Multimodal Analysis of Dyads Collaboratively Exploring Emergent System Microworlds

Abizar Bagasrawala **Jacob Kelter**

INTRODUCTION

Collaborative Learning in Emergent System Microworlds

Constructionism, as an educational philosophy, emphasizes not only the power of learning through doing but also the power of thinking and talking about what you are doing [1], [2]. This suggests a natural connection between research on Constructionist learning environments and collaborative learning.

One type of constructionist learning environment, called a microworld [3], aims to represent specific area of math or science in a way that allows the learner to learn through open-ended or lightly guided exploration [4]. One line of work aims to help learners understand emergent systems through interreacting with and building multi-agent models [5]. Such microworlds have been called "emergent systems microworlds" (ESMs) [6]. Classroom learning using ESMs often follows a progression in which the teacher introduces the ESM, the students then explore the parameter space and build an understanding of the model the phenomenon and then finally extend or modify the model. Little research has specifically investigated the collaborative nature of the learning that takes place during the model exploration phase.

Key Collaboration Concepts and Multimodal Learning Analysis

Barron [7] identified a number of important factors for productive collaborative problem solving including that (1) participants establish a collaborative, as opposed to independent, orientation, (2) there is balance and mutuality in exchanges so that all participants can contribute and (3) participants achieve joint attention. Collaborative *learning* has been framed as problem of how two people construct a shared meaning [8]. Traditionally, these types of studies rely primarily on qualitative methods due to the complexity and nuanced nature of the relevant interactions.

Qualitative analysis will continue to be indispensable for understanding complex human learning, but for the past several years, the field of multimodal learning analytics (MMLA) has begun to leverage multiple quantitative data sources to attempt gain new perspectives on complex learning [9]. Specifically for collaboration, Schneider and Pea have studied how joint visual attention mediates collaboration and dialogue in pairs working remotely through a shared computer interface on a collaborative learning task [10], [11].

Research Questions

This paper is an exploratory study of using MMLA techniques to study collaborative learning using ESMs. Specifically, our research questions are:

1. Can we use MMLA techniques to identify patterns of collaboration when pairs engage with ESMs?
2. Do those patterns differ when pairs use two laptops vs one shared laptop?

The second research question is particularly relevant given wide-spread "one laptop per child" initiatives. Many classrooms now have the resources to have every student working on their own laptop, but this might not actually be ideal for certain collaborative learning activities.

METHODS

Participants and Setting

The participants of the study were undergraduate students. The study took place in a laboratory environment with both participants and both authors present. This paper only reports on the results from a single dyad, Samir and Pablo (pseudonyms). We collected data from two other dyads with different orderings of the task conditions discussed next.

The Task

Participants were asked to explore two ESMs built using NetLogo [12]. One ESM contains a model of a predator-prey ecosystem consisting of wolves, sheep and grass. The other contains a model of traffic jams. The dyad first explored the ecosystem model while sharing a single laptop. Then they explored the traffic model while each using their own laptops. They were still sitting side by side and could view each other's laptops.

Data Collection

We used The Social Signal Interpretation Framework to collect the following data: video of the participants using a webcam, audio, screen capture including cursor position and clicks.

Data Analysis

Speech To Text

To perform our analyses, we needed to transcribe our audio data. We attempted to automate the transcription using Otter.ai and Google Speech-To-Text, but the resulting transcription had poor accuracy. To refine our results, we manually corrected the transcriptions. Furthermore, we noted start and end times for each utterance, which was important for our other analyses.

Silences and Overlaps in Speech

Using transcript data from the previous subsection, we calculated the duration of each utterance, the silences between utterances and the duration of any overlapping speech. This data was intended to shed light on the turn-taking behavior of the participants, specifically if they developed a fluid and reciprocal conversation.

Speech Over Time by Participant

We wanted a measure for participants' speech distribution over time. Specifically, we wanted to look at the amount of speech over bucketed time intervals. Based on the fineness of data we required, we chose to use time intervals of 2 minutes.

To compute amount of speech in each bucket, there were two options: (1) number of words spoken, and (2) time spoken for. Ideally, both measures should be used together, but choosing one makes the analysis more feasible.

We chose (1) for our analysis because it turned out to be a better measure for our dataset. Specifically, for the two participants we analyzed, we felt that number of words was a better measure for how much they contributed to the conversation. If we felt that either participant used more words to communicate less substance, we might have considered using (2) instead, but this was not the case.

Coherence of Successive Utterances

Following [11] we vectorized each utterance using bag-of-words and calculated the cosine similarity between each utterance and the previous 3 utterances of the other participant. We then took the maximum cosine similarity of these three and used it as a measure of the coherence of the current utterance with the preceding ones. This is a very rough measure but should give some sense of whether the participants are building off each other in the conversation or not.

To preprocess our text for bag-of-words, we removed stop-words and stemmed the vocabulary using NLTK. While vectoring words, we used tf-idf so that our model would assess more frequent words as being less important.

For a more sophisticated vectorization method, we tried using Doc2Vec on each utterance, but we couldn't find a suitable pre-trained Doc2Vec model and our dataset wasn't nearly large enough to train our own deep neural network.

Joint Visual Attention

As noted in the introduction, achieving joint visual attention is important for collaboration. We attempted to create an automated process to classify a simple visual state for each participant: whether they were looking at their own laptop or looking at their partner/partner's laptop. To do this, the video was cropped in half and each half fed into OpenFace2 [13]. OpenFace2 returns (among other things) the head angle around each axis in 3D and we used the rotation around the y-axis (the axis of the person's neck) to try to determine where the person was looking. This simple algorithm was moderately successful but failed when the person rotated too far to the side, because then OpenFace stopped recognizing a face at all. As a result, for the purpose of this project, we manually coded the visual state for each participant, noting the time whenever they changed states.

Qualitative Analysis

We have conducted some informal qualitative analysis to help shed light on the quantitative measures we calculated. This consisted of watching videos of the interactions and reading transcripts to then create a narrative description of what happened.

RESULTS

All of the data we collected and analyzed are in Figures 2 and 3 for the one laptop and two laptop conditions respectively. We will first provide interpretations of each sub-figure individually, comparing across the two conditions. Then we will note interesting relationships between the subgraphs. Qualitative analysis and data interspersed for context and to aid our interpretation.

Silences and Overlaps in Speech

Figure 1A and 2A show the time series of individual speech, silences and overlapping speech. The single laptop condition seems to have had both more overlaps and more silences, but it is hard to tell with certainty. More overlaps could be due to tighter joint attention in the one laptop case which results in both people thinking of similar things to say at the same time. However, it could also just have been due to the one laptop condition being first and the participants were still getting used to collaborating or any number of other factors including just randomness. More silences in the one laptop case is likely due to the participants being less comfortable during the first activity than the second.

One important thing to note is that contrary to our expectation, overlaps seemed to be a sign of very tight collaboration. They were usually due to the two participants saying something similar at the same time, a sign of building joint understanding, as opposed to speaking over each other.

Speech Over Time by Participant

Figures 1B and 2B show a time series of amount of speech over time for each participant. Specifically, amount of speech is measured as number of words spoken per 2 minute interval.

Comparing across both figures, both participants had more speech in 2B, the two laptop condition. Specifically, Samir spoke 758 words in the single laptop condition, normalized to 44.59 words per minute, and 765 words in the two laptop

condition, normalized to 63.75 words per minute. As for Pablo, he spoke 857 words in the single laptop condition, normalized to 50.41 words per minute, and 680 words in the two laptop condition, normalized to 56.67 words per minute.

The fact that both participants had more speech in the two laptop condition could be because of the following With the dual option of either interacting with their own model or collaborating with the other person, they had more opportunities to build contributions. However, the difference could also be because participants experienced the two laptop condition after the single laptop condition, making participants more comfortable in the former.

We'd like to draw attention to a specific moment in Figure 1B, the one laptop condition. Between the 10 minute mark and the 14 minute mark, we notice that Pablo speaks more than Samir. Upon watching the video across this 4-minute time interval, we realize that Pablo speaks more because he is speaking what he types. This realization brings up three important points that we address below.

Firstly, it is less likely that one participant would speak what they type in the two laptop condition than in the one laptop condition. We believe this is because, in the two laptop condition, the participant typing would not want to disturb the other participant, who has the option to explore the model on his own laptop while his partner types. The fact that the one laptop condition prevents the other participant from exploring the model while one is typing impacts their collaboration. On one hand, it could be seen as a bottleneck that limits the other participant. On the other hand, by forcing them to pay attention to each other's answers, it could be motivating them to come to consensus, clarify doubts, and ensure they understand what they other person types.

The second point the moment described previously brings up is that it is important to have the context while analyzing speech over time. Seeing that one participant spoke more than the other for a 4-minute interval, one might conclude that the participant speaking more dominated the conversation. However, as we learn from this moment, we need more context to understand what factors led one participant to speak more. Additionally, we learn from qualitatively analyzing the videos that Pablo has a tendency to talk while he's typing, while Samir doesn't. Such individual differences must be factored in while analyzing speech over time.

Coherence of Successive Utterances

Figures 1C and 2C show the cosine similarity of each vectorized utterance with the most similar of the previous three utterances of the other person. This is intended to be a measure of the coherence of the conversation.

One period with particularly high similarity scores is from around 3:00 to 3:30 in Figure 2C, the two laptop condition. This period in the conversation was a very focused discussion about the maximum speed of the cars in the traffic model:

> Pablo: It's gonna just stay, I think, at the **max speed** now.
>
> Samir: Yeah. Oh, it's all just about Yeah, but yeah, because the min **speed** is at the **max speed**. So
>
> Pablo: Yeah
>
> Samir: Yeah. And then as the number of cars increase, it can't it can't reach the **max speed** anymore like this period which is much lower,
>
> Pablo: It reaches the **max speed**. Right?
>
> Samir: It took
>
> Pablo: Because the **max speed** is capped.
>
> Samir: Where it's
>
> Pablo: Look it reaches a **maximum**.
>
> Samir: Mm, Yeah. Okay, but it **reaches a lower max speed.**
>
> Pablo: Yeah, it **reaches a lower max speed.**

In this exchange, the participants are highly focused on understanding when and why the cars reach a maximum speed. The repeated use of the phrase “max speed” and in the last two statements the entire phrase “reaches a lower max speed” led to the high similarity scores.

In contrast, the beginning of the two-laptop condition had quite low similarity until almost the two minute mark. These interactions seemed to be more exploratory and less focused:

> Samir: And I'm just trying this, the deceleration more than the acceleration.
>
> Pablo: Okay, sounds good.
>
> Samir: Right now at 20.
>
> Pablo: Yeah.
>
> Samir: When the red car reaches the max it stops.
>
> Pablo: Yeah, same here. It's
>
> Samir: Over here, I don't know what happend.
>
> Samir: Yeah,
>
> Pablo: I think that's the first one is it's reaching equilibrium kind of thing. But it's oh, but then its weird.
>
> Samir: Yeah, like this.
>
> Pablo: Oh, it a pattern. Do you see it?
>
> Samir: Yeah. Go on.

The participants in this interaction are trying out different parameters and sharing the results. This is a crucial part of the collaboration, but doesn’t necessarily lend itself to focused discussion that results in repeated words and thus high similarity scores.

In the one-laptop condition we saw similar patterns. The period of high similarity statements around the 8 minute mark in Figure 1C were due to a highly focused discussion about the maximum and minimum population levels in the wolf-sheep predation model. In contrast, the low similarity region after the 2 minute mark was a more exploratory part of the collaboration with more general statements without much immediate follow-up discussion.

These findings lead us to the conclusion that high and low similarity regions of the conversation are not signs of good and bad collaboration, at least this pair of participants. Rather, they can signify different phases of the collaboration with low similarity periods corresponding to more exploratory parts of the dialogue and high similarity periods corresponding to more focused sense-making, potentially with some disagreement to be resolved. Both are important parts of collaboratively building understanding.

Joint Visual Attention

Figures 1D and 2D show time series of the attentional states of the two participants. Given the different setups, the relevant states were different. For the single laptop condition, shown in Figure 1D, grey represents looking at the laptop, purple represents looking at the other person and white represents looking at something else. The vast majority of the time, both participants were looking at the laptop. They occasionally pointed to something on the screen and it seems it was easy for them to maintain joint attention on the laptop while they spoke. Occasionally one participant would look at the other, only three times did they both look at each other simultaneously.

For the two laptop condition, the participants spent the majority of the time looking at either their own laptop or the other laptop. So, we coded each participant as in one of two states: either (1) looking at their own laptop or (2) looking at the other person's laptop or the other person. In Figure 2D, red represents looking at Samir or his computer and blue represents looking at Pablo or his computer. At any given time, when both participants were looking at the same laptop, the graph has matching colors. When they were each looking at their own laptop, Pablo's graph is blue and Samir's is red. When they were looking at each other (which was very rare), then Pablo's graph is red and Samir's blue. It is clear from the graph that Pablo spent more time looking at Samir's laptop than vice versa. The period from about 11 to 12 minutes stands out, as both participants switched their attention back and forth quite a bit. This is because they agreed to run the model with different parameters and then were comparing results across the two laptops.

The two laptop condition certainly gives the participants more to keep track of and they have to decide how to allocate their attention. With these participants, it didn't seem to hamper their collaboration and maybe even helped it. It is possible, maybe even likely, that starting with the the two laptop condition would have been somewhat overwhelming, but our participants already had some experience collaborating and exploring NetLogo models before they had to deal with the added complexity of managing their attention across two laptops.

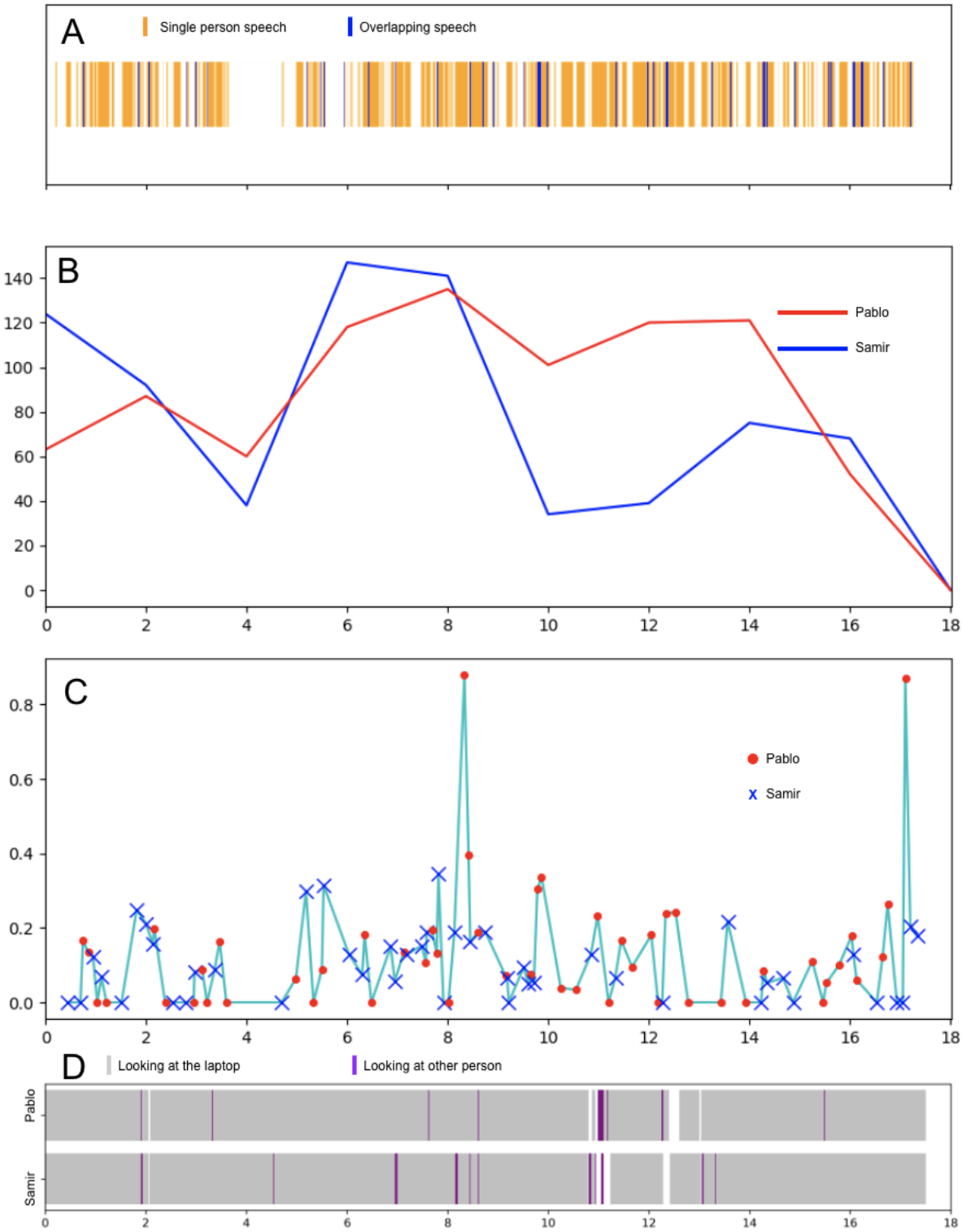

Figure 1: Data from the one laptop condition in which the participants explored the wolf-sheep predation model. All x-axes are minutes into the activity which lasted around 18 minutes **(A)** Visualization of single person speech (yellow), silences (white) and overlapping speech (blue). **(B)** Plot of words spoken each 30 second interval by Pablo (red) and Samir (blue) **(C)** Plot of the cosine similarity of each statement with most similar of previous three statements. **(C)** Visualization of the gaze state of each participant. The vast majority of the time they both were looking at the laptop and occasionally looked at each other.

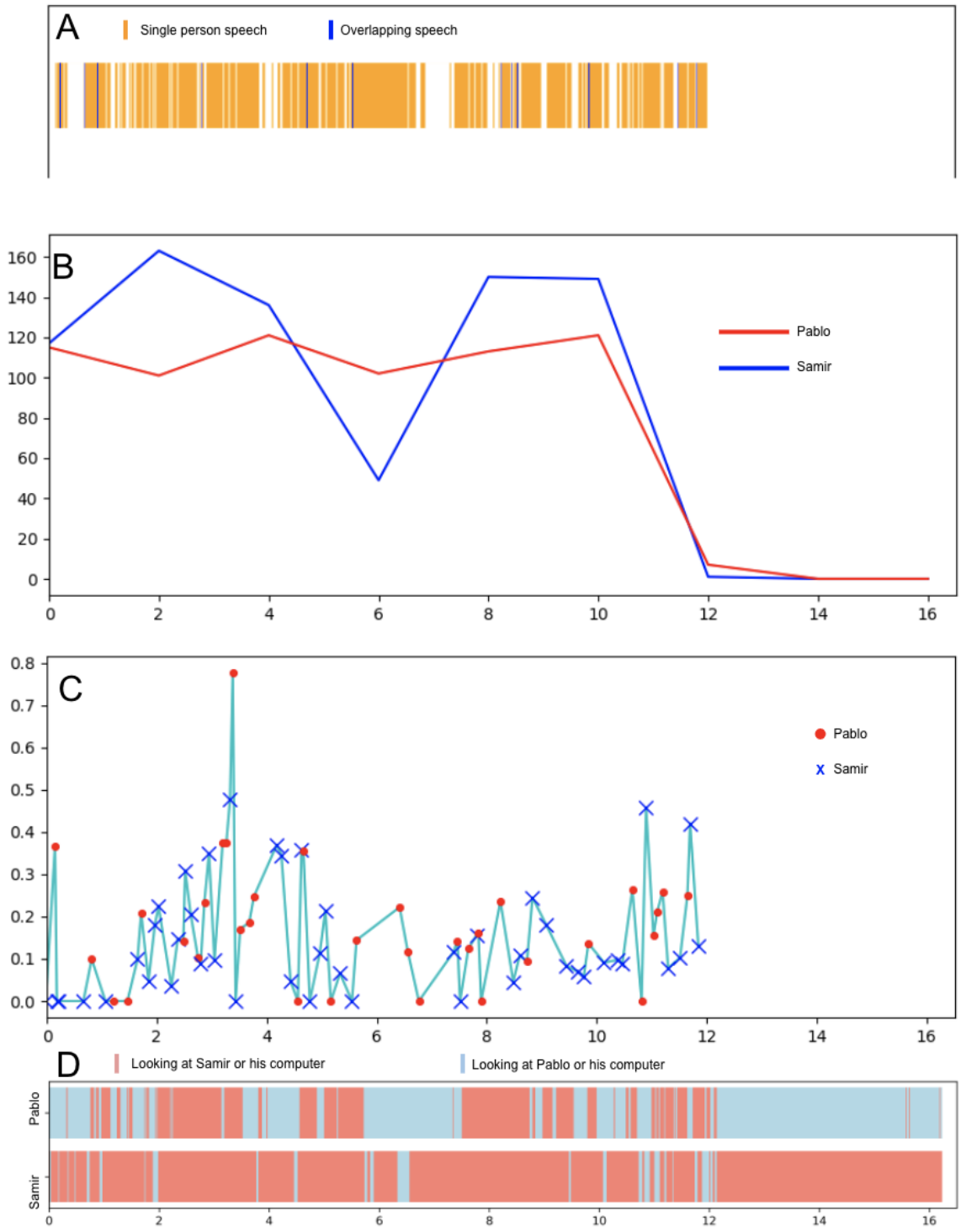

Figure 2: Data from the two laptop condition in which the participants explored the traffic model. All x-axes are minutes into the activity which lasted around 16 minutes. **A**, **B** and **C** are the same plots as in figure 1. **(C)** Visualization of the gaze state of each participant. Blue represents the participant looking at Pablo's laptop or Pablo. Red represents looking at Samir's laptop or Samir. Any point in time when both plots are the same color, they are looking at the same laptop. If the top graph is blue and the bottom is red, they were each looking at their own laptop. If the top graph is red and the bottom is blue they were looking at each other, but this was quite rare.

DISCUSSION

Perhaps the most important finding from this exploratory study is the importance of context and qualitative information for interpreting quantitative data. We initially assumed that overlapping speech would signal poor collaboration, but qualitative analysis revealed otherwise. Similarly, without qualitative analysis, we might have considered low coherence parts of the dialogue poor collaboration and high coherence parts good collaboration. However, qualitative analysis revealed that, at least with this dyad, low and high coherence periods are better thought of as exploratory vs focused discussion, both crucial aspects for ultimately building understanding.

Regarding our first research question we can tentatively answer that it seems possible to identify patterns of collaboration in dyads exploring ESMs, but it will probably be difficult to identify purely quantitative patterns without qualitative context. One pattern that seemed to emerge is initial parts of dialogue with low coherence scores corresponding to exploration followed by dialogue with increasing coherence scores as the conversation focused.

Regarding our second research question, the only clear difference between the two conditions was the visual gaze states. This is an unsurprising finding, but still worth noting. Our dyad managed their attention quite well across the two laptops. It is likely that starting with the one-laptop condition helped them to initially become comfortable collaborating and exploring ESMs. Further work is needed to test this hypothesis.

Limitations and Future Work

While we set up the groundwork for analyzing collaboration between dyads, there is significant work left to be done. Firstly, our analysis was restricted to one dyad because we wanted to develop a qualitative understanding of different questions we could ask using the data collected. The quantitative approaches we present here will have far greater significance when applied to a larger dataset. We believe we have automated significant portions of the analysis and uncovered some general techniques to set the stage for this larger work. Secondly, while we identified patterns and differences in collaboration, we believe it is vital that future work analyzes the impact of collaboration on learning. Such work could help researchers design collaborative environments to encourage specific kinds of learning. Thirdly, while our work is limited to the task of exploring Emergent System Microworlds, we believe our approaches can be extended to other collaborative problem-solving tasks such as pair programming.

Regarding future work for our quantitative analysis, we outline some specific steps. Firstly, we can perform richer analyses by analyzing all data streams together. For example, we can segment the data by attention states or silences-overlaps and see trends across other factors. Secondly, we can use Netlogo logging data to plot when participants setup the model, changed parameters, and run the model, to further segment the data by their activity. Thirdly, we can incorporate more nuanced text analysis such as identifying utterances that are questions, since questions can be seen as a request for collaboration.

Regarding our data collection process, we can use mobile eye trackers instead of our technique to compute joint visual attention. We believe mobile eye trackers will generate more accurate results. Secondly, we can incorporate skin conductance data to give us stress levels and emotional or physical arousal. While we captured this data using Empatica E4, we did not analyze it and believe it could benefit future work. Thirdly, we can use higher-quality mics and use independent component analysis to separate the audio streams. This would help automate away the significant effort and time that manual transcription required.

CONCLUSION

This exploratory study demonstrated some of the potential benefits and challenges of using MMLA to study collaborative learning in Emergent System Microworlds. Multimodal data can help us to see aspects of collaboration we wouldn't otherwise, allowing us to gain a more nuanced understanding of the complex and nuanced interactions in such settings. However, this very complexity means that the quantitative data rarely, if ever, can speak for themselves. Qualitative analysis is still needed to contextualize and make sense of quantitative multi-modal data.

ACKNOWLEDGMENTS

Thanks to Marcelo Worsley for his guidance on this project.

REFERENCES

[1] S. Papert, "Chapter 1: Situating Constructionism," in *Constructionism: research reports and essays, 1985-1990*, I. Harel and S. Papert, Eds. Norwood N.J.: Ablex, 1991.

[2] I. Harel and S. Papert, "Software Design as a Learning Environment," *Interactive Learning Environments*, vol. 1, no. 1, pp. 1–32, Mar. 1990, doi: 10.1080/1049482900010102.

[3] S. Papert, *Mindstorms: children, computers, and powerful ideas*, 2nd ed. New York: Basic Books, 1980.

[4] L. D. Edwards, "Microworlds as Representations," in *Computers and Exploratory Learning*, 1995, pp. 127–154.

[5] U. Wilensky, "Modeling nature's emergent patterns with multi-agent languages," in *Proceedings of EuroLogo*, 2001, pp. 1–6.

[6] S. Dabholkar and U. Wilensky, "Designing Computational Models As Emergent Systems Microworlds To Support Learning Of Scientific

Inquiry," presented at the International Conference to Review Research in Science, Technology and Mathematics Education, 2020.

[7] B. Barron, "Achieving Coordination in Collaborative Problem-Solving Groups," *Journal of the Learning Sciences*, vol. 9, no. 4, pp. 403–436, Oct. 2000, doi: 10.1207/S15327809JLS0904_2.

[8] J. Roschelle, "Learning by Collaborating: Convergent Conceptual Change," *Journal of the Learning Sciences*, vol. 2, no. 3, pp. 235–276, Jul. 1992, doi: 10.1207/s15327809jls0203_1.

[9] M. Worsley, "Multimodal learning analytics' past, present, and, potential futures," in *CEUR Workshop Proceedings*, 2018, vol. 2163.

[10] B. Schneider and R. Pea, "Real-time mutual gaze perception enhances collaborative learning and collaboration quality," *Intern. J. Comput.-Support. Collab. Learn.*, vol. 8, no. 4, pp. 375–397, Dec. 2013, doi: 10.1007/s11412-013-9181-4.

[11] B. Schneider and R. Pea, "Does seeing one another's gaze affect group dialogue A computational approach.," *JLA*, vol. 2, no. 2, pp. 107–133, Dec. 2015, doi: 10.18608/jla.2015.22.9.

[12] U. Wilensky, *NetLogo*. Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL., 1999.

[13] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 2018, pp. 59–66, doi: 10.1109/FG.2018.00019.