Riding an emotional roller-coaster: A multimodal study of young child's math problem solving activities *

Lujie Chen
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA USA
lujiec@andrew.cmu.edu

Xin Li
Shandong Yingcai University
No. 2 Yingcai Road
Shandong, China
lixin@ycxy.com

Zhuyun Xia
Shandong Yingcai University
No. 2 Yingcai Road
Shandong, China
xiazhujun@ycxy.com

Zhanmei Song
Shandong Yingcai University
No. 2 Yingcai Road
Shandong, China
songzhanmei@ycxy.com

Louis-Philippe Morency
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA USA
morency@cs.cmu.edu

Artur Dubrawski
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA USA
awd@cs.cmu.edu

ABSTRACT

Solving challenging math problems often invites a child to ride an "emotional roller-coaster" and experience a complex mixture of emotions including confusion, frustration, joy, and surprise. Early exposure to this type of "hard fun" may stimulate child's interest and curiosity of mathematics and nurture life long skills such as resilience and perseverance. However, without optimal support, it may also turn off child prematurely due to unresolved frustration. An ideal teacher is able to pick up child's subtle emotional signals in real time and respond optimally to offer cognitive and emotional support. In order to design an intelligent tutor specifically designed for this purpose, it is necessary to understand at fine-grained level the child's emotion experience and its interplay with the inter-personal communication dynamics between child and his/her teacher. In this study, we made such an attempt by analyzing a series of video recordings of problem solving sessions by a young student and his mom, the ideal teacher. We demonstrate a multimodal analysis framework to characterize several aspects of the child-mom interaction patterns within the emotional context at a granular level. We then build machine learning models to predict teacher's response using extracted multimodal features. In addition, we validate the performance of automatic detector of affect, intent-to-connect behavior, and voice activity, using annotated data, which provides evidence of the potential utility of the presented tools in scaling up analysis of this type to large number of subjects and in implementing tools to guide teachers towards optimal interactions in real time.

*(Does NOT produce the permission block, copyright information nor page numbering). For use with ACM_PROC_ARTICLE-SP.CLS. Supported by ACM.

Keywords

math problem solving, affect, interaction dynamics, multimodal learning analytics

1. INTRODUCTION

A popular perception of math education in the US schools is often associated with the lack of inspiration and excitement. One of the possible reasons for that is a common perception of math learning as shallow learning activities such as memorizing multiplication tables and procedure learning activities such as long division [10]. This is especially true with elementary level education where learning facts and procedures accounts for most of the curriculum. In contrast, math problem solving activities can take a form of complex learning [10] that often requires the student to take an adventurous emotional and cognitive "roller-coaster" ride when navigating the uncharted land of possible solutions.

Involvement in this type of activities from young age may play a major role in stimulating student's interest in math and more generally in STEM topics. It may also help building self-confidence and perseverance. However, if not done right, it may disengage student due to unresolved frustration and result in an even more negative view of the subject. It is thus important to know what is the right mixture of emotional and cognitive support to be provided in the process, as well as the right amount and the optimum timing of such support. This role of support is consistent with the vision of a Learning Companion [12] which is a computer system that facilitates learning on the side, is watchful for the trajectory and provides appropriate level of support.

In this study, we explore that question by analyzing the fine-grained multimodal behavior cues that could be automatically extracted from video recordings of one-to-one math problem solving sessions in a naturalist environment. Specifically, we explore data driven methods to characterize the temporal dynamics of the child's emotion states as well as patterns of the interaction between the child and the teacher when problem solving processes unfold.

2. RELATED WORK

A substantial amount of prior work on the automatic detection of student's affective states exists primarily in the context of intelligent tutor systems. [2] introduces a "sensor free" detector to infer engagement from the logs of students' interaction with computerized reading tutor using a method called engagement tracing. [15] uses facial expression analysis to infer engagement during interactive cognitive skill training sessions. Using the same sensing modality, [13] studies an array of affective states such as boredom, confusion, delight, flow, frustration and surprise, based on Facial Action Units. [5] leverages multimodal inputs including conversational cues with computer tutors and gross body language as well as facial features to detect distinct affective states.

While the work mentioned above focuses on static modeling of affects, another thread studies dynamics of affective states. [5] characterizes transitions of affective states between confusion, engagement/flow, boredom and frustration during complex learning activities when using computer tutors. [11] uses a hierarchical dynamic Bayesian network to model temporal dynamics of behavior trends such as flow, stuck and off-task, as well as related emotion states such as stress, confusion, boredom and frustration.

Within literature on student and human teacher interaction, [14] applied theory of dynamic systems to model real time teacher-student interactions using videotaped classroom sessions. Quality of interaction was rated and analyzed in terms of content, structure and complementary. [8] uses turn level audio features and contextual information to predict students' high level affect states using a human-human tutoring dialogue corpus.

There are several aspects in which this study differs from relevant prior work: (1) Instead of using computer tutor, we are interested in an "unplugged" scenario where the child is interacting with a real human teacher. This setup allows us to observe the genuine inter-personal communication dynamics which is not available when interacting with a computer tutor. Specifically, help seeking behaviors, a well studied phenomenon with computer tutors, are generalized into Intent-To-Connect (ITC) behaviors manifested by either subtle cues such as eye contacts or head pose changes, or explicit verbal help requests. ITC behaviors carry a richer meaning that exceeds the conventional cognitive support oriented "help seeking". Instead, ITC behaviors can also be used to signal emotional connection for other purposes such as "comfort seeking" or "joy sharing"; (2) The subject in this study is a child at young age. Since children at this age often are not exposed to the social pressure to hide negative emotions such as frustration, this allows observing their emotions with high fidelity, though it also presents unique detection challenges since the frequent baseline body movement are more frequently observed in young children; (3) The problem solving tasks in this study call for the child to take an active role in open exploration, with support from adult only when needed, whereas other studies typically consider a specific task such as cognitive skill training [15]. Consequently, we expect to observe non-baseline affect states at higher level of frequency and intensity; (4) With a few exceptions, most of the existing work relies on signals from a single modality, while this study attempts to integrate multimodal signals available from audio and video data.

```
At a round table there are chairs placed with the same
distance between them.  They are numbered consecutively
1, 2, 3, ?.  Peter is sitting on chair number 11, di-
rectly across from Chris, who is sitting on chair num-
ber 4.  How many chairs are there at the table?
A) 13     B) 14     C) 16     D) 17     E) 22
```

Figure 1: An example of a Math Kangaroo problem

3. DATASET AND USER STUDY

We collected video recordings of one-to-one problem solving sessions between a 9-year-old boy (a third grader) and his mom (the first author of this paper) as his teacher. We chose this setup because this mom and son has worked together on math problem solving for a few years. As result, the mom is used to picking up and reacting optimally to child's behaviors. This is the closest to the desirable model of the "ideal teacher" as we described earlier.

In each of multiple sessions, the child was asked to solve one challenging math problem. We selected the problems from Math Kangaroo[1], an annual international math competition for students in K-12. Using interesting but challenging problems, the goal of this competition is to stimulate students' interest in math problem solving. There are 24 problems in each competition, divided into three sections with gradual increase of difficulty. The problems for this study were selected from the most difficult set of levels 3 and 4 competition geared towards students in third and fourth grades. Those problems assume basic arithmetic skills and background knowledge at the child's grade level. Figure 1 shows an example of a problem used in the study. In all of the sessions, mom tried to optimize the experience of the child by balancing the goal of reducing frustration and providing sufficiently stimulating challenge.

The videos were captured in a home environment using a Logitech 1080P webcam with an integrated microphone. The positions of mom and child make it possible to capture child's non-verbal behavior cues such as head pose and gaze changes when he intends to connect with mom. Both audios and videos were captured for child, whereas only voice was recorded for mom. We recorded a total of 21 sessions, accumulating 141 minutes of raw video with mean length of 6.4 minutes per session, with longest session lasting 14.6 minutes and the shortest only about 2 minutes. In most of the recordings, the child ended with a joyful mood and a sense of accomplishment.

All recordings were manually annotated in ELAN [2][3] for voice activity at utterance level of child and mom. We also annotated child's non-verbal ITC behaviors using cues such as head turn and eye contact as well as verbal cues. Annotation included timestamps of start and end of events. Frame-

[1]www.mathkangaroo.org

[2]http://tla.mpi.nl/tools/tla-tools/elan/, Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands

Table 1: The affective state and problem solving stages and their behavior cues F(facial), HP(head pose), Voc(vocal), Ver(verbal)

Affects	Problem Understanding	Planning	Execution
Confusion	F+Ver		
Frustration		F+HP+Voc	F+HP
Joy		F+Ver	
Engaged			Voc+HP
Disengaged	HP	HP	HP

by-frame emotion states were extracted using FACET Software Development Kit[3]. Head pose and gaze features were extracted using OpenFace framework toolkit[4] [1]. In addition, acoustic features were extracted using COVAREP toolkit (version 1.3.2) [4] every 10ms.

4. QUALITATIVE ANALYSIS

4.1 Problem solving stages and affective states

In his famous book "How to solve it" [9] , the mathematician Gorge Polya proposed four stages of problem solving, a framework widely used in today's math problem solving instructions. In this study we adapt it into a three-stage framework without the last reflection stage, including "problem understanding", "planning" and "execution". Table 1 lists the most likely affective states as well as plausible behavior cues at each problem solving stage based on qualitative analysis of video recordings. Those cues were used to guide the annotation of events as well as informed the feature design for the automated analysis.

There are several "landmark" behavior cues that could be used to identify problem solving stages and transitions. During problem understanding stage, the child reads the problem and asks clarification questions when necessary. The child often ends this stage by saying "okay". Afterwards, the child might be stuck at the planning stage with no idea as for how to proceed, or go on smoothly with a brief planning stage, or in rare cases dive right into the implementation stage. During the implementation stage, the child is often engaged, with his head down, writing on paper, either silently or with fast paced talking suggesting a "flow" experience. After one attempt, he may succeed at solving the problem, or he could find that his answer is obviously wrong.[5] In those cases, he needs to re-enter into the planning stage to find alternative solution, or rework the original plan. The process ends when the correct answer is confirmed in which case the child often exhibits positive emotions such as excitement and joy.

4.2 Interpersonal communication dynamics

The problem solving sessions can be highly interactive between mom and child: the child actively verbalizes his problem solving process and frequently connects with mom through verbal and non-verbal cues which we call "intent-to-connect" behaviors, or, ITC. Verbal ITC cues refer to explicit request for help or questions, while non-verbal ITCs are subtle cues of head pose and/or gaze change.

ITC may carry multiple different meanings, which calls for differentiated responses to achieve best learning outcomes. According to her interpretation, mom's response to ITC may serve a purely cognitive support purpose such as providing scaffolding, or, as in most cases, providing emotional support in the form of "back channel" signals such as "yes", "good", "good thinking". Given the many subtle variation of ITCs that can be considered in modeling response, it is desirable to take into account contextual information such as problem solving stages and emotion states in order to infer the true intent of an ITC.

Figure 2 provides an overview of the events of an example session that illustrates the interplay between interpersonal communication dynamics, including voice activity events (mom's talk and child's talk) and child's ITC behaviors, within the context of problem solving stages transitions and emotion states. As shown in the plot, the session started with the problem understanding stage (1) that is characterized by child's monologue while reading the problem followed by a brief period of pause and thinking. At the same time, confusion and frustration began to kick in (A), after which mom started to intervene by explaining the problem (2), then child entered planning and execution stage (3) that lasts about 3 minutes. Then, at 1 minute into this process, child said "I didn't get it" with head turn, and mom offered help by asking "Do you need help?". However, the child did not take the offer and kept working on his own. Towards the end of this phase, the child exhibited positive emotion of joy. Then mom discovered that child is on the wrong path, so she intervened (4) and the two worked together to correct the error during which time the child showed brief moments of frustration and confusion (C). Afterwards, the session moved into the problem solved stage (5), the child revealed a spike of surprise and moderate joy (D).

5. QUANTITATIVE ANALYSIS

In this section, we present an analytic framework developed to characterize and understand the interplay between dynamics of emotional states as well as interpersonal communication. We first present a method to quantify the relationship between ITC and mom and child's talk. We then present results from analysis of videos using emotion and interaction features. We end this section with predictive modeling of mom's response using multimodal features.

5.1 Interpersonal communication dynamics

5.1.1 Event intensity metric

We use event intensity metric to characterize temporal patterns of intensity of a specific type of event (e.g. child's talk). This metric takes into account both the frequency and duration of an event. To compute the metric, we first convert the annotated duration of the events into discrete sequences sampled uniformly at interval of every 20ms. Binary flag of 1 is assigned to intervals of the event's occurrence and 0 otherwise. A moving sum is then computed from a window centered at the time of interest. The resulting time series of the moving sum of thusly assigned binary flags characterizes

[3]www.emotient.com
[4]https://github.com/TadasBaltrusaitis/OpenFace
[5]Since the problems are formulated as multiple-choice questions, if the answer is not any of the choices provided, then it must be wrong

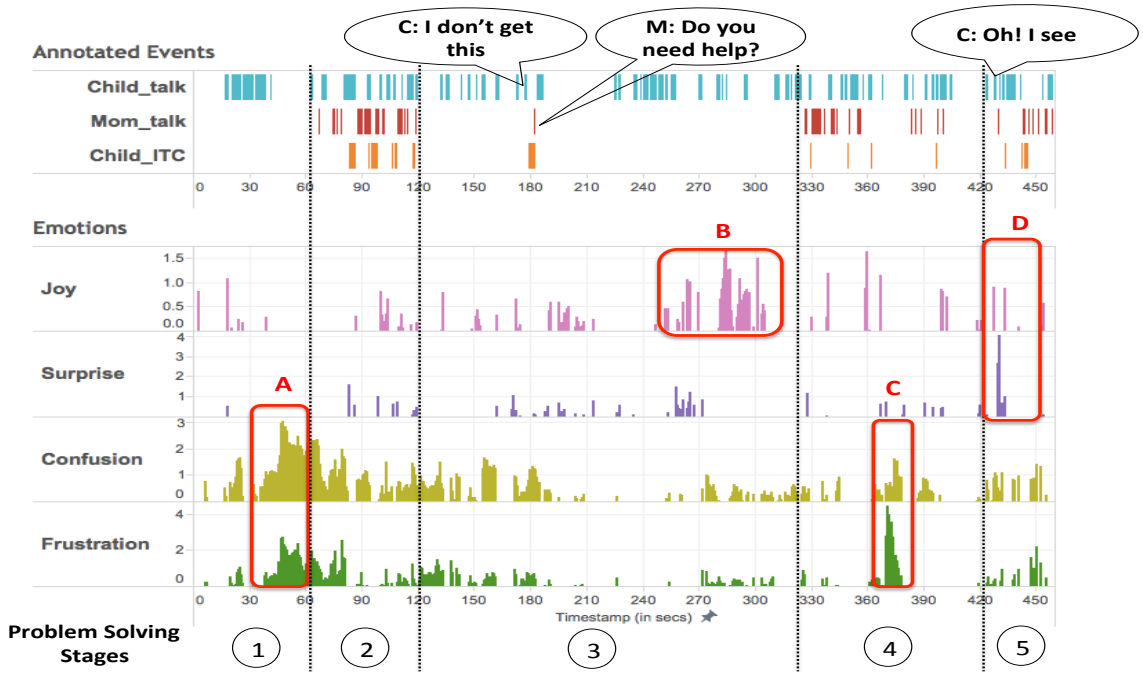

Figure 2: Timeline of annotated events within the context of problem solving stages and affective state transitions. Problem solving stages: (1) problem understanding (2) mom's intervention (3) planning and execution (4) mom's intervention (5) solved ; Emotion states: (A) confused and frustrated (B) joy (C) confused and frustrated (D) joy and surprise; Dialogue legends: C: child; M: mom

temporal intensity distribution of the events. The width of the window determines temporal resolution and smoothness of the temporal patterns.

5.1.2 *Floor sharing metrics*

We characterize temporal patterns of floor sharing between mom and child using normalized metrics of event intensity of mom's talk and child's talk as described above. The formula for mom's sharing of conversation at time stamp t is given as:

$$Mom\_Talk\_Share(t) = \frac{Mom\_Talk(t)}{Child\_Talk(t) + Mom\_Talk(t))} \quad (1)$$

This metric is useful to identify periods of time when mom's intervention dominates or vice versa. Figure 3 shows temporal distribution of floor sharing patterns for each video sorted by video length. It seems apparent that in short videos (presumably representing easy problems), mom did not talk much. However, longer videos often involve larger proportion of mom's talk. It is also interesting to observe that mom's talk often occurs in batches, presumably at the time when child gets stuck so that elaborate explanation is necessary.

5.1.3 *Synchronization of voice activity and ITC*

In this section, we describe a method to quantify synchronization between voice activity (mom's talk and child's talk) and ITC. Figure 4 shows two examples with different synchronization patterns. In the left plot, ITC seems to be more synchronized with child's talk, while in the right plot it is more synchronized with mom's talk which suggests child's attention or engagement . We summarize synchronization as the pairwise correlation among these time series. The result is displayed in the scatter plot in Figure 5 in which each video is plotted as a point labeled with its index. As shown, ITC seems to be more correlated with mom's talk than child's talk as seen from the cluster of points in the upper left quadrant of the plot in Figure 5, with a few exceptions (videos 12, 14 and 32) in which ITC seems to be drifted away from mom's talk and correlate more with child's talk. Incidentally, mom intervened significantly in those videos, which suggests child's disengagement may be induced by mom's higher intensity of teaching.

5.2 Video analysis

In this section, we report the results from video analysis by exploring the pairwise statistical correlations among variables related to interaction dynamics (i.e. voice activity and ITC behaviors) and affective states, as well as the outcome measure, i.e. time taken to solve a problem. For each video, we computed the following variables:

1. Interaction dynamics variables
   - Mom/Child talk ratio (mom-child): The ratio of the accumulative duration of mom's talk versus child's talk.
   - ITC rate: The count of ITC, normalized by video length.

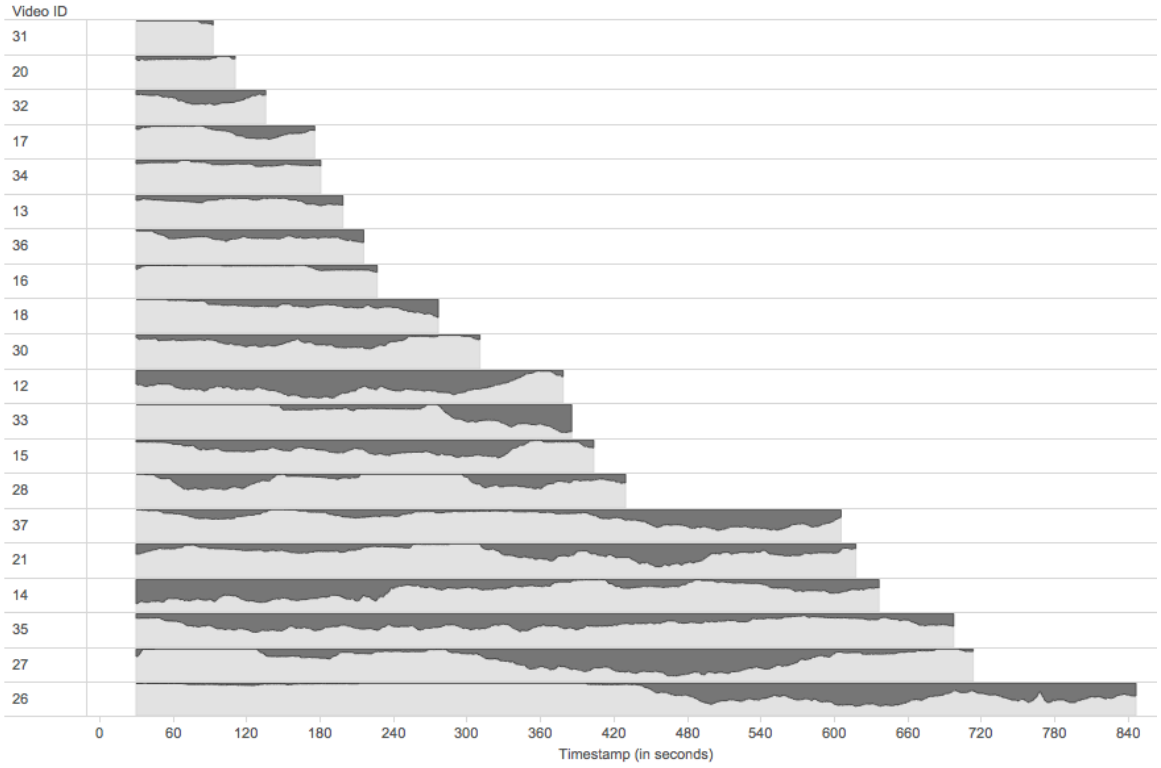

Figure 3: Temporal patterns of floor sharing for each video (dark color: mom, light color: child)

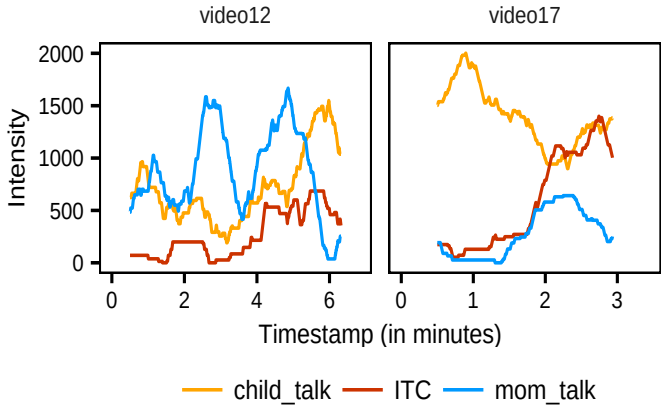

Figure 4: Two example time series plot of events intensity, ITC synchronized more with child's talk (left) or with mom's talk (right)

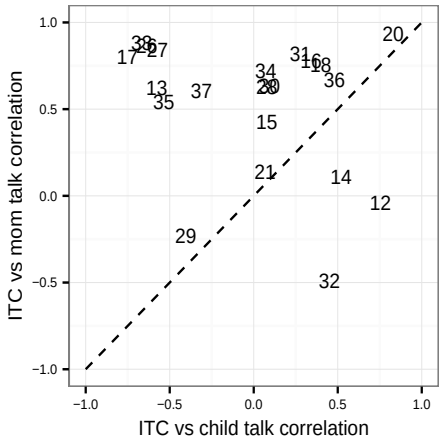

Figure 5: Summary of synchronization : ITC vs child talk (x axis) and ITC vs mom talk (y axis)

- Mom's back channel response rate (mom-BC): Back channel response is defined as a response that lasts less than 2 seconds. This variable represents the count of such response normalized by video length.

2. Affective state variables

   - These are counts of video frames with FACET score greater than 1, normalized by total number of frames during the period of interest for each of the four affect channels including joy, surprise, frustrations and confusion.

In order to further explore the importance of features at the beginning as well as those at the end of a session,we also compute statistical features from two sub-periods of interest: first 30 and last 30 seconds of each video.

We then compute pairwise Pearson correlation among the variables, including outcome. Due to the small number of videos, for each pair of correlations, we performed 1000 iterations of a randomization test [7] under null hypothesis of zero correlation to obtain non-parametric p-values. A sparse graph (Figure 6) is created to summarize the significant correlations among the variables with a p-value cutoff at 5% significance level.

There are several interesting insights that could be derived from this graph. Firstly, there is a significant positive correlation between initial frustration or confusion and the time taken to solve a problem. Since the beginning period is likely to be devoted to problem understanding, this suggests difficulty in understanding of the problem is the first obstacle child may face. His struggle in this period is likely to extend over the entire problem solving process. Secondly, there is a positive correlation between mom/child talk ratio and the video length. This suggests that mom intervenes more in case of hard problems which take longer to solve. Thirdly, child's ITC rate is positively correlated with mom's back channel rates which suggests a level of interaction synchrony between the two. Lastly, there is negative correlation between the overall frustration and joy at the ending period, in other words, more frustrating experience is associated with less joy toward the end, and vice versa.

5.3 Predictive modeling of response

In this section, we report the results from machine learning models used to predict the binary label if there is mom's response within 5 seconds for occurrence of an ITC. The following list explains the features used for the predictive model:

1. Voice activity features:

   - ITC co-occurrence: The count of other ITC within time windows of 2, 5 and 10 seconds respectively for each ITC;
   - Overlap statistics: The number of child talk, mom talk and child or mom talk events that are overlapping a given instance of ITC;

Table 2: Performance of the predictive models of mom's response to child's ITC (leave one video out)

Model	AUC mean	Lower bound of CI	Upper bound of CI
LR	0.594	0.557	0.630
NB	0.617	0.581	0.652
SVM	0.519	0.506	0.531

2. Head pose features : Min, max, mean, median of detection success, confidence, tilt, turn, up-down, within 5 seconds surrounding a given ITC;
3. Features from affect detector: Min, max, mean, median of FACET score for each of the emotion categories (joy, surprise, confusion, frustration and baseline) within the 5 seconds surrounding a given ITC. Negative scores are replaced with 0.

We performed a leave-one-video-out cross-validation experiment to evaluate three different classifiers (logistic regression[LR], naive bayes[NB] and support vector machine[SVM]). The Area Under Curve(AUC) score for each classifier is shown in Table 2 with mean values and 95% confidence intervals. Though the overall performance has much room for improvement, all of the three models perform significantly better than random, which suggests there are indeed predictive signals in the features. A better model might need to incorporate features related to the problem solving state, which may be learned using state space method such as Hidden Markov Models or Conditional Random Fields.

6. VALIDATION OF AUTOMATIC RECOGNITION

6.1 ITC and voice activity recognition

In this section we summarize the results from following recognition tasks:

1. ITC recognition using Openface head pose features. For each video, a random sample of 500 positive frames with ITC and 500 negative frames without ITC were selected, and a model was trained using frame-by-frame head pose features (confidence, Tx, Ty, Tz, Rx, Ry, Rz, up-down, turn and tilt) as inputs;
2. Voice activity recognition using features from COVARAP. One classifier built to discriminate between speaker and non-speaker segments; another classifier to discriminate mom's talk and child's talk. For each task, we random select 500 samples from each class from each video.

In those recognition tasks, we experimented with different types of classifiers including logistic regression, support vector machine, decision tree and naive Bayes, and found logistic regression to show overall superior performance as reported in Table 3. We performed leave-one-video-out cross validation and reported mean AUC scores. We also reported per video performance where we build a dedicated classifier for each video and summarized 10-fold cross-validation AUC score across all videos. As expected, leave-one-video performance is worse than the per video performance for both ITC

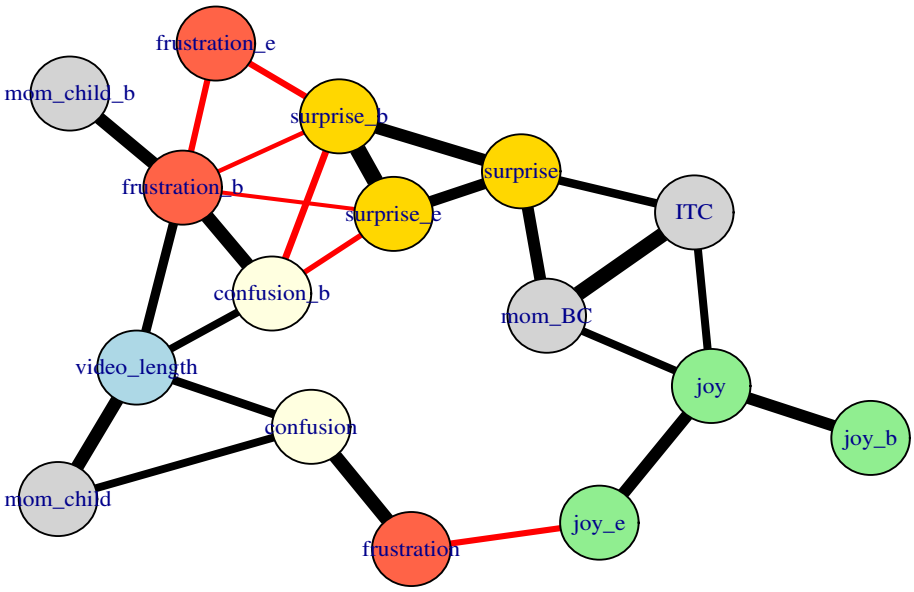

Figure 6: Graph of pairwise correlation of variables. Variable named in the form of "xxx" (e.g. joy) are computed from full length of video; "xxx_b" (e.g. mom_child_b) are computed from first 30 seconds of each video, "xxx_e" (e.g. surprise_e) are computed from the last 30 seconds. Black edges depict positive correlations while red edges represent negative correlations. The width of the edge corresponds to the magnitude of the absolute value of the correlation. The colors of the nodes denote types of variable: Green-Joy, Red-Frustration, Golden-Surprise, Light Yellow-Confusion, Gray-Interpersonal dynamics features, Blue-Outcome

Table 3: AUC scores of models built for ITC and voice activity recognition task

	ITC recognition	Speaker vs. non-speaker	Mom and child talk
Leave one video out CV	0.90	0.81	0.74
Dedicated classifier 10-fold CV	0.92	0.81	0.81

and mom and child talk classification. This suggests that camera and microphone calibration/normalization might have impact on those two tasks, however the speaker and non-speaker classification task seems to be more robust to this issue. Overall, performance of ITC detection is satisfactory, while the voice activity recognition task leaves room for improvement, using a higher quality microphone for each participant might be beneficial.

6.2 Affect detection

In this section, we report validation results for affect labels produced by FACET. We randomly selected 30 top-scored frames (at least 10 seconds apart) from each of the affect class (joy, surprise, frustration, confusion and baseline), and requested labels from two independent annotators who were blinded from FACET labels. Table 4 shows Cohen's Kappa for each affect label (when treated as a binary labeling task) as well as the overall score. As shown, the inter-rater agreement is relatively high for both joy and surprise, though the annotator's agreement with FACET is higher for joy

Table 4: Validation scores of FACET's affect detection (Cohen's Kappa)

Affect	annotator1 vs FACET	annotator2 vs FACET	annotator1 vs annotator2
joy	0.70	0.57	0.73
surprise	0.48	0.43	0.71
confusion	0.30	0.51	0.41
frustration	0.11	0.36	0.44
baseline	0.58	0.42	0.44
overall	0.35	0.46	0.41

than surprise. Confusion and frustration are two of the most challenging affects to detect as compared to joy and surprise, possibly due to the fact that confusion and frustration are easily mistaken for each other, as evidenced by the low inter-rater agreement score. This suboptimal performance may also be attributable to the fact that FACET is trained on faces from general population rather than specifically on young children. A detection algorithm that would incorporate transfer learning and age based customization will possibly improve the performance.

7. CONCLUSION AND FUTURE WORK

In this study, we analyzed data from the 21 video recordings of a nine year old boy while he was working through challenging math problems that demand high order cognitive skills to understand, plan, execute and solve the problems on his own, with only limited and mostly passive support from his mom.

We have shown qualitatively that there are clusters of non-baseline emotions rolling throughout the problem solving process, with the strongest representation from emotion class of joy, surprise, confusion and frustration. This observation confirmed our hypothesis that this type of active exploration indeed facilitates a unique experience of riding an "emotional roller coaster".

We also explored various analytical approaches to characterize the interpersonal dynamics between mom and child as well as the interplay with ITC behaviors. Our video analysis reveals some interesting associations between voice activity, ITC and emotional context.

Lastly, we built a classification model to predict whether there is mom's response within 5 seconds of a given ITC. The recognition task results show promise for automatic annotation of ITC and voice activity in order to scale up the presented analysis. Those findings collectively provide initial evidence for the feasibility of building affect sensitive computer tutor by mining multimodal signals as demonstrated in this study.

The key contributions of this paper include the new framework for fine-grained analysis of affect dynamics during student's interaction with a human teacher, the use of multimodal signals in truly dynamic settings, and demonstration of the utility of the proposed approach to automatically detect behaviors and predict emotions.

We consider multiple thrusts of future work. With the current data set, we envision the following tasks worth consideration: (1) Learn latent dynamic model for problem solving state recognition so that it can be used to improve predictive model of ITC; (2) Explore the possibility of automatic transcription with Automatic Voice Recognition system, and explore sentiment analysis of mom's response; (3) Explore the utility of prosody features of speech signals to complement the current visual-cues based affect detection. Another research direction involves extending this study to more subjects so that inter-subject variation can be observed and modeled. In addition, we would also like to explore the possibility of transferring models learned from one child to another. It is also of interest to explore the correlation between metrics gathered in this study with psychological instruments such as grit scales [6]. Last but not least, we envision our current work to be a foundation for a future tool to guide teachers towards optimal interactions with their students in real time.

8. ACKNOWLEDGMENTS

We would like to thank Liangke Gui for help extracting features from FACET and COVARAP and Tadas Baltrusaitis for helping with Openface features. This work has been partially supported by NSF (1320347).

9. REFERENCES

[1] T. Baltrušaitis, P. Robinson, and L. P. Morency. OpenFace: an open source facial behavior analysis toolkit. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2016.

[2] J. Beck. Engagement tracing: using response times to model student disengagement. *Proceeding of the 2005 conference on Artificial ...*, pages 88–95, 2005.

[3] H. Brugman and A. Russel. Annotating Multimedia/ Multi-modal resources with ELAN. *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*, 2004.

[4] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer. COVAREP âĂŞ A COLLABORATIVE VOICE ANALYSIS REPOSITORY FOR SPEECH TECHNOLOGIES Computer Science Department , University of Crete , Heraklion , Greece Phonetics and Speech Laboratory , Trinity College Dublin , Ireland TCTS Lab - University of Mons , Belgium A. pages 960–964, 2014.

[5] S. K. D'Mello and A. Graesser. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, 20(2):147–187, 2010.

[6] A. L. Duckworth, C. Peterson, M. D. Matthews, and D. R. Kelly. Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology*, 92(6):1087–1101, 2007.

[7] E. S. Edgington. *Randomization Tests*. Marcel Dekker, Inc., 1986.

[8] K. Forbes-riley. Predicting emotion in spoken dialogue from multiple knowledge sources. *Proceedings of the 4th Meeting of the North American Chapter of the Association for Computational Linguistics: : Human Language Technologies,*, pages 201–208, 2004.

[9] P. Gorge. *How to Solve It*. Princeton University Press, 1945.

[10] A. Graesser, Y. Ozuru, and J. Sullins. What is a good question? In M. G. McKeown & L. Kucan, editor, *Threads of coherence in research on the development of reading ability*, pages 112–141. Guilford, New York, New York, USA, 2009.

[11] I. Jraidi, M. Chaouachi, and C. Frasson. A hierarchical probabilistic framework for recognizing learners' interaction experience trends and emotions. *Advances in Human-Computer Interaction*, 2014, 2014.

[12] A. Kapoor, S. Mota, and R. W. Picard. Towards a Learning Companion that Recognizes Affect. *AAAI Fall symposium*, (543):2–4, 2001.

[13] B. Mc, S. D'Mello, B. King, P. Chipman, K. Tapp, and A. Graesser. Facial Features for Affective State Detection in Learning Environments. *29th Annual meeting of the cognitive science society*, pages 467–472, 2007.

[14] H. J. M. Pennings, J. van Tartwijk, T. Wubbels, L. C. a. Claessens, A. C. van der Want, and M. Brekelmans. Real-time teacher-student interactions: A Dynamic Systems approach. *Teaching and Teacher Education*, 37:183–193, 2014.

[15] J. Whitehill, Z. Serpell, Yi-Ching Lin, A. Foster, and J. R. Movellan. The Faces of Engagement: Automatic Recognition of Student Engagementfrom Facial Expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, 2014.