Speech Is Not Enough: Interpreting Nonverbal Indicators of Common Knowledge and Engagement

Derek Palmer[1], Yifan Zhu[1], Kenneth Lai[1], Hannah VanderHoeven[2], Mariah Bradford[2], Ibrahim Khebour[2], Carlos Mabrey[2], Jack Fitzgerald[2], Nikhil Krishnaswamy[2], Martha Palmer[3], James Pustejovsky[1]

[1]Brandeis University
[2]Colorado State University
[3]University of Colorado Boulder
{derek.palmer@colorado.edu, jamesp@brandeis.edu}

Abstract

Our goal is to develop an AI Partner that can provide support for group problem solving and social dynamics. In multiparty working group environments, multimodal analytics is crucial for identifying non-verbal interactions of group members. In conjunction with their verbal participation, this creates an holistic understanding of collaboration and engagement that provides necessary context for the AI Partner. In this demo, we illustrate our present capabilities at detecting and tracking nonverbal behavior in student task-oriented interactions in the classroom, and the implications for tracking common ground and engagement.

Introduction

Our goal is developing an AI partner that can provide beneficial information or suggestions to groups of collaborators in real time. Essential to this process is an accurate interpretation of two dimensions of the AI partner's environment: the working group's knowledge of the topic, and the current social dynamics of the group. Multi-modal analysis offers unique analysis of vital non-verbal cues (Dey and Puntambekar 2023). Also, the more complex and novel the environment, the less reliable automatic speech recognition is. Multi-channel input is crucial for useful input to AI partners. A demo video is at https://youtu.be/WzajCzOYggg.

The Task Our Institute for Student-AI Teaming (iSAT), aims to develop an AI Partner that can intervene positively in collaborative problem solving groups of students (D'Mello et al. 2024). Student group productivity can be heavily influenced by social dynamics (Moulder, Duran, and D'Mello 2022). Social dynamics can be positive or negative based on the behavior and level of engagement of each individual member (Adams-Wiggins and Dancis 2022). Social cohesion is positive social dynamics manifesting as high levels of engagement for all group members culminating in constructive progress towards the group's goal. Negative social cohesion can be either low levels of engagement of the group with little progress towards the goal or unconstructive interactions. Automatic speech recognition is a vital part of identifying both positive and negative social situations. However, as the amount of topics, participants, and background noise expands, ASR accuracy decreases (Cao et al. 2023b), increasing reliance on multimodal analysis, especially with speaker cohorts such as children with minimal training data. Also, many non-verbal interactions are simply inaccessible to ASR and critical to capture via multi-modal analysis.

Tracking gestures such as pointing can be indispensable in building real-time understanding of a group's common knowledge (Khebour et al. 2024b; Tu et al. 2024). Tracking each individual's posture over time, in particular leaning in or leaning out, is a powerful indicator of a group's engagement level (Adams-Wiggins and Dancis 2022). Joint visual attention is critical to contextualize both common knowledge and group engagement. All together these modalities help pinpoint intervention opportunities and avoid unconstructive interruptions (Cao et al. 2023a). We illustrate multi-modal analysis in both dimensions: 1) a knowledge support analysis with our Fibonacci weights exercise; 2) contrasting levels of engagement only observable via multimodal analysis for our simulated classroom environment.

Our Setup

The physical task space consists of a table with task-relevant objects on it and 3 participants seated around it. The task is recorded using an Azure Kinect RGBD camera, and an MXL AC-404 ProCon microphone. The Kinect automatically tracks 32 joints per body, covering head, torso, and limbs, returning 3 position and 4 orientation values for each.

Gaze detection uses the direction of participants' noses as a proxy. Using the joints of bodies extracted by the Azure Kinect SDK, we take the average position of both ear joints, which results in a point roughly behind the nose, and use a vector connecting this point and the nose joint to indicate gaze direction. We extend this vector (purple) into 3D space to see which objects participants' gazes are landing on.

Posture detection determines the participants' positions (left, middle, right) using the x-coordinate of the pelvis of each body. Each participant's position and orientation information is then flattened. The vectors are stacked and then input into a two-layer feedforward neural network. We train three such models, one for each participant position.

Gesture recognition primarily concerns pointing detection. We use a lightweight 2-stage method from VanderHo-

Figure 1: Object detection in Weights Task

even, Blanchard, and Krishnaswamy (2023, 2024) with features extracted from the video using MediaPipe (Lugaresi et al. 2019). First we detect if a gesture is in the "stroke" phase (following Kendon (1997)) and then classify the gesture's shape. As with gaze, to infer the target objects of a pointing gesture, we calculate a "pointing frustum" (Kranstedt et al. 2006) from the extended digit into 3D space (blue and orange in the video). Objects intersecting this frustum are considered "selected" (highlighted green).

Real-time object detection in the video is performed using a FasterRCNN ResNet-50-FPN model (Lin et al. 2017), initialized using the default ResNet-50-FPN weights from TorchVision, then trained over annotations of object bounding boxes for 10 epochs using SGD with learning rate $1e-3$, momentum $9e-1$, and weight decay $5e-4$. Batch size was 32 and input size was $3\times416\times416$.

Video Content

Our two scenarios (Figs. 1 and 2) prioritize different aspects of multimodal information processing: knowledge support often includes the evaluation of specific objects and classroom discussions do not. The context for knowledge support is equally dependent on joint visual attention, gesture and domain specific object detection. For social cohesion, joint visual attention to speakers and posture are more valuable.

Scenario 1: Fibonacci Weights Task To better evaluate the accuracy and utility of object detection, we developed a situated collaborative task wherein participants infer the weights of a set of differently weighted blocks with the use of a balance scale (Khebour et al. 2024a). The increases in weight adhere to the Fibonacci series. A series of lab experiments provided video data for annotation and training purposes (see Fig. 1). The expectation is that object detection provides valuable input for AI Partners offering knowledge support. With minimal amounts of training data we can port to similar new objects specific to new domains.

Scenario 2: Simulated Classroom Project Planning The simulated classroom content we are demoing is a project planning scenario from Lesson 4 of the Sensor Immersion Curriculum Unit developed by the SchoolWide Labs project (Biddy et al. 2021). In the lesson each student comes into the group as the sole expert on one of three specific sensor types. The students answer factual questions about the capabilities of their sensors and brainstorm about potential

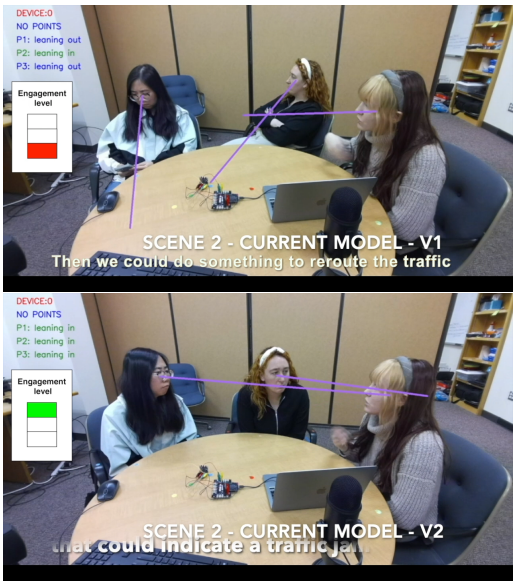

Figure 2: Contrasting engagement levels in simulated project planning

beneficial projects involving multiple sensors in real-world environments. An AI Partner (Cao et al. 2023a) guides and supports these conversations, via a chat window.

Deciding which problem to address at any given time is a compromise between what can be detected, theories of behavior based on behavioral/cognitive models, and defined lesson goals. Non-verbal Multimodal analysis is primarily useful for determining whether social dynamics are positive or negative, relative to the specific domain under investigation. Our AI partner has several knowledge support and social cohesion states it attempts to recognize, as defined and vetted by educational researchers (Zhang et al. 2024). An example of a state, Dominated Discussion, is the detection of a single group member talking for 30 seconds or more. Our demo illustrates the necessity of using multi-modal analysis to accurately identify whether the dominated discussion results in disengaged participants (Fig. 2).

Conclusion and Future Work

We have demonstrated the importance of integrating non-verbal behavior recognition into the modeling and interpretation of multi-party dialogues when there is an intervention objective. With an eye on portability, our next focus for object detection will be devices common to most settings such as tablets, laptops and phones. We expect that our approach to modeling social cohesion will port to any AI partner included in a working group of 3 or more, given a document summarizing the specific topic and a task specific model of engagement (Moore 2016; Kofod-Petersen, Wegener, and Cassens 2009). This could include business, government, health, or education settings, such as board meetings, working task forces, training exercises, etc. An opportunity also exists for decreasing the time intensive labor of video annotation tasks for qualitative research purposes.

Acknowledgements

We are extremely grateful to the brilliant actors in our demo, Rui Zhang, Sierra Rose and Brooklyn Clines. This material is based in part upon work supported by the National Science Foundation (NSF) under subcontracts to Colorado State University and Brandeis University on award DRL 2019805 (Institute for Student-AI Teaming), and by Other Transaction award HR00112490377 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program. Approved for public release, distribution unlimited. Views expressed herein do not reflect the policy or position of the National Science Foundation, the Department of Defense, or the U.S. Government.

References

Adams-Wiggins, K. R.; and Dancis, J. S. 2022. Marginality in inquiry-based science learning contexts: the role of exclusion cascades. *Mind, culture, and activity*, 29(4): 356–373.

Biddy, Q.; Gendreau Chakarov, A.; Bush, J.; Hennessy Elliott, C.; Jacobs, J.; Recker, M.; Sumner, T.; and Penuel, W. 2021. A Professional Development Model to Integrate Computational Thinking Into Middle School Science Through Codesigned Storylines. *Contemporary issues in technology and teacher education*, 21(1).

Cao, J.; Dickler, R.; Grace, M.; Bush, J. B.; Roncone, A.; Hirshfield, L. M.; Walker, M. A.; and Palmer, M. S. 2023a. Designing an AI Partner for Jigsaw classrooms. In *Proceedings of the Workshop on Language-Based AI Agent Interaction with Children (AIAIC'2023)*.

Cao, J.; Ganesh, A.; Cai, J.; Southwell, R.; Perkoff, E. M.; Regan, M.; Kann, K.; Martin, J. H.; Palmer, M.; and D'Mello, S. 2023b. A Comparative Analysis of Automatic Speech Recognition Errors in Small Group Classroom Discourse. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. ACM.

Dey, I.; and Puntambekar, S. 2023. Examining nonverbal interactions to better understand collaborative learning. In *Proceedings of Computer Support for Collaborative Learning 2023*, 273–276. International Society of the Learning Sciences.

D'Mello, S. K.; Biddy, Q.; Breideband, T.; Bush, J.; Chang, M.; Cortez, A.; Flanigan, J.; Foltz, P. W.; Gorman, J. C.; Hirshfield, L.; Ko, M.; Krishnaswamy, N.; Lieber, R.; Martin, J.; Palmer, M.; Penuel, W. R.; Philip, T.; Puntambekar, S.; Pustejovsky, J.; Reitman, J. G.; Sumner, T.; Tissenbaum, M.; Walker, L.; and Whitehill, J. 2024. From learning optimization to learner flourishing: Reimagining AI in Education at the Institute for Student-AI Teaming (iSAT). *AI Magazine*, 45(1): 61–68.

Kendon, A. 1997. Gesture. *Annual review of anthropology*, 26(1): 109–128.

Khebour, I.; Brutti, R.; Dey, I.; Dickler, R.; Sikes, K.; Lai, K.; Bradford, M.; Cates, B.; Hansen, P.; Jung, C.; Wisniewski, B.; Terpstra, C.; Hirshfield, L. M.; Puntambekar, S.; Blanchard, N.; James, P.; and Krishnaswamy, N. 2024a. When Text and Speech are Not Enough: A Multimodal Dataset of Collaboration in a Situated Task. *Journal of Open Humanities Data*, 10(1).

Khebour, I. K.; Lai, K.; Bradford, M.; Zhu, Y.; Brutti, R. A.; Tam, C.; Tu, J.; Ibarra, B. A.; Blanchard, N.; Krishnaswamy, N.; and Pustejovsky, J. 2024b. Common Ground Tracking in Multimodal Dialogue. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3587–3602. Torino, Italia: ELRA and ICCL.

Kofod-Petersen, A.; Wegener, R.; and Cassens, J. 2009. Closed Doors – Modelling Intention in Behavioural Interfaces. Tapir Akademisk Forlag, Trondheim, Norway.

Kranstedt, A.; Lücking, A.; Pfeiffer, T.; Rieser, H.; and Wachsmuth, I. 2006. Deixis: How to determine demonstrated objects using a pointing cone. In *Gesture in Human-Computer Interaction and Simulation: 6th International Gesture Workshop, GW 2005, Berder Island, France, May 18-20, 2005, Revised Selected Papers 6*, 300–311. Springer.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.-L.; Yong, M.; Lee, J.; et al. 2019. Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, volume 2019.

Moore, A. 2016. *Lovers, wrestlers, surgeons: a contextually sensitive approach to modelling body alignment and interpersonal engagement in surgical teams.*, 257–285. ISBN 9781781790502.

Moulder, R. G.; Duran, N. D.; and D'Mello, S. K. 2022. Assessing multimodal dynamics in multi-party collaborative interactions with multi-level vector autoregression. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, 615–625.

Tu, J.; Rim, K.; Ye, B.; Lai, K.; and Pustejovsky, J. 2024. Dense Paraphrasing for Multimodal Dialogue Interpretation. *Frontiers in Artificial Intelligence*, 7.

VanderHoeven, H.; Blanchard, N.; and Krishnaswamy, N. 2023. Robust motion recognition using gesture phase annotation. In *International conference on human-computer interaction*, 592–608. Springer.

VanderHoeven, H.; Blanchard, N.; and Krishnaswamy, N. 2024. Point target detection for multimodal communication. In *International Conference on Human-Computer Interaction*, 356–373. Springer.

Zhang, R.; Cao, J.; Dey, I.; Foltz, P.; Palmer, M.; Tissenbaum, M.; Biddy, Q.; Doherty, E.; Bodzianowski, M.; Palmer, D.; and Hirshfield, L. 2024. "JIA Fueled My Ideas": Designing an Interactive AI Partner for Assisting Small Group Collaborations among Students Aged 12-17. In *Manuscript submitted for publication.*