Real-Time Monitoring of ECG and GSR Signals during Computer-Based Training

Keith W. Brawner[1] and Benjamin S. Goldberg[1]

[1] United States Army Research Laboratory
Human Research and Engineering Directorate
Learning in Intelligent Tutoring Environments (LITE) Laboratory
Simulation and Training Technology Center, Orlando, FL 32826
{keith.w.brawner, benjamin.s.goldberg}@us.army.mil

Abstract. The potential of Intelligent Tutoring Systems (ITSs) to influence learning may be greatly enhanced by the tutor's ability to accurately assess the student's state in real-time and then use this state as a basis to provide timely feedback or alter the instructional content. In order to maximize the ITS' potential to influence learning, the physiological state of students needs to be captured and assessed. Electrocardiogram (ECG) and Galvanic Skin Response (GSR) data has been shown to be correlated to physiological state data, but the development of real-time processing and analysis of this data in an educational context has been limited. This article describes an experiment where nineteen participants interacted with the Cultural Meeting Trainer (CMT), a web-based cultural negotiation trainer. Metrics of mean, standard deviation, and signal energy were collected from the GSR datastream while instantaneous and average heart rate were collected from the ECG datastream using a windowing technique around important interactions. Our analysis assesses these measures across three interaction scenarios. The findings of this experiment influence the appropriateness of instructional intervention, and drive the development of real-time assessment for education.

Keywords: Intelligent Tutoring, Affective Computing, Physiological Sensing, Scenario-Based Training, Instructional Intervention

1 Introduction

Technology-driven instruction has led to a culture of learning that extends beyond the confines of the conventional classroom. With continual advancements in computing resources and artificial intelligence, computer-based instruction has evolved into a means for providing tailored and personalized educational experiences. This is achieved through the application of Intelligent Tutoring Systems (ITS) that monitor student interactions in real-time and adapt learning events to the individual. ITSs, in certain domains such as mathematics, have been shown to be more effective than traditional classroom instruction [1]. This capability is propagated through web-based systems that produce a one sigma difference, on average, in performance and reduce the need for training support personnel by 70%, and operating costs by 92% [2].

However, expert human tutoring has shown to produce two standard deviations of improvement [3]. Tutors sense and make decisions based upon observations relating

adfa, p. 1, 2011.

to affective states, and are then used by tutors to direct flow and difficulty [4]. This promotes efficiency and thoroughness in decision making and problem solving [5]. While humans sense affect naturally, ITSs must assess the user via sensors. An affect-sensitive ITS monitors the emotional state of the user in order to provide intervention, if appropriate. Sensor technology advancements offer a unique opportunity with this approach, as student interactions and physiological variables can be monitored. This allows for an ITS to respond to an individual student's affective needs, which can improve learning outcomes [6].

There is a strong link between affect, cognition, and learning [7]. Electrocardiogram (ECG) and Galvonic Skin Response (GSR) signals, specifically, have been shown to be significant factors in emotional aspects. Several researchers are beginning to believe the claims that GSR [8] and ECG [17] data are appropriate for response to ITS interactions, which are the sources of measured data in this paper.

If a student is monitored in real-time and assessed to be in a state which is not conducive to learning, there is still the issue of what type of instructional interaction to apply for correction. Two possible methods of instructional intervention that can be implemented within scenario-based training are to reduce specificity of task or provide an unexpected response. It is expected that the response to these types of variations is observable within the ECG and GSR metrics.

2 Methodology

Each participant interacted with the CMT, a web-based system prototype for training bilateral negotiations. The game characteristics are representative of Middle Eastern culture, with scenario interactions presented through static dialogue. Each of the participants experienced 5-6 minute conversations with three individuals, in randomly assigned order. A baseline measurement and break period of 120 seconds was included between each of these interactions. Interactions with the three characters corresponded to information gathering assignments at a hospital following an insurgency attack. The first of these tasks was Well-Defined with No Interruption (WDNI) and involved maintaining small talk with an in-house physician. The second task, which was Ill-Defined with No Interruption (IDNI), was a conversation with the lead physician to gain information without making firm commitments. The third task, which was Ill-Defined with an Interruption (IDI), was intended to gain US support and identify hospital needs with the hospital administrator. The character interrupts discussion by speaking out of turn when an answer is attempted.

The methodology of this paper is heavily based upon the previously reported pilot study [9]. However, there are two large deviations: the type of data collected and the population group of interest. The first difference between experiments is that this study focuses on the ECG and GSR datasets. The second is that this study focused on a population of interest: current cadets of the United States Military Academy (USMA) at West Point.

Thirty-five cadets volunteered as subjects for this study. Following informed consent and collection of demographics, each participant was fitted with ECG and GSR sensors from the Biopac system. Due to noise in data collection and erroneous tagging of gameplay events, only nineteen sets of usable data were identified. These errors in data quality are that of the collection apparatus and the controlling software program, and are not believed to be systemic to GSR data collection methods. Of the nineteen cadets collected, 15 were males (age $\mu = 19.8$, $\sigma = 1.15$) and 4 were females (age $\mu = 19.25$, $\sigma = 0.96$). Participants reported intermediate (58%) or basic (42%) skill with computer games, with none claiming mastery.

The physiological data was collected using a BIOPAC MP 150 system at a 500 Hz sampling rate. This rate allows for the capture of individual heart beats, and meets requirements for GSR analysis. Each participants' signal is preprocessed for areas of interest. These data points, in order, are taken before, after, and halfway between each system interaction. These samples are sixteen seconds (8000 samples) in duration. Sixteen seconds is sufficient time to extract an instantaneous Heart Rate Variability point, and to perform a power analysis in the GSR signal.

Fig. 1. – Areas of analysis interest

The ECG signal has had the following features extracted: the heart rate between the closest two heartbeats to the event, and the averaged heart rate over the interval. The GSR signal, which responds slower to change, has had the following features extracted: the mean, standard deviation, and energy within the interval. All feature extraction in this paper has been performed with the idea of a real-time adaptive ITS in mind, and represents signals that can be communicated to real-time algorithms to determine whether an intervention is required. This is intended to be used in the Generalized Intelligent Framework for Tutoring (GIFT) [10] system, which uses both real-time sensor and performance data to drive personalized instructional intervention.

ECG Signal. The ECG signal is processed for real-time QRS detection in accordance with original work on the subject [11]. The signal passes through a slightly improved second-order band pass filter. It then has the derivative taken, is squared, moving window integrated (MWI), and thresholded for heartbeat detection, shown below:

- Filter Response: $\frac{s*w_0}{s^2+s*\frac{w_0}{Q}+w_0^2}$ (with a center frequency of 5 and a Q value of 4)
- $\frac{\mathrm{d}}{\mathrm{dx}} = y(nT) = \frac{1}{8} * T[-x(nT-2T) - 2x(nT-T) + 2x(nT+T) + x(nT+2T)]$
- Squaring: $y(nT) = [x(nT)]^2$
- MWI: $y(nT) = \left(\frac{1}{N}\right) * [x(nT-(n-1)T + x(nT-(N-2)T + \cdots + x(nT)]$
  (N is 30 samples, or a 3.6 millisecond delay for this work)

GSR Signal. The fundamental GSR data item of interest within the window is the change in response to stimulus. As such, the features that have been extracted over the window are the mean, standard deviation, and signal energy [13]. This is completed by using the steps of smoothing ($y[n] = \frac{1}{\tau}x(n) + \frac{\tau-1}{\tau}y(n-1)$), normalization ($s(t) = \frac{s(t)-\mu_s(t)}{\sigma_s(t)}$), and second difference energy ($\sqrt{\int_t \frac{d^2}{dt^2}(s(t))}$).

3 Results

The post-processed set of ECG and GSR data was used for statistical analysis. Both within-subject and between-subject tests were run looking for statistically reliable differences in the calculated metrics across treatments. The variability in scenario manipulations is hypothesized to produce varying levels of arousal, which should be represented in the collected data. It is important to note that self-reported measures of engagement, via the Independent Television Commission-Sense of Presence Inventory instrument [13], and mood, via the Self-Assessment Manikin [14], were collected following the completion of each scenario, but there was minimal variance in responses between treatments. This analysis focuses on the recorded physiological data.

Analysis showed ECG data to display minimal variance over time and across scenarios, including the IDNI scenario. This can be seen when looking at the correlations between ECG metrics (Instantaneous Heartbeat Rate: [IDI vs. IDNI $r = 0.945$, $p<.0001$; IDI vs. WDNI $r = 0.871$, $p<.0001$; and IDNI vs. WDNI $r = 0.771$, $p<.0001$] and Average Heartbeat Rate [IDI vs. IDNI $r = 0.943$, $p<.0001$; IDI vs. WDNI $r = 0.904$, $p<.0001$; and IDNI vs. WDNI $r = 0.846$, $p<.0001$]). Due to this factor, the results highlight the GSR data.

A non-directional t-Test ($\alpha = .05$) was used to compare the average for all three GSR outputs to identify scenarios that produced significant differences in GSR metrics. Interestingly, results show reliable differences in all metrics when comparing the ill-defined treatments against the well-defined. When evaluating IDI against WDNI, significant differences were found for the average of the windowed-mean (IDI [M = 2.272, SD = 1.08] and WDNI [M = 2.555, SD = 1.23], $t(18) = -2.643$, $p<.025$), the average standard deviation (IDI [M = .027, SD = .019] and WDNI [M = .041, SD = .035], $t(18) = -2.323$, $p<.05$), and the average signal energy (IDI [M = 387787.3, SD = 373776.2] and WDNI [M = 261590.1, SD = 268921.3], $t(18) = 2.414$, $p<.05$). Similarly, the test looking at IDNI compared with WDNI had analogous results, with the exception of the windowed-mean, which reported a p-value just above the .05 threshold. For the two remaining measures, the average standard deviation (IDNI [M = .0234, SD = .016] WDNI [M = .0408, SD = .035], $t(18) = -2.472$, $p<.025$), and the average signal energy (IDNI [M = 373610.4, SD = 315170.5] WDNI [M = 261590.1, SD = 268921.3], $t(18) = 2.965$, $p<.01$) all show statistically reliable differences.

To examine detectable differences within individual subjects, a repeated-measure analysis of variance was conducted, which allows for the observance of data variability created by individual differences. As seen in the between-subject analysis, all three GSR metrics are reporting to be reliably different. The result shows the scenario to have a main effect on the windowed-mean, $F(3, 15) = 4.184$, $p<.05$, the average standard deviation, $F(3, 15) = 4.787$, $p<.025$, and the average signal energy, $F(3, 15) = 3.643$, $p<.05$. Upon further analysis, a pairwise comparison was used to identify the scenarios to have the largest effect on collected GSR data. Of all the compared treatments, only two pairs were reported as being significantly different. Results show individuals in the WDNI condition output had significantly higher GSR scores in the windowed-mean when compared to the IDI treatment, with a mean difference between the two scenarios of 0.283, p=.05. As well, participants in the WDNI condition output had significantly lower signal energy scores when compared to the IDNI treatment, with a mean difference between the two scenarios (p=.025).

4 Discussion and Future Work

The experiments described above are intended to examine several effects. The first of these is that ECG and GSR measurements will be able to discern a difference between well- and ill-defined scenarios. The second is that, between the ill-defined scenarios, the interjection of an interruption will have an effect of the participants' further responses.

The combination of self-paced instruction, web-based interaction, static character pictures, and text feedback failed to vary heart rate, and lowered survey response across all scenarios, but represents typical web-based instruction response. There were no reportable differences in dependent variables between the IDI and IDNI scenarios. This is an indication that the instructional event of interrupting the user had no effect on their arousal levels. It is an interesting conclusion that within the context of this training environment, this intervention is shown to be an inappropriate instructional strategy to increase engagement. The authors continue to believe that interruption is still a valid strategy among more engaging applications.

Significant differences in the windowed measurements of mean, standard deviation, and signal power were found between the well- and ill-defined scenarios. GSR is a measurement of anxiety, arousal, boredom, frustration, or stress [15]. Interaction scenarios without clear goals, such as in the ill-defined interaction context, are likely to produce lower levels of arousal. This is supported by work examining the relation between performance and stress through compensatory control of one's attention and effort [16]. This effect is observed without regard to self-reported immersion and heart rate response. While it is noted that USMA cadets may have less of a response to being interrupted, there is a clear difference when not given a specific mission.

Future work to assess real-time changes in trainee affect is motivated by the ability of the GSR signal to detect significant differences among experiences. This is encouraging when combined with the wide availability of low-cost GSR sensors. There is additional research being conducted to investigate alternate low-cost sensors, with promising results, and the data stream feature extractions created as part of this work are intended for use within GIFT [10].

Acknowledgements. The authors would like to thank two organizations for this work. We would like to thank SoarTech for the design of the data collection testbed and scenarios. We would also like to thank Dr. Michael Matthews of the Behavior Sciences and Leadership Department at the United States Military Academy at West Point.

5 References

1. Verdú, E., Regueras, L.M., Verdú, M. J., De Castro, J.P., Pérez, M.A.: Is Adaptive Learning Effective? A Review of the Research. In: Proceedings of the 7th WSEAS International Conference on Applied Computer & Applied Computational Science (ACACOS '08), pp. 710-7150. WSEAS Press, Stevens Point, Wisconsin (2008)
2. Woolf, B.: Reasoning about Teaching and Learning. In: 10th Conference of Spanish Association for Artificial Intelligence (CAEPIA 2003) and the 5th Conference on Technology Transfer (TTIA 2003), San Sebastian, Spain. (November 2003)
3. Bloom, B. S.: The 2 sigma problem. The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13(6), 4-16 (1984)
4. Lehman, B., Matthews, M. D'Mello, S., Person, N., 2008. What are you feeling? Investigating student affective states during expert human tutoring sessions. In: Proceedings of Intelligent Tutoring Systems Conference (ITS), Montreal, Canada (June).
5. Isen, A.M.: Positive Affect and Decision Making. In: Lewis, M., Haviland, J.M. (eds.) Handbook of Emotions, pp. 261-277. Guilford Press, New York (1993)
6. Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., Picard, R.: Affect-Aware Tutors: Recognizing and Responding to Student Affect. International Journal of Learning Technology. 4, 129-164 (2009)
7. D'Mello, S.K., Craig, S.D., Gholson, B., Franklin, S., Picard, R.W., Graesser, A.C.: Integrating Affect Sensors in an Intelligent Tutoring System. In: Affective Interactions: The Computer in the Affective Loop Workshop at 2005 International Conference on Intelligent User Interfaces, pp. 7-13. AMC Press, New York (2005)
8. Handri, S., Yajima, K., Nomura, S., Ogawa, N., Kurosawa, Y., Fukumura, Y. : Evaluation of Student's Physiological Response Towards E-Learning Courses Material by Using GSR Sensor. In: Computer and Information Science (ICIS), 2010 IEEE/ACIS 9th International Conference, pp.805-810. IEEE, Los Alamitos, CA (2010)
9. Goldberg, B., Sottilare, R., Brawner, K., Holden, H.: Predicting Learner Engagement during Well-defined and Ill-defined Computer-Based Intercultural Interactions. In: Proceedings of the Affective Computing and Intelligent Interaction Conference (ACII) 2011, Memphis, Tennessee (October 2011)
10. Sottilare, R., Holden, H., Brawner, K., Goldberg, B.: Challenges and Emerging Concepts in the Development of Adaptive, Computer-based Tutoring Systems for Team Training. In: I/ITSEC Emerging Concepts Group Proceedings, Orlando, Florida (2011)
11. Pan, J., Tompkins, W.J.: A real-time QRS detection algorithm. IEEE Trans. Biomed. Eng. BME, 32(3): 230-236 (1985)
12. Grundlehner, B., Brown, L.. Penders, J., Gyselinckx, B.: The design and analysis of real-time, continuous arousal monitor. In: Proceedings of the 6th International Workshop on Wearable and Implantable Body Sensor Networks (2009)
13. Lessiter, J., Freeman, J., Keogh, E., Davidoff, J.: A Cross-Media Presence Questionnaire: The ITC-Sense of Presence Inventory. Presence, 10(3), 282-297 (2001)
14. Bradley, M.M., Lang, P.J.: Measuring emotion: The Self-Assessment Manikin and the Semantic Differential. Journal of Behavior Therapy and Experimental Psychiatry. 25(1), 49-59 (1994)
15. Kapoor, A., Burleson, W., & Picard, R. W. Automatic prediction of frustration. International Journal of Human-Computer Studies, 65, p. 724-736 (2007)
16. Hockey, G.R.J. A state control theory of adaptation to stress and individual differences in stress management. In G.R.J. Hockey, A.W.K. Gaillard & M.G.H. Coles (Eds.), Energetics and Human Information Processing. Dordrect: Martinus Nijhoff (1986)