Automatic Detection of Collaborative States in Small Groups Using Multimodal Features

Mariah Bradford*, Ibrahim Khebour*, Nathaniel Blanchard, and Nikhil Krishnaswamy

Colorado State University, Fort Collins, CO 80523, USA
mbrad@colostate.edu

Abstract. Cultivating collaborative problem solving (CPS) skills in educational settings is critical in preparing students for the workforce. Monitoring and providing feedback to all groups is intractable for teachers in traditional classrooms but is potentially scalable with an AI agent who can observe and interact with groups. For this to be feasible, CPS moves need to first be detected, a difficult task even in constrained environments. In this paper, we detect CPS facets in relatively unconstrained contexts: an in-person group task where students freely move, interact, and manipulate physical objects. This is the first work to classify CPS in an unconstrained *shared* physical environment using multimodal features. Further, this lays the groundwork for employing such a solution in a classroom context, and establishes a foundation for integrating classroom agents into classrooms to assist with group work.

Keywords: Collaborative Problem Solving · Multimodal · Natural Language Processing · Small Groups

1 Introduction

Working in teams is an essential skill in the workforce, which the education system needs to prepare students for. Such practices have been formalized into pedagogical techniques of *collaborative problem solving* (CPS) wherein students learn by working together to achieve a shared goal. With CPS, peers can develop a "positive interdependence" [8], but doing so depends on having an effective group dynamic that does not fall into dysfunction. With proper facilitation, this can be avoided, but this role usually falls to the teacher, and with a single teacher and many small groups, such facilitation becomes intractable.

For some group-facilitation tasks, such as aligning group goals, an artificially intelligent agent can be a useful tool to help teachers manage groups. A prerequisite to agent interaction with a group is that the agent must be able to observe the group and detect its state. Sun *et al.* [13] recently proposed a novel coding scheme where each component of CPS could occur simultaneously, rather than being treated as distinct phenomena (e.g., social or cognitive indicators). They

* These authors contributed equally to this work.

define an alternative framework composed of three main facets: *construction of shared knowledge*, *negotiation/coordination*, and *maintaining team function*, which are defined by *indicators* (e.g., "proposes specific solutions"). In this work, we adopt the Sun *et al.* framework from [13] and use it to annotate a novel collaborative problem-solving task: the Weights Task (Section 2.1).

A number of works have trained machine learning models to detect CPS in controlled, virtual environments [1, 12]. We extend these findings, showing that CPS facets can still be automatically detected in in-person group work, despite the increased complexity. Several works have explored what features are important for detecting in-person CPS states [4, 5, 9], and we took inspiration from these works in the design of our study. Our work presents a novel physical, in-person shared task, and we utilized many of the recommended features to establish our baseline on this data. [1]

2 Methods

2.1 Data Collection

Weights Task We collected audiovisual data of small groups collaborating on an in-person, shared, physically-grounded problem-solving task, known as the Weights Task. An example still can be seen in Figure 1. In this task, triads are given five colored cubes of different weights, a balance scale, and a worksheet to track answers. We identify the weight of one block, and ask them to use the balance scale to identify the weights of the remaining blocks. When participants have identified the weights of all five blocks, we remove the balance scale and provide a new block of unknown weight to participants. They must then try to identify the weight of the mystery block. To successfully do this, they must infer the pattern in the block weights. They have two attempts. We then ask participants for the weight of a hypothetical next block in the sequence, according to the pattern. They again have two attempts. Recording ends at the end of this second attempt. A prior version of this task was described in [3] — our version extends those methods in several ways to elicit more collaborative moves.

Thirty participants were recruited for this study. All participants were over the age of 18, spoke fluent English, and were drawn from the student population of Colorado State University. Participants were 20% female and 80% male. When asked to identify their ethnicity, 60% of participants identified as Caucasian, 10% identified as Hispanic or Latino, and 30% identified as Asian. Participants indicated a range of native languages including English, Hindi, Assamese, Gujarati, Bengali, Telugu, Persian, Malayalam, Urdu, and Spanish. The full dataset consists of ten triads completing the Weights Task. Recordings average 16 minutes. Audio recording used an MXL AC-404 Procon microphone as advised by findings from [3]. Each audio recording was processed using Google's Voice Activity Detection (VAD) [11] to automatically segment audio files into utterances, with only one speaker per segment. Next, we transcribe the segmented audio files using Google's automatic speech recognition (ASR). After preprocessing, there were a total of 1,822 utterances, with an average length of 4.26 seconds.

[1] Supplemental material can be found here: https://github.com/Blanchard-lab/aied_2023_suppmat

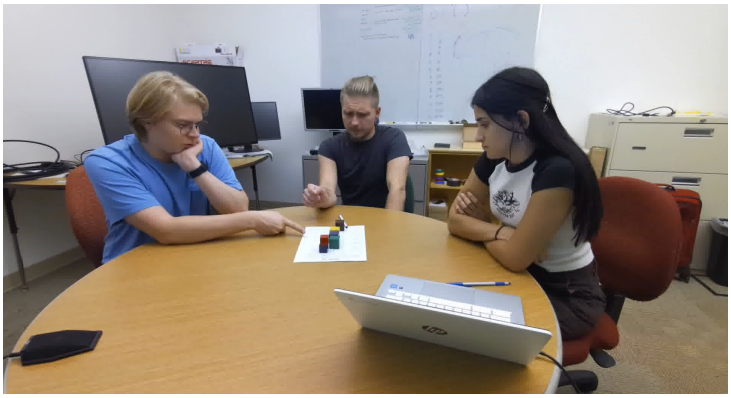

Fig. 1: The Weights Task

2.2 Annotations

Utterances from each group were annotated for collaborative problem solving (CPS) in accordance with the CPS framework developed by Sun *et al.* [13]. Each utterance was annotated by two trained coders. Table 1 shows the average number of occurrences of each facet per group.

Table 1: Descriptive statistics of all 1,822 utterances across all groups

	Average	SD	Min	Max
All utterances	182.20	80.51	90	380
# None	62.10	30.50	40	141
# Construction of shared knowledge	69.20	25.49	33	131
# Negotiation/Coordination	68.10	27.12	28	126
# Maintaining team function	37.10	23.59	9	73
Time (s)	4.26	2.84	0.84	23.64

2.3 Verbal Features

Verbal features comprised features corresponding to the words in each utterance as transcribed by Google's ASR. Each group's utterance-level transcripts were preprocessed for formatting (including removing newlines and periods and surrounding the utterance with BERT's required `[CLS]` and `[SEP]` tokens), and then fed into the BERT Transformer model [6] to retrieve the sentence embedding for each utterance. To expedite computation, we use the BERT-SMALL model first published in [14] and made available on the HuggingFace platform. Therefore the embedding size is 512 dimensions.

2.4 Prosodic Features

Prosodic features here refers to the non-linguistic features of speech. Each group's audio files were processed using openSMILE to extract prosodic features of speech — e.g., features relating to frequency, amplitude, and balance. We used the extended feature set predefined by Eyben *et al.* [7]. This feature set aims to be minimalist while still effective. After processing, each utterance has an associated total of 88 prosodic features, such as loudness and spectral flux.

2.5 Model Training

All models were trained and evaluated using leave-one-group-out cross-validation. We trained and evaluated three types of models for evaluation: a random forest (RF) and AdaBoost (AB) classifier where optimal hyperparameters were identified with Hyperopt's [2] guided TPE search, and a neural network (NN) classifier where hyperparameters were identified with a grid search.

3 Results

Table 2 shows the results of binary CPS facet classification with respective standard deviations given in Table 3. As discussed in our related works, binary classification (presence or absence of a CPS facet) is important since some utterances may contain multiple collaborative components [13]. Area Under the Receiver Operating Characteristic Curve (AUROC) was computed using test results from every utterance.

Table 2: Weighted average AUROC for binary classification

	Construction of shared knowledge			Negotiation/ Coordination			Maintaining team function		
Modalities	RF	AB	NN	RF	AB	NN	RF	AB	NN
Verbal	.814	.804	.829	.788	.783	.791	.712	.689	.678
Prosodic	.832	.796	.714	.730	.710	.595	.661	.649	.598
Verbal + Prosodic	.840	.818	.794	.785	.794	.760	.720	.699	.645

Table 3: Standard deviations of weighted average AUROC across all 10 groups for binary classification

	Construction of shared knowledge			Negotiation/ Coordination			Maintaining team function		
Modalities	RF	AB	NN	RF	AB	NN	RF	AB	NN
Verbal	.044	.037	.040	.054	.052	.057	.082	.079	.079
Prosodic	.038	.051	.118	.055	.056	.094	.077	.074	.091
Verbal + Prosodic	.035	.044	.143	.054	.052	.099	.076	.088	.095

4 Discussion

In many cases, we achieve results comparable to or even exceeding those reported in [12], even though our shared environment and task are noisier and our data size is smaller (30 participants compared to 111).

We often observe that performance with feature combinations does not significantly exceed that with verbal (linguistic) features alone. In these cases, the utterances or numerical representations thereof usually carry sufficient information to classify or detect CPS facets most of the time.

4.1 Qualitative Error Analysis

In order to identify instances where multimodal features helped with CPS facet detection, we examined 50 random samples where predictions by the random

forest model were wrong using verbal features only and correct when using verbal+prosodic. 32% of utterances required prosody to clarify intent (e.g., one participant said "that's 100" to indicate the team had done well, not to posit that a block weighed 100 grams) — half of this set were instances of participants asking questions. 16% of the 50 sample utterances contained interruptions (e.g., Participant 2: "So even when they're equal it leans—"; Participant 1: "It leans slowly that way"). For the remaining samples, it was not immediately clear how the inclusion of prosodic information led to correct CPS identification, but within this subset, we noted that 78.9% of utterances were correctly classified by the prosodic classifier, indicating the prosodic signal alone was sufficient and they did not benefit from the combination of verbal and prosodic features. Finally, we found 374 out of 1,822 (20%) utterances were misclassified by all models — these utterances clearly require additional features or modalities.

5 Limitations, Future Work, and Conclusion

In this study, we have used several tools for automatic feature extraction and trained multiple machine learning classifier models to detect collaborative problem solving (CPS) in small groups. We achieve promising results on multimodal detection of CPS in a challenging in-person setting: a task that requires real-time interaction with physical objects.

There are several limitations to this effort that future work could address. While this is in line with other efforts, we are currently identifying CPS at the facet-level (the most coarse-grained) and future work will need to identify CPS at the sub-facet and indicator levels. While our participants exhibit a range of ethnic, national, and linguistic backgrounds, all participants still tend to satisfy most conditions of the WEIRD paradigm [10]. We are not utilizing any visual features but the weights task itself is ripe to take advantage of body pose and block interactions — some work on incorporating gesture was recently conducted by [15]. We experiment with only one CPS task and it would behoove future research to explore task-agnostic CPS classification. Our dataset is collected outside of classrooms — classroom environments will inevitably be subject to additional noise. Future work will need to consider how robust our solutions are to classroom noise. Finally, the ultimate goal of this work is to create an AI agent to assist small groups — this work is an important milestone, but much remains to be done before an agent can actively interact with a group.

References

1. Avci, U., Aran, O.: Predicting the Performance in Decision-Making Tasks: From Individual Cues to Group Interaction. IEEE Transactions on Multimedia **18**(4), 643–658 (Apr 2016). https://doi.org/10.1109/TMM.2016.2521348
2. Bergstra, J., Yamins, D., Cox, D.D.: Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In: Proc. of the 30th International Conference on Machine Learning (2013)

3. Bradford, M., Hansen, P., Beveridge, J.R., Krishnaswamy, N., Blanchard, N.: A deep dive into microphone hardware for recording collaborative group work. In: Proceedings of the International Conference on Educational Data Mining (2022)
4. Castillon, I., VanderHoeven, H., Bradford, M., Venkatesha, V., Krishnaswamy, N., Blanchard, N.: Multimodal Features for Group Dynamic-Aware Agents. In: Interdisciplinary Approaches to Getting AI Experts and Education Stakeholders Talking Workshop at AIEd. International AIEd Society (2022)
5. Cukurova, M., Zhou, Q., Spikol, D., Landolfi, L.: Modelling collaborative problem-solving competence with transparent learning analytics: is video data enough? In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge. pp. 270–275 (2020)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (May 2019). https://doi.org/10.48550/arXiv.1810.04805
7. Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., Truong, K.P.: The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing **7**(2), 190–202 (Apr 2016). https://doi.org/10.1109/TAFFC.2015.2457417
8. Graesser, A.C., Fiore, S.M., Greiff, S., Andrews-Todd, J., Foltz, P.W., Hesse, F.W.: Advancing the Science of Collaborative Problem Solving. Psychological Science in the Public Interest **19**(2), 59–92 (Nov 2018). https://doi.org/10.1177/1529100618808244
9. Grover, S., Bienkowski, M., Tamrakar, A., Siddiquie, B., Salter, D., Divakaran, A.: Multimodal analytics to study collaborative problem solving in pair programming. In: Proceedings of the Sixth International Conference on Learning Analytics Knowledge. pp. 516–517 (2016)
10. Henrich, J., Heine, S.J., Norenzayan, A.: The weirdest people in the world? Behavioral and brain sciences **33**(2-3), 61–83 (2010)
11. Karrer, R.: Google WebRTC Voice Activity Detection (VAD) module (2022), https://www.mathworks.com/matlabcentral/fileexchange/78895-google-webrtc-voice-activity-detection-vad-module
12. Stewart, A.E.B., Keirn, Z., D'Mello, S.K.: Multimodal modeling of collaborative problem-solving facets in triads. User Modeling and User-Adapted Interaction **31**(4), 713–751 (Sep 2021). https://doi.org/10.1007/s11257-021-09290-y
13. Sun, C., Shute, V.J., Stewart, A., Yonehiro, J., Duran, N., D'Mello, S.: Towards a generalized competency model of collaborative problem solving. Computers & Education **143**, 103672 (2020), https://www.sciencedirect.com/science/article/pii/S0360131519302258
14. Turc, I., Chang, M.W., Lee, K., Toutanova, K.: Well-read students learn better: On the importance of pre-training compact models. arXiv preprint arXiv:1908.08962 (2019)
15. Vanderhoeven, H., Blanchard, N., Krishnaswamy, N.: Robust motion recognition using gesture phase annotation. In: International Conference on Human-Computer Interaction (HCII). Springer (2023)