Development of an Affect-Sensitive Agent for Aplusix

Thor Collin S. ANDALLAZA, Ma. Mercedes T. RODRIGO

Ateneo Laboratory for the Learning Sciences, Department of Information Systems and Computer Science, Ateneo de Manila University, Loyola Heights, Quezon City, Philippines
{tandallaza, mrodrigo}@ateneo.edu

Abstract. We compared two versions of an affect-sensitive embodied conversational agent for Aplusix, an intelligent tutoring system for algebra. Version 1 of the agent was able to detect and respond to user affect, but it responded too quickly and too frequently. The second version of the agent featured new student models for detecting and responding to student affective states, which is less sensitive compared to the first version. We conducted a field test with students to determine its effect on learning and learning experience in comparison to using Aplusix alone and Aplusix with the version 1 agent. Results show that version 2 provided significantly fewer interventions to engaged students, more evaluations of engagement, fewer evaluations of boredom, and was generally preferred over version 1.

Keywords: affect, Aplusix, embodied conversational agent, intelligent tutoring systems, learning, motivation

1 Introduction

Embodied conversational agents (ECAs) are computer programs that are capable of autonomous action within their environment [14] and are able to interact with users or other agents in a manner similar to human face-to-face conversation [5].

In recent years, the intelligent tutoring systems (ITSs) community has been adopting ECAs to address the non-cognitive aspects of learning. The work of Rebolledo-Mendez et al. [13], for instance, uses an ECA named Paul to address student motivation. Graesser et al.'s [7], [8] AutoTutor responded to the learner emotions (confusion, frustration, and boredom). Leelawong et al. [10] introduced a teachable agent called Betty that allowed students to learn about ecosystems and interact with the agent through the use of concept maps. Finally, Mastuda et al. [12] created SimStudent, an agent that allowed students to hone their knowledge of mathematics through a reversal of roles – the student being the tutor, and the agent being the tutee.

Our study's medium term objective was to create an emotionally intelligent ECA for Aplusix, an ITS for algebra [6]. We wanted to make this ECA correctly identify and respond to student affect, as well as be able to direct and sustain a student's motivation to learn. To do this, we created and field tested student models that will allow the ECA to function properly.

To guide us in our study, we attempted to answer the following questions:

1. What is the appropriate timing of the agent's interventions given an observed affective state?
2. How do we determine the effectiveness of the agent in improving the learning of students?
3. How do we determine the effectiveness of the agent in improving the learning experience of students?

2 Prior Work

Recent studies have paved the way for the creation of an ECA for Aplusix [6], an ITS for algebra. The study of Lagud and Rodrigo [9] examined the relationship of student learning and affective profiles while using Aplusix. Affective states were coded by human observers who worked in pairs for each student. A total of 3,640 observations were taken from the experiment with a high inter-rater reliability: Cohen's κ=0.63 [3], [9]. The study determined that high performing students experienced engaged concentration [3] the most, while low performing students were more susceptible to boredom and confusion.

Another study created detectors for off-task behavior when using Aplusix [4], and another created an initial agent framework for the Aplusix ECA [11]. The most recent study by Andallaza and Jimenez [1] created version 1 of a fully-functional ECA, Grimace, which integrated the findings of the previous three studies.

The model used by Grimace that allowed it to evaluate a student's affective state was a set of threshold values for two features: the number of steps the student took to solve a problem, and the duration (in seconds). The purpose of the thresholds was to define three groups of students: high performing, average, and low performing. Grimace classified students into one of these three groups by observing the students current values and comparing it to the thresholds of each group. The classification then determined what affective state the student was in – high performing students were evaluated as engaged, average as confused, and low performing as bored, based on the patterns observed in previous work [9]. It then fired responses particular to the observed affective state in order to sustain (in the case of engaged concentration) or to change (in the case of confusion and boredom) that state.

Version 1 of Grimace had several limitations. Grimace responded too quickly and too frequently to the student's actions [1]. Since then, we have been refining the student models the agent uses determine affective state and timing of responses.

3 Methods

In this section, we shall describe the Aplusix ITS in greater detail, the steps taken to improve the existing student models, the method used to evaluate these new models prior to field testing, and finally, the procedure for conducting the field test.

3.1 Aplusix

As mentioned earlier, Aplusix (Figure 1) is an intelligent tutoring system that covers various topics in algebra, such as simplification of expressions and factoring. Students are able to tackle each topic in problem sets of 10 items, where each set is of varying difficulty [6], [9].

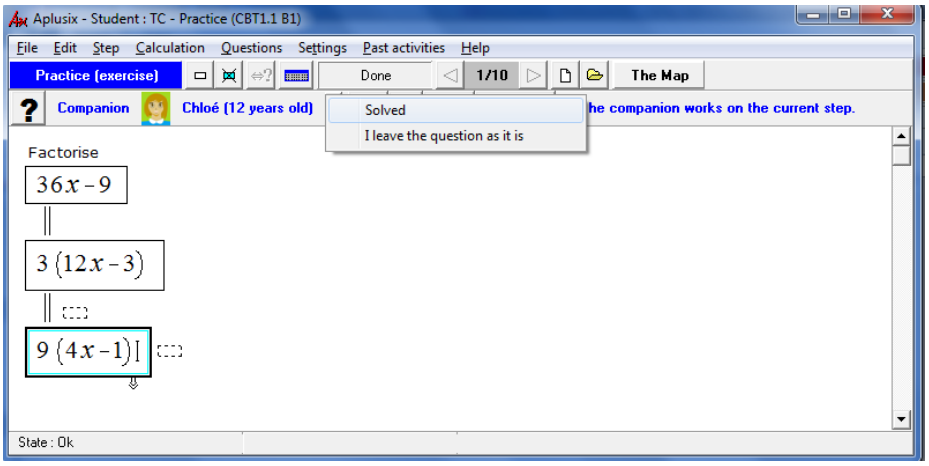

Fig. 1. The Aplusix ITS editor.

Upon choosing a problem set to solve, Aplusix displays an advanced editor that allows students to solve items step-by-step, as if writing solutions using paper and pencil. In addition, the ITS provides a real time visual feedback on their current progress, either through black parallel lines to indicate equivalent steps, i.e. the adjacent steps are moving towards a possible solution, or red parallel lines with an X to indicate an incorrect step. Finally, Aplusix also generates reports on current student progress in the attempt to resolve the problem, as well as having domain-based agents in the form of Chloe, Julien, and Olivia that students may interact with to get hints or the final solution to their current problem.

3.2 Experimentation with the Model Sensitivity

In the latest iteration of developing the agent's models for affect, we reviewed and refined the Aplusix log dataset compiled from Lagud and Rodrigo's [9] study and ran a terciles analysis of the data. This was done by sorting the data and dividing it into three groups of equal size – high, average, and low groups. For each group, we then computed for the minimum, maximum, and mean values that distinguished each group from the others. The result of this analysis generated four new models, each with varying levels of sensitivity. The first two were based on a per student data analysis similar to Lagud and Rodrigo's study [9], and last two used the per problem type analysis approach in Andallaza and Jimenez's study [1]. The features used for the creation of these models were the same features used in the first version of the agent –

number of steps and duration. A summary and comparison of the models in terms of number of steps and duration are shown in Tables 1 and 2 respectively. Note that for models 3 and 4, only those for problem type B1 are shown for presentation.

Table 1. Summary of Threshold Values for the Four Models, Number of Steps

Affective state		Model 1	2	3	4
Engaged Concentration	Min	25	25	25	2
	Max	63	63	51	33
	Mean	47	47	39	21
Confusion	Min	64	64	52	34
	Max	95	95	82	67
	Mean	79	79	67	48
	Intervention Threshold	n/a	133	n/a	n/a
Boredom	Min	100	100	85	68
	Max	578	578	578	1032
	Mean	154	154	147	156
	Intervention Threshold	n/a	200	n/a	n/a

Table 2. Summary of Threshold Values for the Four Models, Duration

Affective state		Model 1	2	3	4
Engaged Concentration	Min	46.8	46.8	34.09	0.1
	Max	114	114	100.99	58.7
	Mean	87.575	87.575	73.764	34.437
Confusion	Min	115.8	115.8	105.42	58
	Max	237	237	221.05	138.7
	Mean	154.46	154.46	147.742	88.886
	Intervention Threshold	n/a	133	n/a	n/a
Boredom	Min	238.8	100	226.82	118.4
	Max	2403.6	578	2403.8	2403.8
	Mean	427.7	154	409.55	373.21
	Intervention Threshold	n/a	200	n/a	n/a

Model 1 generated terciles based on the number of steps and duration taken by the student for the entire 40-minute observation period conducted in Lagud and Rodrigo's study [9]. On the other hand, Model 2 was similar to the first, but with the addition of another value for boredom and confusion called an intervention threshold. This threshold, which was based on the observed incidence of the two affective states,

prevented the agent from firing interventions despite knowing that the student is either confused or bored. It was therefore expected that this version of the models would be the least responsive of the four. Model 3 used student data on the problem type level, where the values were computed for each student based on the entire portion of data where the student worked on that particular problem type. The last version, Model 4, followed the same idea as the terciles used in the first version of the agent, where for each problem type, each student attempt is taken as a discrete transaction. The analysis in effect generated smaller values, which made this version the most responsive among the four models.

3.3 Initial Testing and Evaluation of Student Models

To evaluate how the models functioned prior to field testing, the study modified the existing agent program to allow the agent to read in data from an input file containing Aplusix interaction logs taken from a previous study [9]. Given this set of logs, the agent outputted the student's affective state and determined whether or not an intervention is needed. The output of the test, instead of appearing in the agent console, was printed out in a text file containing the number of interventions, the frequency of interventions, what logs triggered the interventions, and the evaluated affective state for each intervention.

We collected a total of 166,284 Aplusix interaction logs. Table 3 shows the number of interventions, the mean time between interventions, as well as the percentage of time the models evaluated the student to be in a particular state.

Table 3. Descriptive statistics of the interactions per condition.

Model	Number of inter-ventions	Mean time between interven-tions	% Engaged	% Confused	% Bored	% Neutral
1	24,921	21.479	33.47	7.10	38.51	21.47
2	24,914	21.466	33.46	17.13	19.08	21.46
3	31,830	16.384	34.41	4.96	43.56	16.38
4	39,864	10.553	33.07	0.74	55.16	10.55

We observed that out of the four models, only Model 2's evaluations of confusion came closest to the observed incidence for the affective state (15%). Other models gave percentages that were too small compared to the original values taken from human observations. In addition, Model 2 gave the least evaluations for boredom among all the other models. We therefore selected Model 2 to take for field testing, along with Model 4 which was the closest model to that which was used on the old agent. From this point onwards, we shall refer to the agent using Model 4 as Grimace v.1, and the agent using Model 2 as Grimace v.2.

3.4 Field Testing of the Models

The agents were tested with first year high school students from a public school in Metro Manila. The population consisted of 39 males and 51 females with ages ranging from 12 to 14, an average age of 12.53, and a modal age of 12. The students were taking up introductory algebra at the time of the experiment, but none were familiar nor have used Aplusix in the past. These students were randomly assigned into one of three groups – a control group, which used Aplusix without the agent, an experimental group which used Aplusix along with Grimace v.1, and an experimental group which used Aplusix with Grimace v.2.

The experiment began with a pre-test consisting of 10 factoring problems. The problems were of difficulty level B1 (factorization with integer coefficients) of Aplusix. After the pre-test, the students were each given a handout on how to use Aplusix to read for five minutes. Students were allowed to ask questions regarding the software, but were not allowed to interact with the software during this time. The students were then asked to interact with Aplusix (and with the agent for experimental groups) for 45 minutes. During this time, the agent generated interaction logs of the session, which included the student's action, the agent's evaluation of the student's affective state, and, if any, its response to the student. Immediately after the interaction, the students were administered a post-test containing a different set of 10 factoring problems, but of the same difficulty as the pre-test. Finally, for the experimental groups, an Agent Perception Survey based on a study by Baker [2] was given to evaluate the agent. The survey contained a set of eight statements which described the agent, and the students were asked to rate from 1-6 how much they disagreed or agreed with the statements.

4 Results

Throughout the experiment, we were able to collect a total of 45,402 transactions between the students and the two agents, with 22,121 transactions between the students and Grimace v.1 and 23,281 between the students and Grimace v.2.

In analyzing the agent interaction logs from the field test, we observed two properties for each affective state: the frequency of responses/interventions, and the agent's observed incidence of that affective state. The frequency of interventions for an affective state was computed by getting the number of logs for the affective state with interventions and dividing it by the total number of logs for that affective state. Table 4 shows the means and standard deviations of the frequency of interventions per affective state.

Table 4. Results of the Frequency of Interventions analysis

Affective state	Grimace v.1 Mean (SD)	Grimace v.2 Mean (SD)
NEU	0.04 (0.07)	0.03 (0.05)
FLO	0.82 (0.012)	0.25 (0.013)
CON	0.034 (0.03)	0.029 (0.04)
BOR	0.61 (0.17)	0.56 (0.21)

The results showed that the old agent intervened the most when students were engaged, while the new agent intervened the most when students were bored. This was interesting to note because previous work [3] indicated that boredom and confusion, not engagement, require the most amount of intervention. Apart from their association with poorer learning, these affective states tended to persist [3]. Unfortunately, the independent samples two-tailed t-test revealed that the difference between the frequency values of the two agents was only significantly different for engaged concentration ($t(58) = 19.29$, $p < 0.01$).

For the observed incidence of each affective state, the values were computed by taking the number of logs for a particular state and dividing it by the total number of logs generated by the agent (see Table 5).

Table 5. Results of the Incidence of Affective States analysis

Affective state	Grimace v.1 Mean (SD)	Grimace v.2 Mean (SD)
NEU	0.03 (0.03)	0.02 (0.02)
FLO	0.19 (0.11)	0.30 (0.15)
CON	0.10 (0.05)	0.11 (0.06)
BOR	0.67 (0.16)	0.56 (0.20)

As with the computations for frequency, the values per condition were compared using an independent samples two-tailed t-test. The analysis showed that boredom was the most frequently detected affective state. However, Grimace v.2 evaluated students as bored significantly fewer times than Grimace v.1 ($t(58) = 2.45$, $p = 0.02$). Moreover, Grimace v.2 evaluated students as engaged significantly more frequently than the Grimace v.1 ($t(58) = -3.12$, $p = 0.003$).

4.1 Impact on Learning

The pre-test mean scores of all three groups fell under the same range, i.e. the intervals of all groups overlap ($M = 2.3$, 95% CI [1.22, 3.48] for the control group; $M = 1.77$, 95% CI [0.72, 2.82] for the Grimace v.1 group; $M = 1.77$, 95% CI [0.89, 2.65] for the Grimace v.2 group). This meant that there is no significant difference in their performance in the pre-test, and thus all of the students across all groups were generally of the same level of knowledge and ability, albeit the mean scores were very low. Unfortunately, the computed post-test mean scores, which improved across all

groups, still overlapped with each other (M = 6.7, 95% CI [5.62, 7.78] for the control group; M = 6.77, 95% CI [5.63, 7.77] for the Grimace v.1 group; M = 6, 95% CI [4.84, 7.16] for the Grimace v.2 group). In addition, a One-Way Analysis of Variance (ANOVA) of the learning gains of each group (M = 0.55, SD = 0.35 for the control group, M = 0.58, SD = 0.32 for the old agent group, and M = 0.52, SD = 0.33 for the new agent group) indicated no significant difference among the groups ($F(2,87) = 0.25$, $p = 0.78$).

4.2 Impact on Learning Experience

Table 6 shows a breakdown of the mean scores for each statement. Note that scores for negative statements have been inverted for the t-test. A t-test of the scores indicated that there were no significant differences on individual criterion. However, we noted that the overall trend per criterion was in favor of Grimace v.2. A paired t-test showed a significant difference ($t(7) = -4.50$, $p = 0.002$) in the mean scores for each statement, showing an overall preference for the new agent.

Table 6. Agent Perception Survey Mean Scores (5=strongly agree; 1=strongly disagree). Statements with (*) refer to negative statements. Score values were reversed.

Statement	Grimace v.1 Mean Score	Grimace v.2 Mean Score
Grimace is friendly.	4.93	5.1
Grimace is smart.	5.26	5.43
Grimace treats people like individuals.	4.7	5.2
Grimace ignores my feelings. (*)	2.47	2.57
I feel that Grimace, in his own unique way, genuinely cares about my learning.	5.33	5.5
Grimace wants me to do well in class.	5.6	5.73
Grimace is irritable. (*)	3.04	3.5
I would like it if Grimace was part of my regular tutor.	4.86	5.1

5 Discussion

The purpose of the study was to create a new version of the Aplusix agent that timed its responses and evaluations appropriately, significantly influenced the student's ability to learn, and significantly made the student's learning experience better. Based on the results, Grimace v.2 performed better than Grimace v.1 in all of the three aspects, but was not significantly better in every criterion.

In terms of appropriate timing of responses, Grimace v.2 responded significantly less often to engaged students. This was important because those who were in this affective state were already performing well [3], [9], and thus required little or no intervention in comparison to their peers. In addition, previous experience with Grimace v.1 [1] indicated that excessive interventions towards engaged students only

make the student irritated at the agent. Grimace v.2 also significantly tended to evaluate the students more as engaged and less as bored. This allowed the agent not only to fire fewer interventions, but also provide more positive feedback, especially when students correctly solved items in the ITS.

However, there was still much that could be done to improve the ability of the agent to influence the student's ability to learn. Grimace v.2, like its predecessor, only had a limited set of responses, all of which were purely motivational in nature and did not take into consideration the current problem being answered. Although Aplusix currently has domain-based agents in place to handle any cognitive needs of the student, perhaps one possible improvement is to make the responses more adapted towards math learning. In addition, further improvement can still be done to the models to more accurately reflect the incidences noted from human observation, as well as general improvement in the agent's appearance, among others.

Nevertheless, results show that there is an overall preference for Grimace v.2 over v.1, indicating that there was indeed a better learning experience. This, we believe, is a step towards a development of a true emotionally intelligent agent for algebra that is capable of improving and sustaining motivation, and in the long run, achievement.

Acknowledgements. We thank the Ateneo Laboratory for the Learning Sciences for its unique contributions to our research, in particular Marc Armenta and Paul Contillo for their assistance during our field testing. We especially thank Dr. Joseph Beck of Worcester Polytechnic Institute in Worcester, MA and Dr. Ryan Baker of Columbia University in New York, NY for their valuable insights and support. We thank Department of Science and Technology Philippine Council for Industry, Energy, and Emerging Technology Research and Development (PCIEERD) for making this research a reality through the grant entitled, "Development of Affect-Sensitive Interfaces".

References

1. Andallaza, T.C.S., Jimenez, R.J.M.: Design of an Affective Agent for Aplusix. Undergraduate thesis, Ateneo de Manila University, Quezon City (2012)
2. Baker, R.S.J.d.: Designing Intelligent Tutors That Adapt to When Students Game the System. Doctoral Dissertation, Carnegie Mellon University, Pittsburgh (2005)
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. Intl. Journal of Human-Computer Studies, vol. 68 (4), 223-241 (2010)
4. Bate, A.E.: Automatic Detection of Off-Task Behavior While Using an Intelligent Tutoring System for Algebra. Master's thesis, Ateneo de Manila University, Quezon City (2010)
5. Cassell, J., Bickmore, T. Campbell, L., Vilhjálmsson, H., Yan, H.: Human conversation as a systems framework: Designing embodied conversational agents. Embodied Conversational Agents. The MIT Press, USA, pp. 29-63 (2000)

6. Chaachoua, H., Nicaud, J., Bronner, and Bouhineau, D.: APLUSIX, a Learning Environment for Algebra, Actual Use and Benefits. In: 10th International Congress on Mathematics Education (2004)
7. Graesser, A., D'Mello, S., Strain, A. Computer Agents that Help Students Learn with Intelligent Strategies and Emotional Sensitivity. Philippine Computing Journal Dedicated Issue on Affect and Empathic Computing vol. 6 (2), pp. 1-8 (2011)
8. Graesser, A., Person, N., Harter, D., The Tutoring Research Group. Teaching Tactics and Dialog in AutoTutor. Intl. Journal of Artificial Intelligence vol. 12 (3), pp. 257-279 (2001)
9. Lagud, M.C.V., Rodrigo, M.M.T.: The affective and learning profiles of students while using an intelligent tutoring system for algebra. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010. LNCS, vol. 6094, pp. 255-263. Springer, Heidelberg (2010)
10. Leelawong, K., Biswas, G..: Designing Learning by Teaching Agents: The Betty's Brain System. Intl. Journal of Artificial Intelligence in Education vol. 18 (3), pp. 181-208 (2008)
11. Lim, S.A.: Towards a Framework for Developing Motivational Agents in Intelligent Tutoring Systems. Master's thesis, Ateneo de Manila University, Quezon City (2010)
12. Matsuda, N., Keiser, V., Raizada, R., Tu, A., Stylianides, G., Cohen, W., Koedinger, K: Learning by teaching SimStudent: Technical accomplishments and an initial use with students. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010. LNCS, vol. 6094, pp. 317-326. Springer, Heidelberg (2010)
13. Rebolledo-Mendez, G., Du Boulay, B., Luckin, R.: Motivating the Learner: An empirical evaluation. In: 8th International Conference on Intelligent Tutoring Systems, pp. 545-554 (2006)
14. Wooldridge, M., Jennings, N: Intelligent agents: Theory and practice. The Knowledge Engineering Review vol. 10 (2), pp. 115-152 (1995).