Modeling User Affect from Causes and Effects

Cristina Conati, Heather Maclaren

Computer Science Department, University of British Columbia
2366 Main Mall, Vancouver, BC, V6T 1Z4, conati@cs.ubc.ca

Abstract. We present a model of user affect to recognize multiple user emotions during interaction with an educational computer game. Our model deals with the high level of uncertainty involved in recognizing a variety of user emotions by probabilistically combining information on both the causes and effects of emotional reactions. In previous work, we presented the performance and limitations of the model when using only causal information. In this paper, we discuss the addition of diagnostic information on user affective valence detected via an EMG sensor, and present an evaluation of the resulting model.

1 Introduction

Several studies have reported correlations between student affect and learning (see [1] for an overview) suggesting that educational systems may be more effective if they can trigger appropriate student affective states. Taking student affect into account could be especially beneficial for systems that, like educational (edu-) games, rely heavily on student emotional engagement to be effective. The long-term goal of our research is to devise emotionally intelligent agents for edu-games that model both student affect and learning, and generate adaptive interventions aimed at balancing the two [2]. In this paper, we focus on the model of student affect that we built for one such agent included in an edu-game on number factorization.

The model is based on a framework that uses Dynamic Decision Network (DDN) to leverage information on both the possible causes and the observable effects of the user's affective reaction [2]. In previous work, we built the model's part that reasons from causes to emotions (*predictive model*) and found that it can achieve reasonable accuracy [3,20]. As expected, however, we also found limitations hard to overcome by using causal information only. In this paper, we investigate the instantiation of the part of the model that reasons from effects to emotions (*diagnostic model*) by monitoring the valence of the user emotional state (i.e., positive or negative) via an Electromyography (EMG) sensor. We show that this addition significantly improves model accuracy in detecting strong user emotions during the interaction.

While other work has looked at combining causal and diagnostic information for affect detection (e.g.,[4-6]) to our knowledge ours is the first attempt to provide an explicit comparison between a model that uses both sources vs. a model that uses causal information only. This comparison is important to assess whether it is worthwhile using the potentially more costly and intrusive technology necessary to obtain diagnostic information on user behaviors, as opposed to causal information that can be usually gathered from naturally occurring interaction events. Our approach is also unique with respect to using information on student goals as a source of causal evidence. McGuiggan et al. [6] proposed an affective student model that also includes goal-related information in its assessment. However, in their application goals are explicitly given to students, whereas in ours they are not, requiring the model to do goal recognition.

Another distinguishing feature of our work is that we consider multiple, rapidly changing and possibly overlapping emotions, as often experienced by students playing educational games. In contrast, most work on affect recognition has focused on detecting one specific emotion (e.g., [4-6]), lower-level affective measures of valence and arousal (e.g.,[7,9]) or overall emotional predisposition over a complete interaction (e.g., [10, 11]). One exception is the work by D'Mello et al., [12], which used dialogue features as predictors of student's boredom, confusion, flow and frustration during interaction with a dialogue-based tutoring system. There are three main differences between this work and ours. First, in [12] the target emotions are treated as mutually exclusive, which they mostly are, with the exception perhaps of confusion and frustration. We try to capture potentially overlapping emotions, adding an additional level of complexity to the modeling task. Second, [12] targets longer-term states that some researchers may classify as *moods*, i.e., states that are less specific than simple emotions, less likely to be triggered by a particular stimulus, and lasting [10]. We see these longer-term affective states as being complementary to the more instantaneous emotions we focus on, as we discuss in a later section. Finally, the approach in [12] does not include an explicit representation of causes of affect, thus providing less information than our approach for an agent to decide how to best deal with the student's emotions.

We begin by describing our general framework for affective user modeling. Next, we introduce the edu-game we use as a test–bed for model development. We then summarize the performance of the predictive part of the model, and compare it with an extended model that uses data from an EMG sensor as diagnostic evidence on student affective valence. We conclude by discussing future work.

2 The Affect-Modeling Framework

Figure 1 shows a high-level representation of two time-slices in our DDN-based framework for modeling user affect [2]. Each time slice represents the system belief over relevant elements of the world after an interaction event of interest, such as a user's action (left slice) or an action from an interface agent (right slice). As the figure shows, the network can combine evidence on both the causes and effects of emotional reactions to assess the user's emotional state after each event.

The sub-network above the nodes *Emotional States* is the predictive component of the framework, representing the relations between emotional states and their possible causes as described in the OCC cognitive theory of emotions [15]. According to this theory, emotions derive from one's appraisal of the current situation (consisting of events, agents, and objects) with respect to one's goals and preferences. For instance, depending on whether an event (e.g., the outcome of an interface agent's action) fits or does not fit with one's goals, one will feel either joy or distress in relation to the event. If the current event is caused by a third-party agent, one will feel admiration or reproach toward the agent; if that agent is oneself, one will feel either pride or shame. Based on this structure, the OCC theory defines 22 different emotions.

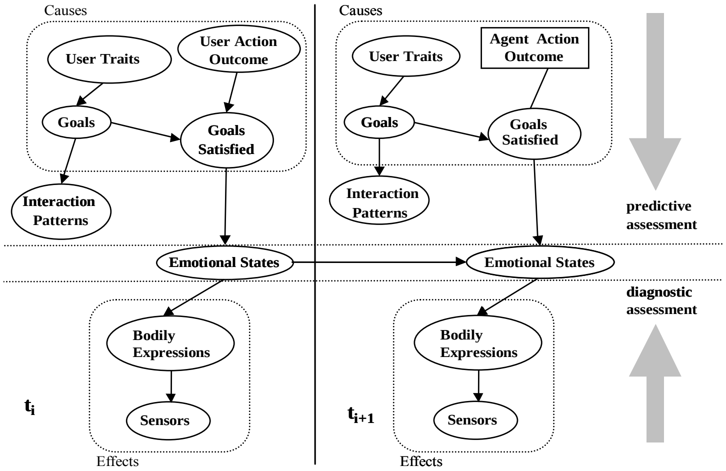


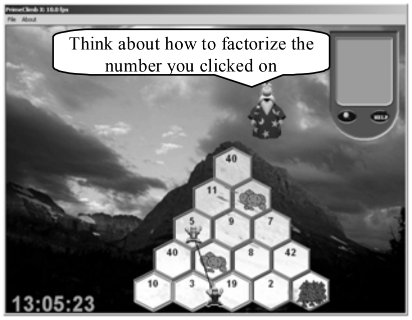
Figure 1: High-level representation of the DDN for affective user modeling

We based our model on the OCC theory because its intuitive representation of the causal nature of emotions lends itself well to devising computational models that can assess not only which emotions a user feels but also why. Thus, an agent's ability to adequately respond to these emotions is enhanced. For instance, if the agent can recognize that the user feels a negative emotion because of something wrong the user has done (*shame* by OCC definition) it may provide hints aimed at making the user feel better toward herself. If the agent recognizes that the user is upset because of its own behavior (*reproach* by OCC definition), it may take actions to make amends. These specific interventions are not possible with approaches that cannot assess the reasons underlying user emotions (e.g. [12]). Another distinguishing feature of the OCC theory is that it mostly captures emotions that are instantaneous reactions to specific events, as opposed to the longer-term affective states such as *frustration*, *boredom*, *confusion* and *flow* targeted by other researchers. We see these states as complementary to those captured by the OCC model in that instantaneous emotions can contribute to creating longer-term affective states. Ideally, an affective user model should be able to capture all these different affective dimensions. However, we decided to focus initially on instantaneous emotions since by acting on them an agent can still impact longer terms affective states.

Our OCC-based DDN includes variables for goals that a user may have during the interaction with a system that includes an interface agent (nodes *Goals* in Figure 1). The events subject to the user's appraisal are the outcomes of the user's or the agent's actions (nodes *User Action Outcome* and *Agent Action Outcome* in Figure 1). Agent actions are represented as decision variables in the framework, indicating points where the agent decides how to intervene. The fit of events with user's goals is modeled by the node class *Goals Satisfied*, which in turn influences the user's *Emotional States* (we call this part of the model *appraisal-subnetwork).* Assessing user goals is not trivial, especially if asking the user about them during interaction is too intrusive, as is the case during game playing. Thus, our DDN also includes nodes (the *goal-assessment subnetwork*) to infer user goals from their interaction patterns and relevant traits (e.g., personality).

The sub-network below the nodes *Emotional States* is the model's diagnostic part, representing the interaction between emotional states and their observable effects. *Emotional States* directly influence user *Bodily Expressions*, which in turn affect the output of *Sensors* that can detect them. Our framework is designed to modularly combine data from any available sensor, and gracefully degrade in the presence of partial or noisy information. We used this framework to build an affective user model for an edu-game on number factorization, which we describe in the next section.

3 The Prime Climb Educational Game



Think about how to factorize the number you clicked on

Figure 2: Prime Climb

Prime Climb is an educational game designed to help 6th and 7th grade students practice number factorization. Two players must cooperate to climb a series of mountains that are divided in numbered sectors (see Figure 2). Each player should move to a number that does not share any common factors with her partner's number, otherwise she falls. Prime Climb includes a pedagogical agent that can both respond to explicit student help requests, and provide unsolicited hints when the student does not seem to be learning from the game [13]. Currently the agent decides when and how to intervene based solely on a probabilistic model that assesses how the player's factorization knowledge evolves during game playing. We have evidence that this knowledge-aware agent can stimulate learning, but we believe that the agent could be more effective if it could respond to user emotions that we observed during game playing. These emotions include feelings generated by the player's performance in the game (i.e., *pride/shame* in the OCC theory) or by the agent's interventions (i.e., *admiration/reproach*, in the OCC theory). Thus, the affective user model we are designing for Prime Climb assesses these emotions, as well as emotions towards game states (i.e., *joy/regret* in the OCC theory) to help the agent take both affect and learning into account when deciding how to act. While other emotions in the OCC model may be relevant, for instance emotions toward one's partner during game play, we decided to add more emotions only after verifying the viability of our approach with the six listed above.

We adopted an iterative design and evaluation approach in building the affective model, starting with the predictive part. In the next section, we briefly summarize the definition of this predictive part and results on its accuracy, to provide the basis for the extensions we describe later.

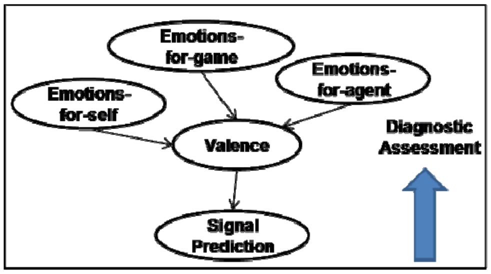
4 Definition and Evaluation of the Predictive Model

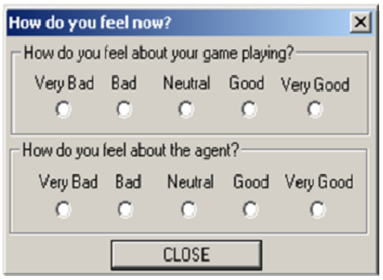
Many components of the predictive part of the Prime Climb's user model were derived from empirical evaluations [3, 20]. Based on student reports after game playing, we identified six high-level non-mutually exclusive goals (*Have Fun, Avoid Falling, Beat Partner, Learn Math, Succeed By Myself and Wanting Help*), represented by separate binary variables in the model. Note that while *Succeed By Myself* and *Wanting Help* intuitively seem mutually exclusive, we observed that they can in fact co-exist for students who express a general preference to succeed by themselves but end up wanting help during especially challenging episodes. We then used interaction data to identify (i) the dependencies among student personality traits, goals and interaction patterns in order to define the goal assessment network; and (ii) the dependencies between the outcomes of student/agent actions and goal satisfaction in order to define the appraisal network. Each of the three emotion pairs in the appraisal network is represented by a binary node (*emotion-for-game, emotion-for-self and emotion-for-agent*, see figure 3 left). This structure was chosen because the two emotions in each pair are mutually exclusive and thus are best represented by a binary node; however, since students may simultaneously feel emotions in the different pairs, a separate node is required to represent each.

An evaluation of the predicted model [3, 20], showed that it performs reasonably well in capturing emotions towards the game (69.5% accuracy), but less so in capturing emotions towards the agent (56.6%), mainly because of problems in capturing regret. In-depth analysis showed that this inaccuracy is due to the model not being able to capture the shifts that some students experience between the goals *Succeed-by-myself* and *Wanting Help* at critical times of game playing. This confusion in turn causes the model to misinterpret how the user appraises the agent's interventions and the impact of user's appraisal on her affect toward the agent. The problem is a consequence of the fact that the model currently represents student goals as static. Modeling how goals evolve during interaction is a form of plan recognition, which is difficult to do without explicitly asking students about their goals Thus, we decided to explore the alternative of improving the model by adding a diagnostic component that captures the player's affective valence via EMG sensors. We look at one sensor, as opposed to directly combining multiple sensors as others have done (see [14] for an overview), because we want to understand the potential of specific sensors as individual sources of affective information in this domain, with the long-term goal of modularly combining evidence in the diagnostic part of the model, depending upon which sensors are available/suitable to use.

5 Adding the EMG Signal to the Affective Model

EMG sensors measure muscle activity by detecting surface voltages that occur when a muscle is contracted. When placed on the corrugator muscle on the forehead, the signal gets excited by movements such as frowning and eyebrow raises. Previous studies (e.g., [16]) report that greater EMG activity in this area tends to be associated with expressions of negative affect. Thus, we decided to experiment with this source of diagnostic evidence, as a way to help the model capture instances of student's reproach. We incorporate this evidence into what we call from now on the *combined model*, as follows. We add two new nodes to each time slice: *Valence* and *Signal Prediction* (see Figure 3, left), both binary. The *Valence* node represents the combined model's overall prediction for the student's affective valence; the *Signal prediction* node encodes whether the EMG signal is positive/negative at a time of interest (as we describe in more detail in a later section). The conditional probability table (CPT) for *Valence* given *Emotional States* is defined so that the probability that valence is positive/negative is proportional to the number of positive/negative emotion nodes. The CPT for *Signal Prediction* given *Valence* represents the probability of observing an EMG prediction of positive or negative valence, given the student's actual affective valence. To instantiate this CPT, we ran a user study to collect both EMG evidence and accompanying affective labels, as described next.




How do you feel now?

How do you feel about your game playing?

Very Bad ○ Bad ○ Neutral ○ Good ○ Very Good ○

How do you feel about the agent?

Very Bad ○ Bad ○ Neutral ○ Good ○ Very Good ○

Figure 3: Adding EMG data to the model (left); emotion self-report dialogue box (right)

Data Collection

Forty-one 6th and 7th grade students from two local schools participated in our study. The study took place in the schools, where for logistic reasons we were limited to a 30-minute session per student. An experimenter placed an EMG sensor on each participant's forehead, and showed a demo of Prime Climb with the emotion self-report mechanism described below. Participants were told that the game included an agent that would try to understand their needs and help them play the game better. The students were encouraged to provide their feelings whenever their emotions changed so that the agent could take them into account when providing help. Next, each participant played Prime Climb with an experimenter as a climbing companion. The experimenter was instructed to play as neutrally as possible, trying to avoid both making mistakes and leading the climb too much. This set up was adopted to avoid the strong emotions toward the partner that we often observed when students play together, given that our affective model currently does not model these emotions.

During game playing, the Prime Climb agent autonomously generated hints to help the student learn from the game, based on the existing model of student learning [13]. At any point during the interaction, students could volunteer information on their emotional states via the dialogue box shown in Figure 3 (right). If students tend not to volunteer self-reports, the dialog box pops up unsolicited, requiring students to input self-reports with a frequency adjusted to balance the amount of data collected and the level of interference generated. The emotion dialog box only elicits information on student emotions towards the game and the agent because dealing with three pairs of emotions turned out to be too confusing for our young subjects. The emotion-self-report approach (which we have used in several previous studies [3, 20]) was chosen because, during interaction with Prime Climb, user emotions are varied, ephemeral and rapidly changing, making it hard for our young users to describe their emotions after the interaction, as it has been done by other researchers. Another commonly used method to obtain affective labels is to have observers post-annotate videos of the interactions based on the users' visible reactions (e.g., [12]). However, when we tried to use this approach we found that observers often had a hard time discriminating among feelings with equal valence in our two different emotion pairs (e.g., reproach toward the agent vs. distress toward the game).

The log files from the study include all relevant game events (e.g., a student's successful climbs and falls, agent interventions), the student's reported emotions and the EMG signals sampled at 32 Hertz. These log files were analyzed to generate a set of datapoints of the form <*affective valence, signal prediction*>*,* as we describe next.

Creating Predictions from the EMG Signal

A datapoint <*affective valence, signal prediction*> is created for each event in the logs that can be associated with an emotion self-report, with value for *affective valence* (positive or negative) derived from that self-report and value for *signal prediction* (also positive or negative) computed by analyzing the EMG signal in the four seconds following the event. The period of four seconds was chosen based on [16], to allow for enough time to detect a response in the signal while avoiding recording the student's reaction to subsequent events. The analysis yielded 196 datapoints, which we used to instantiate the CPT for the *Signal Prediction* node in Figure 3 by calculating the frequencies of the various combinations of *signal prediction/affective valence* value pairs in the data set.

To obtain the values of *signal prediction* for our datapoints, we used the mean of the raw EMG signal as the base for signal analysis, because it is a measure that has consistently shown a reliable mapping with affective valence (e.g., [4,11). The standard method for generating valence predictions from EMG is to compare the EMG signal over the interval of interest against a baseline recorded during a resting period before the experiment. Due to limitations on time with the students, in our study we could not set up an idle "resting time" that we could use as a baseline. Thus, we resorted to using as a baseline the mean of the EMG values recorded during the entire interaction. That is, given a datapoint associated with interface event *e*, the prediction produced by the EMG signal following *e* is computed as:

```
signal prediction(e) = positive
     if mean(EMG_e) < mean(EMG_all)  (1)
negative, otherwise.
```

where *EMG_e* is the set of EMG values recorded during the 4 seconds following *e* and *EMG_all* is the set of EMG values recorded during the entire interaction. Our choice of using the overall signal mean as a threshold for signal prediction is based on the experimenters' observations that many students experienced both positive and negative affect at some point during the interaction. Because negative affect is often associated with greater EMG activity in the forehead muscles [16, 17], the overall EMG mean of a student who experienced both positive and negative affect would be higher than the mean in those intervals where the student did not experience negative affect. This non-standard baseline is bound to misclassify students who experienced only positive affect during interaction. However, when we checked the performance of Equation 1 as a classifier for affective valence on our dataset, we found that this method could still allow us to add useful information to the model (see next section). Thus, we decided to continue with our investigation, by comparing the performance of the combined model with the predictive model described earlier.

6 Evaluating the Combined Model

Each model's performance is assessed via a simulator that replays event logs from Prime Climb interactions with that model. Model accuracy is computed via 100-fold random resampling, a cross-validation method commonly used with limited datasets [18]. We divided the evaluation into two steps. In the first step, we evaluate model performance on 83 datapoints obtained from self-reports that were either clearly positive or clearly negative. These are self-reports in which students indicated a positive (negative) emotion toward both game and agent, or in which one reported emotion was strongly positive (negative) and the other was neutral; we will call these data points *clear-valence* from now on. In the second step, we analyze model performance on the less-investigated assessment of multiple emotions with unclear and possibly conflicting valence, represented in our dataset by 99 self-reports. We excluded from our analysis 14 reports that received neutral answers for both emotion questions. These points are certain to be misclassified by our models, which currently can't represent neutral affect.

The first step above, focusing on clear valence datapoints, is meant to verify whether we can replicate previous results from the literature on using the EMG on the corrugator muscle as valence predictor. These results were mainly obtained with clear valence affective states. As part of this step, we checked the performance of Equation 1 as a classifier for affective valence on the 83 clear-valence datapoints. The method achieved 89% accuracy in classifying datapoints with negative affective valence, indicating that, despite our less than ideal baseline, evidence from the EMG signal may still be a good detector of negative effect and help us improve the model's assessment of Reproach in the presence of clear valence emotions. As expected, Equation1 does not perform as well in classifying positive data points, reaching only 48% classification accuracy. Thus, any positive results obtained with this method should be considered as a lower bound on the potential of including EMG evidence in the Prime Climb model.

We tested model performance on clear-valence datapoints as follows. For each of the 100 folds in the cross-validation, we divided the set of students into a training and a test set of equal size using random selection. The clear-valence datapoints in the training set were used to define the CPT for the *Signal prediction* node in the combined model. The event logs in the test set were run through the simulator, first with the predictive and then with the combined model. For each data-point in the test-set, model prediction was compared with the corresponding self-reported emotion.We used an analogous procedure to test the models on datapoints with ambiguous valence (second evaluation step above).

Table 1: Accuracies on clear-valence data (†significantly different from predictive model)

	Accuracy % (Clear-Valence Data Points)			N
	Predictive	Combined	Baseline	
Joy	74.80	79.10†	100	74
Distress	53.48	56.70	0.00	5
Macro Avg.	64.14	67.90†	50.00	
Micro Avg.	72.58	76.92†	91.03	
Admiration	83.49	83.18	100	67
Reproach	39.11	63.02†	0.00	9
Macro Avg.	61.30	73.10†	50.00	
Micro Avg.	76.86	79.2†	84.67	

For each emotion pair, we report model accuracy on both the positive and the negative emotion. Since there is a trade-off between these measures, we also need a measure of combined accuracy. Two common choices include micro-average (the percentage of cases correctly classified over all the test instances) and macro-average (the average of the accuracies for each class). Micro-averages are a commonly used measure of classification accuracy, but they produce a somewhat biased picture in the presence of classes with unbalanced size, because the accuracy over classes with few data-points is overshadowed by the accuracy over larger classes. Macro-averages are considered an adequate way to overcome this short-coming (e.g., [19]); they give fair weight to classes with few instances, when it is important that their instances are correctly detected. This is exactly the case in our work: we often see far fewer negative than positive data points (see Table 1), however, it is crucial for the model to detect these negative emotions since they may compromise the player's overall attitude towards the game. Given the nature of our dataset, macro-averages are a more appropriate measure of the model's overall accuracy, and so we will base our discussion on this measure. However, we report both micro- and macro- average for sake of completeness. We also report the performance of a standard baseline, i.e., a model that always predicts the most likely emotion. However, comparison with this baseline is not very meaningful, given the unbalance in our data. The baseline tends to have a high micro-average, because its perfect performance in capturing positive emotions off-balances its non-existent performance on the negative data points. Still, it would be hard to argue for a model that cannot capture negative affect, as reflected by its poor macro-average, consistently lower than those of both affective models.

Results

Clear-valence datapoints. We start by comparing the predictive and combined models on the clear-valence dataset. All measures of statistical significance are based on a two-tailed paired-samples t-test with $\alpha = 0.05$. As Table 1 shows, the combined model performs significantly better than the predictive model for Joy ($t(99)=4.59$, $p<.001$, $d=.92$) and Reproach ($t(99)=8.84$, $p<.001$, $d=1.78$). The increase in Reproach results in a significant increase of the model's macro average for emotions towards the agent (t(99)=8.62, p<.001, d=1.38), with a large effect size. The increase for Joy results in a significant increase of the model's macro average for emotions towards the game (t(99)=2.11, p=.038, d=.26), with small effect size.

Thus, we achieved our goal of improving the assessment of reproach by including diagnostic evidence in the model, at least for clear-valence datapoints. Essentially, when the student feels strong reproach and has no other conflicting emotion, the strong evidence of negative affect from the EMG sensor propagates to the *emotion-for-agent* node, overriding the more indirect (and incorrect) goal-based assessment from the causal part of the model. It is also encouraging to see that the poor performance of the EMG as a classifier for positive valence (see section 5) did not transfer to the combined model. In this case, the limitations of the EMG signal in detecting positive affect are compensated by the predictive model, with no negative, and actually some positive impact, on accuracy.

Ambiguous-valence datapoints. Accuracy results on the ambiguous-valence datapoints are not as encouraging. As Table 2 shows, there are significant decreases in both Joy (t(99)=-10.87, p<.001, d=-2.19) and Distress (t(99)=-2.55, p<.001, d=-.51). There is no relevant change for Reproach. The model's macro and micro-average for emotions towards the agent increase significantly (t(99)=8.03, p<.001, d=.84) because of an increase in admiration, but they are still below baseline accuracy. Although these results are disappointing, they are not surprising. Previous work showing the effectiveness of EMG in predicting valence usually investigated the mapping between EMG and clear valence emotions. Our ambiguous-valence data points, on the other hand, correspond to states were students reported mild or even conflicting emotions. Mild emotions are likely to generate more subtle facial expressions, difficult to capture by monitoring only the movements of the corrugator muscle. As for conflicting emotions, their overall valence may depend on which of the emotions involved is stronger. In our model, any evidence of overall valence coming from diagnostic data is propagated upwards to all the emotion pairs, biasing them in the same direction and causing a misclassification for any pair that had opposite valence, unless there is strong evidence coming from the causal model to correct the trend.

Table 2: Accuracies on ambiguous-valence data
(†significantly different from predictive model)

	Accuracy % (Clear-Valence Data Points)			N
	Predictive	Combined	Baseline	
Joy	83.66	75.15†	100	51
Distress	43.82	38.72	0.00	15
Macro Avg.	63.74	56.44†	50.00	
Micro Avg.	75.69	66.71†	79.35	
Admiration	58.58	71.70†	0.00	28
Reproach	25.36	25.11	100	33
Macro Avg.	42.11	48.41†	50.00	
Micro Avg.	42.67	49.10†	51.57	

The problem with capturing mild emotions is likely to be solved by increasing the model's ability to capture valence-related behaviors with the addition of other sensors linked with affective valence (e.g. an heart-rate monitor, EMG sensors on the frontalis muscle, or on the zygomatic major muscle). This solution, however, is unlikely to ease the problem with capturing conflicting emotions, because the problem is due to valence not carrying enough information to tease out the individual emotions. In this case, the only viable solution seems to be improving the accuracy of the diagnostic model, the only component that can provide direct information on the user's individual emotions.

Discussion and Conclusion

In this paper, we evaluated the addition of diagnostic information to an affective user model to detect players' emotions while interacting with Prime Climb, an edu-game for number factorization. The model combines information on causes and effects of users' affect to recognize multiple, possibly overlapping and rapidly changing emotions. While there are approaches to recognizing one specific user emotion or emotion valence/arousal, ours is one of the few models targeting the recognition of multiple emotions, and is unique in dealing with possibly overlapping emotions.

We have presented results of comparing a model that uses only causal information on game state, against a model that also includes information on user affective valence detected via an EMG sensor placed on the user's forehead. While approaches combining diagnostic and predictive inference have received substantial attention, our contribution is an ablation study that compares two versions of the model to understand the effects of each source of evidence. We showed that EMG information can significantly improve the model's accuracy in cases where the students' affective state has clear valence. Given that our method for signal processing relies on a less-than-ideal baseline, this result is a lower bound of what this approach can achieve. We also discussed the limitations of our approach in the presence of emotions with milder or conflicting valence, and presented two avenues of future work to overcome them. In particular, we are planning to (i) include other sources of valence information to detect emotional states expressed more subtly; and (ii) explore ways to capture the evolution of player goals during game playing, to refine the model assessment of conflicting emotions. We also plan to add sensors to capture arousal, so that the agent can gauge the actual impact of the user's emotions on game playing and learning. Other future work includes adding to the model the capability of assessing emotions toward a partner, and showing the effectiveness of an agent that has detailed information on user affect.

References

1. Craig, S.D., Graesser, A. C., Sullins, J., & Gholson, B. (2004). Affect and Learning: An Exploratory Look into the Role of Affect in Learning with AutoTutor. *Journal of Educational Media*, 29.
2. Conati C. (2002). Probabilistic Assessment of User's Emotions in Educational Games . *Journal of Applied Artificial Intelligence,* 16 (7-8), p. 555-575.
3. Conati C and Mclaren H. (2005). Data-driven Refinement of a Probabilistic Model of User Affect. *Proceedings of UM2005 User Modeling: Proceedings of the Tenth International Conference.*

4. Bosma, W. and E. André (2004). Exploiting Emotions to Disambiguate Dialogue Acts. *IUI '04, Int Conf. on Intelligent User Interfaces*.
5. Kapoor, A. and R.W. Picard. (2005). Multimodal Affect Recognition in Learning Environments. *13th Annual ACM Int Conf. on Multimedia*.
6. McQuiggan, S., S. Lee, and J. Lester (2007). Early Prediction of Student Frustration. *2nd Int. Conf. on Affective Computing and Intelligent Interactions*.
7. Zakharov, K., A. Mitrovic, and L. Johnston. (2008). Toward Emotionally Intelligent Pegadogical agents. *ITS08, Intelligent Tutoring Systems, 9th Int. Conf*.
8. Thayer, R. E. (1989). *The biopsychology of mood and arousal*. New Yok, NY: Oxford University Press.
9. Prendinger, H., J. Mori, and M. Ishizuka (2005). Recognizing, Modeling, and Responding to Users' Affective States. *UM'05 10th Int. Conf. on User Modeling*.
10. Yannakakis, G.N., J. Hallam, and H.H. Lund (2008). Entertainment Capture through Heart Rate Activity in Physical Interactive Playgrounds. *User Modeling and User-Adapted Interaction*,18(1-2).
11. Mandryk, R.L., K.M. Inkpen, and T.W. Calvert (2006). Using Psychophysiological Techniques to Measure User Experience with Entertainment Technologes. *Journal of Behavior and Information Technology*, 25.
12. D'Mello, S.K., Craig, S.D., Witherspoon, A. W., McDaniel, B. T., and Graesser, A. C. (2008). Automatic Detection of Learner’s Affect from Conversational Cues. *User Modeling and User-Adapted Interaction*, 18(1).
13. Conati C. and Zhao X. (2004). Building and Evaluating an Intelligent Pedagogical Agent to Improve the Effectiveness of an Educational Game. *Proc. of IUI '04, Int. Conf. on Intelligent User Interfaces*, Island of Madeira, Portugal, p. 6-13.
14. Kim, J. and E. André. Emotion (2006). Recognition Using Physiological and Speech Signal in Short-Term Observation. *PIT'06, Perception and Interactive Techologies*, Kloster Irsee, Germany: Springer.
15. Ortony, A., G.L. Clore, and A. Collins (1988). The Cognitive Stucture of Emotions. 1988: Cambridge University Press.
16. Lang, P., Greenwald, M., Bradley, M .& Hamm, A. (1993). Look at Pictures: Affective, Facial, Visceral, and Behavioral Reactions. *Psychophysiology, 30*.
17. Scheirer, J., Fernandez, R., & Picard, R. W. (1999). *Expression Glasses: A Wearable Device for Facial Expression Recognition.* Proceedings of CHI'99, Human Factors in Computer Systems, Pittsburgh, PA.
18. Mitchell, T. (1997). Machine Learning: McGraw Hill.
19. Sebastiani, F., (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1): p. 1-47.
20. Conati C and Maclaren H. (2009). Empirically Building and Evaluating a Probabilistic Model of User Affect. *User Modeling and User-Adapted Interaction* (to appear).