Modeling Confusion: Facial Expression, Task, and Discourse in Task-Oriented Tutorial Dialogue

Joseph F. Grafsgaard, Kristy Elizabeth Boyer,
Robert Phillips*, and James C. Lester

Department of Computer Science, North Carolina State University
Raleigh, North Carolina, USA
*Dual affiliation with Applied Research Associates, Inc.
Raleigh, North Carolina, USA
{jfgrafsg, keboyer, rphilli, lester}@ncsu.edu

Abstract. Recent years have seen a growing recognition of the importance of affect in learning. Efforts are being undertaken to enable intelligent tutoring systems to recognize and respond to learner emotion, but the field has not yet seen the emergence of a fully contextualized model of learner affect. This paper reports on a study of learner affect through an analysis of facial expression in human task-oriented tutorial dialogue. It extends prior work through in-depth analyses of a highly informative facial action unit and its interdependencies with dialogue utterances and task structure. The results demonstrate some ways in which learner facial expressions are dependent on both dialogue and task context. The findings also hold design implications for affect recognition and tutorial strategy selection within tutorial dialogue systems.

Keywords: Affect, tutorial dialogue, tutorial strategies.

1 Introduction

Recent years have seen a growing recognition of the role that affective computing can play in providing students with highly adaptive and effective learning experiences [1,2]. These investigations highlight the importance of affect in tutorial interactions and have contributed to an emerging understanding of learner emotions [2-7]. To date, a number of systems have incorporated affect, recognizing and responding to it in pedagogically beneficial ways [8-10]. However, the field has not yet seen the emergence of a contextualized model of affect that explains when learners are likely to experience particular emotions and what the impacts of affective states are on learning outcomes.

This paper presents a novel approach to analyzing student emotion, as evidenced by facial expressions, during computer-mediated human task-oriented tutorial dialogues. In particular, we focus on all occurrences of a specific facial *action unit* [11] that has been shown to correlate with confusion in learning [12,13], as well as with anger, fear, and mental effort in other settings [14,15]. Concentrating on this single, highly relevant facial action unit reveals important interdependencies between facial expression, dialogue, and task structure. We discuss ways in which tutorial

dialogue systems can leverage these contextual models of student affect to inform such behaviors as question asking and adaptive delivery of feedback.

2 Related Work

Research on emotion during learning within the AI in Education community has focused on predictive models of student affect [9,10,13,16,17], affective adaptations within intelligent tutoring systems [1,6,18], and understanding student affect during tutoring sessions [2-5,7]. Prior studies on understanding student affect during learning have aimed to identify the presence and characteristics of student emotions and transitions between them. Confusion and flow have been observed to positively effect learning gains, while boredom has a negative impact [3]. A state of stuck may be an important negative parallel to the state of flow [18]. Learners may transition in particular ways among the emotions of boredom, confusion, curiosity, delight, eureka, flow, and frustration, as shown in several studies [2-4,6].

Facial expressions provide a natural window onto student affect. Automated tracking of facial features and head movement has been shown to predict self-reported frustration [10], as well as confidence, interest, and excitement [9,19]. Studies of facial expression in learning contexts found that learner emotions are discernible through facial features [5,20] and that facial and discourse features diagnose confusion more accurately than gross body language [21]. Particular facial configurations have been found to correlate with learner emotions, and facial *action unit 4* (AU4), the Brow Lowerer, has been most strongly correlated with confusion [12,13].

The current work focuses on the affective state of confusion as evidenced by AU4 and extends previous work by applying a focused manual facial annotation approach to tutoring sessions in their entirety. This paper contributes to the body of empirical results on facial expressions of emotion by examining how the context of dialogue and learning task are associated with student displays of a highly relevant facial action unit, AU4.

3 Corpus and Facial Action Analysis

A corpus of human-human tutorial dialogue was collected during a tutorial dialogue study. Students solved an introductory computer programming problem and carried on computer-mediated textual dialogue with a human tutor. The original corpus consists of 48 dialogues and was previously annotated with dialogue acts and subtask structure [22]. Facial recordings of students were collected using built-in webcams. The tutors were not shown the student facial videos. Video quality was ranked based on how completely each student's face was visible within the frame, and the fourteen highest quality videos were used in this analysis. They have a total running time of eleven hours 55 minutes and include dialogues with three female subjects and eleven male subjects.

The facial videos were annotated manually using the Facial Action Coding System (FACS), which enumerates the possible movements of the face through a set of facial action units [11]. The FACS coders viewed videos from start to finish, pausing at observed instances of AU4 activation (Figure 2). Facial movements were encoded as events with a start frame and an end frame. A certified FACS coder [14] annotated all

fourteen videos. A second certified FACS coder annotated six videos. After the tagging was complete, the sessions were discretized into one-second intervals. Cohen's kappa for inter-coder agreement on AU4 across all one-second intervals was κ=0.86, which indicates very good reliability. Excerpts from the fully annotated corpus are shown in Figure 1. Displays of AU4 were noted during a total of 53 minutes of the approximately 12 hours of video, with high variance across individual students (*min*=0 seconds; *max*=33 minutes).

Excerpt 1		
14:07:03	Tutor:	ok, so that's closer [LUKEWARM FDBK]
14:07:23	Tutor:	but you are currently saying, i want the value at position i to be the same as the value at position i + 1 [STATEMENT]
	Student:	BUGGYTASKACTION
14:07:43	Tutor:	instead of wanting the value at position i to be one more than the current value at position i [STATEMENT]
Excerpt 2		
17:44:41	Tutor:	okay, good so far [POSITIVEFDBK]
17:44:47	Tutor:	except there's a typo in that loop condition [NEGCONTENTFDBK]
	Student:	CORRECTTASKACTION
17:45:08	Tutor:	now that we have n, how can we change the loop condition for c? [ASSESSINGQUESTION]
	Student:	FACIALEXPRESSION: AU4
Excerpt 3		
15:43:26	Tutor:	well you have one error, it's underlined in red [NEGATIVECONTENTFDBK]
	Student:	FACIALEXPRESSION: AU4, CORRECTTASKACTION
15:43:35	Tutor:	yup [POSITIVEFDBK]
	Student:	FACIALEXPRESSION: AU4, INCOMPLETETASKACTION
15:44:01	Tutor:	so far so good, let's fix the return statement and then we should probably check if the first two problems work by running it [LUKEWARM FDBK]

Figure 1. Tutoring session excerpts

Figure 2. Student displays of facial action unit 4 (AU4, Brow Lowerer)

The annotated facial action data were merged with the previously annotated dialogue acts and task actions to form a chronological record of task actions, dialogue, and student displays of AU4 that were then used to empirically explore dependencies between events. Table 1 displays the relative frequencies for student task action tags that occurred at the same time as AU4. Statistically significant differences are in bold.[1] Students were significantly less likely to display AU4 while engaging in on-track, INCOMPLETE task actions. Students were also more likely to display AU4 during a BUGGY or CORRECT task action, and less likely during DISPREFERRED task actions (which technically meet the problem specifications but circumvent the pedagogical goals of the task), though these differences were not statistically significant.

Table 2 displays the analogous relative frequencies of tutor dialogue acts across all sessions compared with the relative frequencies of only those dialogue acts that were followed by a student display of AU4 within ten seconds. The results indicate that students were significantly less likely to display AU4 immediately following tutor EXTRA-DOMAIN moves, LUKEWARM FEEDBACK, and QUESTIONS.

Table 1. Student AU4 during task actions[2]

Student Task Action	Relative Freq. of Task Action *(stdev)*[3]	Rel. Freq. of Task Action With Student AU4 Present *(stdev)*	*p*-value (paired *t*-test, *N*=13)
BUGGY	0.578 *(0.156)*	0.602 *(0.333)*	0.7808
DISPREFERRED	0.057 *(0.106)*	0 *(0.001)*	0.0773
INCOMPLETE (ON-TRACK)	**0.154 *(0.143)***	**0.082 *(0.151)***	**0.0076**
CORRECT	0.809 *(0.183)*	0.856 *(0.176)*	0.2943

Table 2. Student AU4 following tutor dialogue acts

Tutor Dialogue Act	Relative Freq. of Tutor Act (*stdev*)	Rel. Freq. Of Tutor Act With Student AU4 w/in 10 Sec. (*stdev*)	*p*-value (paired *t*-test, *N*=11)
ASSESSING QUESTION	0.097 *(0.075)*	0.177 *(0.233)*	0.2510
EXTRA DOMAIN	**0.055 *(0.057)***	**0.009 *(0.020)***	**0.0227**
GROUNDING	0.063 *(0.081)*	0.020 *(0.052)*	0.2007
LUKEWARM CONTENT FDBK	0.031 *(0.025)*	0.012 *(0.028)*	0.0680
LUKEWARM FDBK	**0.023 *(0.021)***	**0 *(0)***	**0.0047**
NEGATIVE CONTENT FDBK	0.094 *(0.053)*	0.153 *(0.191)*	0.3117
NEGATIVE FDBK	0.016 *(0.013)*	0.006 *(0.014)*	0.0819
POSITIVE CONTENT FDBK	0.032 *(0.030)*	0.051 *(0.107)*	0.55
POSITIVE FDBK	0.150 *(0.069)*	0.162 *(0.317)*	0.9040
QUESTION	**0.049 *(0.060)***	**0.004 *(0.012)***	**0.0363**
STATEMENT	0.391 *(0.119)*	0.406 *(0.254)*	0.8221

[1] Because of the limited sample size and the goal of highlighting trends that warrant future study, a statistical correction for multiple tests was not applied.

[2] Sample sizes *N* reflect only students who displayed AU4 during task action segments (Table 1) or within ten seconds of any dialogue act (Table 2). Else the corresponding probability could not be calculated.

[3] Task action segments may contain multiple tags and therefore do not sum to one.

4 Discussion

These results indicate that student expressions of AU4 are dependent on both the dialogue and task context. This action unit is highly relevant for tutoring because of prior findings that it is correlated with confusion, negative emotions, and mental effort [12-15]. A contextual understanding of this action unit during learning may hold a number of important insights for developing affective tutoring systems.

4.1 Interpretation

After tutor EXTRA-DOMAIN dialogue acts, students were significantly less likely to display AU4, which is consistent with an understanding of EXTRA-DOMAIN moves as conversational and unrelated to the learning task. Students were also less likely to display AU4 following tutor LUKEWARM FEEDBACK, a finding that may at first seem counterintuitive. However, as demonstrated by Excerpt 1 of Figure 1, these tutors often used LUKEWARM FEEDBACK to encourage students. Finally, students were less likely to display AU4 immediately following a tutor QUESTION. This finding may also seem counterintuitive given the expectation that question answering may induce confusion, or at least require mental effort, on the part of the student. However, the lack of this facial expression following tutor questions is consistent with a prior observation that the non-expert tutors in this corpus rarely posed deep reasoning questions, but instead tended to ask shallow questions that could be answered quickly [23]. We hypothesize that when working with expert tutors, the statistical relationship between tutor questions and student expressions of AU4 may be reversed.

Some other trends warrant discussion although they did not display statistically significant relationships. For example, tutor ASSESSING QUESTION dialogue moves were more likely to be followed by student AU4 (Figure 1, Excerpt 2). Such questions ask students to reflect on what they already know. For novice students, being asked directly about their knowledge may have produced genuine confusion as they worked to reconcile their emerging knowledge of specific target concepts with their pre-existing knowledge. A similar phenomenon may explain why students were more likely to display AU4 after NEGATIVE CONTENT FEEDBACK (Figure 1, Excerpt 3). Out of all types of feedback, this type may be most likely to place students into cognitive disequilibrium [7].

A statistically significant dependence also emerged between student INCOMPLETE, on-track task actions and AU4. Students were less likely to display AU4 while engaged in these task actions. This finding is likely related to the cognitive-affective state of flow, in which the student is actively focused and making progress on the learning task [24].

4.2 Design Implications

These findings have important implications for the design of intelligent tutoring systems in two dimensions: affect recognition and tutorial strategy refinement. First, affect recognition involves inferring the student's emotional state based on a variety of predictors. *A priori* knowledge that a particular emotional state is more or less likely given the context of the dialogue or task may narrow the state space under consideration by an affect recognition model, potentially increasing efficiency and

accuracy. Second, understanding which student emotions are likely to follow particular tutor moves or problem-solving events can help an ITS select cognitive strategies or affective interventions that are likely to guide students toward affective states conducive to learning.

The results presented here suggest particular ways in which ITSs may leverage knowledge of student affect to provide highly adaptive, affect-informed feedback. For example, the type of question the system poses may directly impact whether the student displays confusion-related facial expressions. Shallow questions are unlikely to produce a cognitive-affective state of confusion, while deep reasoning and assessment questions are more likely to do so. Additionally, when providing feedback on student errors, indirect approaches such as LUKEWARM FEEDBACK may not be sufficient to help novice students become aware of their mistakes or misconceptions. NEGATIVE CONTENT FEEDBACK, in which student errors are explicitly pointed out and a hint is given, appears more likely to accomplish this. Finally, the low probability of observing AU4 during student INCOMPLETE, on-track work emphasizes the importance of sensitivity during possible times of student flow, when a system may choose not to interrupt.

4.3 Limitations

The study has two primary limitations. First, the number of tutoring sessions is small due to the time-intensive manual tagging approach, which for each coder required up to ten hours per hour of video.[4] While manual annotation is time-intensive, it nevertheless serves as a valuable part of achieving complete coverage of tutoring sessions and establishing a foundation on which highly reliable automated techniques can be built. A second limitation lies in the structure of the tutorial dialogue itself, namely, that student utterances are approximately half as numerous as tutor utterances. With a larger number of student utterances, a correlational analysis analogous to that reported in Table 2 could reveal patterns of dependence between student utterances and AU4.

5 Conclusion

Affect plays a central role in learning, and developing a clear understanding of learner emotions can lead to improved affect recognition and adaptation by intelligent tutoring systems. In particular, understanding the interdependencies between facial expression, dialogue, and task structure may hold important insights for designing affective tutoring systems. The work reported here has examined student facial expression, in particular AU4 (Brow Lowerer), during computer-mediated human task-oriented tutorial dialogue. The findings demonstrate that the occurrence of this confusion-related facial expression is dependent on both dialogue and task context. The results indicate that students are less likely to display AU4 immediately following tutor questions, lukewarm feedback, and extra-domain dialogue acts, as well as during incomplete, on-track task actions. Leveraging knowledge of these

[4] This annotation approach considers only a subset of FACS action units. It is significantly faster than full FACS coding, which requires up to sixty hours per hour of video.

patterns can help tutoring systems better recognize student affect and select strategies or interventions that encourage desirable affective states.

This work constitutes a first step toward a comprehensive catalogue of fine-grained facial configurations during learning and their relationships with the tutoring context. Employing a fine-grained approach that focuses on a single facial action unit highlights several important directions for future work. First, facial action coding is a domain-independent approach that can be used to compare the occurrence of student emotions across tutoring corpora. Second, promising work on automatic facial action tagging indicates that in the near future, this type of fine-grained investigation will no longer require manual annotation [25]. Finally, the Core Affect framework [26] provides a promising model by which comprehensive facial annotations and contextual features may be utilized to identify emotions without prior semantic assumptions. Together, these lines of investigation will contribute to the design of the next generation of affectively aware tutorial dialogue systems.

6 Acknowledgements

This work is supported in part by the NC State University Department of Computer Science along with the National Science Foundation through Grants IIS-0812291, DRL-1007962 and the STARS Alliance Grant CNS-0739216. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the participants, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

7 References

1. Woolf, B.P., Burleson, W., Arroyo, I., Dragon, T., Cooper, D.G., Picard, R.W.: Affect-Aware Tutors: Recognizing and Responding to Student Affect. International Journal of Learning Technology 4, 129-164 (2009).
2. D'Mello, S.K., Lehman, B., Person, N.: Monitoring Affect States During Effortful Problem Solving Activities. Int. J. Artif. Intell. Educ. 20, (2010).
3. Baker, R.S.J.d., D'Mello, S.K., Rodrigo, M.M.T., Graesser, A.C.: Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive-Affective States during Interactions with Three Different Computer-Based Learning Environments. International Journal of Human-Computer Studies 68, 223-241 (2010).
4. Lehman, B., D'Mello, S.K., Person, N.: The Intricate Dance between Cognition and Emotion during Expert Tutoring. Proceedings of the 10th International Conference on Intelligent Tutoring Systems. pp. 433-442 (2010).
5. Afzal, S., Robinson, P.: Natural Affect Data - Collection and Annotation in a Learning Context. Proceedings of the International Conference on Affective Computing and Intelligent Interaction. pp. 1-7 (2009).
6. Robison, J.L., McQuiggan, S.W., Lester, J.C.: Evaluating the Consequences of Affective Feedback in Intelligent Tutoring Systems. Proceedings of the International Conference on Affective Computing and Intelligent Interaction. pp. 37-42 (2009).
7. Graesser, A.C., Olde, B.A.: How Does One Know Whether a Person Understands a Device? The Quality of the Questions the Person Asks When the Device Breaks Down. Journal of Educational Psychology 95, 524-536 (2003).
8. D'Mello, S.K., Picard, R.W., Graesser, A.C.: Toward an Affect-Sensitive AutoTutor. IEEE Intelligent Systems 22, 53-61 (2007).

9. Cooper, D.G., Muldner, K., Arroyo, I., Woolf, B.P., Burleson, W.: Ranking Feature Sets for Emotion Models used in Classroom Based Intelligent Tutoring Systems. User Modeling, Adaptation, and Personalization. pp. 135-146 (2010).
10. Kapoor, A., Burleson, W., Picard, R.W.: Automatic Prediction of Frustration. International Journal of Human-Computer Studies 65, 724-736 (2007).
11. Ekman, P., Friesen, W.V., Hager, J.C.: Facial Action Coding System. A Human Face, Salt Lake City, USA (2002).
12. Craig, S.D., D'Mello, S.K., Witherspoon, A., Graesser, A.: Emote Aloud During Learning with AutoTutor: Applying the Facial Action Coding System to Cognitive-Affective States During Learning. Cognition & Emotion 22, 777-788 (2008).
13. McDaniel, B.T., D'Mello, S.K., King, B.G., Chipman, P., Tapp, K., Graesser, A.C.: Facial Features for Affective State Detection in Learning Environments. Proceedings of the 29th Annual Meeting of the Cognitive Science Society. pp. 467-472 (2007).
14. Ekman, P., Friesen, W.V., Hager, J.C.: Facial Action Coding System: Investigator's Guide. A Human Face, Salt Lake City, USA (2002).
15. Cohn, J.F., Zlochower, A.J., Lien, J., Kanade, T.: Automated Face Analysis by Feature Point Tracking Has High Concurrent Validity with Manual FACS Coding. Psychophysiology 36, 35-43 (1999).
16. Conati, C., Maclaren, H.: Empirically Building and Evaluating a Probabilistic Model of User Affect. User Modeling and User-Adapted Interaction 19, 267-303 (2009).
17. McQuiggan, S.W., Lee, S., Lester, J.C.: Early Prediction of Student Frustration. Proceedings of the Second International Conference on Affective Computing and Intelligent Interactions. pp. 698-709 (2007).
18. Burleson, W.: Affective Learning Companions: Strategies for Empathetic Agents with Real-Time Multimodal Affective Sensing to Foster Meta-Cognitive and Meta-Affective Approaches to Learning, Motivation, and Perseverance. MIT Ph.D. thesis (2006).
19. Kaliouby, R., Robinson, P.: The Emotional Hearing Aid: An Assistive Tool for Children with Asperger Syndrome. Universal Access in the Information Society 4, 121-134 (2005).
20. Afzal, S., Robinson, P.: Modelling Affect in Learning Environments - Motivation and Methods. Proceedings of the International Conference on Advanced Learning Technologies. (2010).
21. D'Mello, S.K., Graesser, A.C.: Multimodal Semi-Automated Affect Detection from Conversational Cues, Gross Body Language, and Facial Features. User Modeling and User-Adapted Interaction 20, 147-187 (2010).
22. Boyer, K.E., Phillips, R., Ingram, A., Ha, E.Y., Wallis, M.D., Vouk, M.A., Lester, J.C.: Characterizing the Effectiveness of Tutorial Dialogue with Hidden Markov Models. Proceedings of the 10th International Conference on Intelligent Tutoring Systems. pp. 55-64 (2010).
23. Boyer, K.E., Lahti, W.J., Phillips, R., Wallis, M.D., Vouk, M.A., Lester, J.C.: An Empirically-Derived Question Taxonomy for Task-Oriented Tutorial Dialogue. Proceedings of the Second Workshop on Question Generation. pp. 9-16 (2009).
24. Csikszentmihalyi, M.: Flow: The Psychology of Optimal Experience. Harper-Row, NY (1990).
25. Calvo, R.A., D'Mello, S.K.: Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. IEEE Transactions on Affective Computing 1, 18-37 (2010).
26. Russell, J.A.: Core Affect and the Psychological Construction of Emotion. Psychological Review 110, 145-172 (2003).