oya Aran, Dinesh Babu Jayagopi, Marianne Schmid Mast, and Daniel Gatica-Perez. 2013. Emergent leaders through looking and speaking: From audio-visual data to multimodal recognition. Journal on Multimodal User Interfaces. https://doi.org/10.1007/s12193-012-0101-0

[51] Richard C. Schmidt and Michael J. Richardson. 2008. Dynamics of interpersonal coordination. Understanding Complex Systems. https://doi.org/10.1007/978-3-540-74479-5_14

[52] Shung J. Shin, Tae Yeol Kim, Jeong Yeon Lee, and Lin Bian. 2012. Cognitive team diversity and individual team member creativity: A cross-level interaction. Academy of Management Journal. https://doi.org/10.5465/amj.2010.0270

[53] Tanmay Sinha and Justine Cassell. 2015. We click, we align, we learn: Impact of influence and convergence processes on student learning and rapport building. In INTERPERSONAL 2015 - Proceedings of the 1st ACM Workshop on Modeling INTERPERsonal SynchrONy And infLuence, co-located with ICMI 2015. https://doi.org/10.1145/2823513.2823516

[54] Daniel Spikol, Emanuele Ruffaldi, Lorenzo Landolfi, and Mutlu Cukurova. 2017. Estimation of Success in Collaborative Learning Based on Multimodal Learning Analytics Features. Proceedings - IEEE 17th International Conference on Advanced Learning Technologies, ICALT 2017: 269–273. https://doi.org/10.1109/ICALT.2017.122

[55] Angela Stewart and Sidney K. D'Mello. 2018. Connecting the dots towards collaborative aied: linking group makeup to process to learning. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-319-93843-1_40

[56] Angela E.B. Stewart, Zachary A. Keirn, and Sidney K. D'Mello. 2018. Multimodal modeling of coordination and coregulation patterns in speech rate during triadic collaborative problem solving. In ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction. https://doi.org/10.1145/3242969.3242989

[57] Angela E.B. Stewart, Hana Vrzakova, Chen Sun, Jade Yonehiro, Cathlyn Adele Stone, Nicholas D. Duran, Valerie Shute, and Sidney K. D'Mello. 2019. I say, you say, we say: Using spoken language to model socio-cognitive processes during computer-supported collaborative problem solving. Proceedings of the ACM on Human-Computer Interaction. https://doi.org/10.1145/3359296

[58] Angela E B Stewart, Mary Jean Amon, Nicholas D Duran, and Sidney K D'Mello. 2020. Beyond Team Makeup: Diversity in Teams Predicts Valued Outcomes in Computer-Mediated Collaborations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20), 1–13. https://doi.org/10.1145/3313831.3376279

[59] Hana Vrzakova, Mary Jean Amon, Angela Stewart, Nicholas D. Duran, and Sidney K. D'Mello. 2020. Focused or stuck together: Multimodal patterns reveal triads' performance in collaborative problem solving. ACM International Conference Proceeding Series: 295–304. https://doi.org/10.1145/3375462.3375467

[60] Hana Vrzakova, Mary Jean Amon, Angela E.B. Stewart, and Sidney K. D'Mello. 2019. Dynamics of Visual Aention in Multiparty Collaborative Problem Solving using Multidimensional Recurrence antification Analysis. In Conference on Human Factors in Computing Systems - Proceedings. https://doi.org/10.1145/3290605.3300572

[61] Miaomiao Wen, Diyi Yang, and Carolyn Penstein Rosé. 2015. Virtual teams in massive open online courses. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-319-19773-9_124

[62] Jacqueline Kory Westlund, Sidney K. D'Mello, and Andrew M. Olney. 2015. Motion Tracker: Camera-based monitoring of bodily movements using motion silhouettes. PLoS ONE. https://doi.org/10.1371/journal.pone.0130293

[63] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing.

[64] IBM Watson speech-to-text. Retrieved June 1, 2020 from https://www.ibm.com/cloud/watson-speech-to-text

[65] 2011. Self-Efficacy of First Year University Physics Students: Do Gender and Prior Formal Instruction in Physics Matter? International Journal of Innovation in Science and Mathematics Education.