2011

Combining Behavioral and Biometric Measurements for Automated Performance Assessment

Chris Forsythe

Robert A. Abbott

Susan M. Stevens-Addams

Michael Haass

Laura Matzen

See next page for additional authors

COMBINING BEHAVIORAL AND BIOMETRIC MEASUREMENTS FOR AUTOMATED PERFORMANCE ASSESSMENT

Chris Forsythe, Robert A., Abbott, Susan M. Stevens-Adams, Michael Haass, Laura Matzen, Kiran Lakkaraju
Sandia National Laboratories
Albuquerque, NM, USA

Technologies are needed enabling more cost-effective military aviation training. Automated performance assessment has been advanced as one approach to enable instructors to make more effective use of simulation-based training systems. Recent experimental research will be reviewed illustrating that automated techniques produce student assessments comparable to human appraisals of student performance and employed within an after-action debrief, resulted in more effective training, as compared to a baseline after-action debrief capability. These studies used the E-2 Enhanced Deployable Readiness Trainer (EDRT), a medium-fidelity simulation trainer employed for training E-2 Hawkeye Naval Flight Officers (NFOs). This paper will summarize further developments to combine behavioral with biometric measures as a basis for automated performance assessment. In particular, speech communications and EEG were assessed as two-person teams of E-2 NFOs completed relatively complex mission scenarios on the EDRT. Using biometric measures, it was possible to distinguish the performance of expert and novice teams, providing a proof-of-principle of feasibility.

As military aviation systems become increasingly complex, training has become a significant cost-driver in the life cycle of aviation platforms. Consequently, there is need for technology innovations that increase the effectiveness of military aviation training, while lessening the demands on human instructors, and their support staff.

Automated performance assessment has been advanced as a technology that should allow reductions in the manpower required to support training operations. This assertion is based on current practices which require instructors observe and grade student performance as students complete missions within simulation-based trainers. For complex operations, the instructor-to-student ration may reach one-to-one, with the need for human role players creating even greater manpower requirements. In theory, by using automated techniques to assess certain facets of student performance, there should be a reduction in the cognitive workload for instructors enabling training to be accomplished with fewer instructors. Furthermore, automated performance assessment is well suited for performance measures that involve continuous attention to detailed facets of student performance (e.g. situationally-dependent duration of radio communications), providing a basis for objectifying such measures.

Previous research has established that the performance assessments obtained using automated techniques are accurate, as compared to equivalent measures obtained through human observation and assessment of student performance (Stevens et al, 2010, Stevens et al, 2010). This research utilized the E-2 Enhanced Deployable Readiness Trainer (EDRT), a medium-fidelity simulation trainer used to train E-2 Hawkeye Naval Flight Officers (NFOs), and focused on key performance measures appropriate for entry-level NFO training (e.g. fleet protection, or preventing enemy aircraft from coming within close enough proximity to pose a threat to a Naval

carrier group). Furthermore, when presented as a component of an instructor's after-action debrief, more effective training was achieved with automated performance assessments, as compared to a condition employing a baseline after-action debrief capability (Stevens et al, 2010). For these efforts, the input to the automated performance assessment consisted of readily available data, and specifically, the geometric relationships between entities within the simulation (e.g. relative positions, directions and speeds of enemy and friendly entities) and distinct transactions as students operated the simulator (e.g. labeling of entities, depressing foot pedal actuator for radio). The following paper considers extension of these data sources to include analysis of speech communications and biometric measurement of brain activity, presenting a proof-of-principle demonstration in which speech and EEG serve as inputs to automated performance assessment.

Automated Expert Modeling and Student Evaluation (AEMASE)

AEMASE has been advanced as both an approach to automated performance assessment, and as specific algorithmic instantiations of this approach (Abbott, 2006). As an approach, AEMASE consists of a three-step process. First, an expert demonstrates desired behavior within either a simulator or instrumented environment. Key to this step is the prior identification of key performance parameters (i.e. features) underlying task performance. Based on data from experts, machine learning techniques are employed to derive a model of expert performance. The specific techniques employed may vary depending on the performance measure. In many cases, performance measures have been modeled using a vector-based representation combining different features within a multi-dimensional parameter space. For example, student's performance for fleet protection (i.e. preventing enemy aircraft from posing a threat to a Naval carrier group) may be modeled as a vector that combines the features for each enemy aircraft: (1) distance from carrier group; (2) angle off and (3) velocity. In the third step, during a training exercise, data is fed into the expert model which provides predictions concerning appropriate courses of action. The actual performance of the student is then compared to these predictions and the difference between predictions generated by the expert model and the student's actual behavior provide the basis for assessing the student's performance.

While various approaches have been employed for automated performance assessment, such as intelligent tutoring systems concepts (e.g. Corbett, 2001), there is a distinction worth noting. The expert models used in the AEMASE approach are based on statistical analysis of data produced as experts perform within a representative task environment. One of the costliest elements of most automated performance assessment concepts is knowledge engineering (i.e. expert interviews, task decomposition, etc.) required to derive a detailed model of expert performance. AEMASE does require some degree of knowledge engineering, but this is primarily restricted to steps associated with identifying performance measures, and associated data features, and obtaining sufficient instances of expert performance. Thus, AEMASE provides a more cost-effective approach for system development, and given interface features that allow users to readily modify expert models, AEMASE streamlines the process for later updating the system.

Accuracy and Utility of AEMASE Automated Performance Assessment

To date, the most extensive implementation of the AEMASE approach has been for training E-2 NFOs. To assess the accuracy and utility of AEMASE automated performance assessments, laboratory studies have been conducted using the E-2 EDRT simulation trainer. In these studies, test subjects were recruited from the employee population of Sandia National Laboratories with demographics comparable to entry-level E-2 NFOs. Subjects then underwent a program of training to provide them with the basic skills needed to complete relatively complex, yet entry-level E-2 mission scenarios. This training consisted of an 8-hour classroom session taught by a reservist E-2 NFO and five sessions on the simulation trainer focused on the development and practical application of key skills. Students were then brought back for a final data collection session in which their proficiency was assessed as they completed two missions requiring an integration of the knowledge and skills attained in the earlier training sessions.

In the first of two studies, the objective was to compare automated assessments with those of human instructors. For this study, three performance measures were chosen that were each deemed to be highly relevant to the training objectives for an entry-level E-2 NFO. The first concerned fleet protection, or the effectiveness with which students recognized potential threats (i.e. enemy aircraft) to a Naval carrier group and committed friendly aircraft to intercept approaching enemy aircraft within a timely manner. The second measure involved the timeliness with which commercial aircraft were identified and labeled. The third addressed situation awareness and management of the battlespace and in particular, whether students recognized and responded correctly to a developing gap in their air defenses. With each measure, there was good correspondence between the automated and instructor assessments, with values of 100%, 95% and 83% respectively for the three measures.

A second study compared the performance of students trained using an after-action debrief featuring automated performance assessment to students trained with a baseline after-action review capability (i.e. scenario capture and replay). Subject trained with the AEMASE after-action debrief exhibited superior performance for performance measures that included fleet protection, the accuracy and latency for labeling commercial aircraft and the timeliness with which the warfare commander was informed following successful downing of enemy aircraft. There was no difference between groups for the measure of situation awareness and battlespace management described above, and it was concluded that this skill was too complex given the limited training provided to test subjects.

Incorporation of Voice and Biometric Data

Previous studies focused on automated performance assessment using readily available data concerning location of entities within the simulation scenario and student transactions with the simulation trainer. Also, in these studies, training and student assessments occurred on an individual level, outside the context of team operations. Given the degree to which tasks of the

E-2 crew involve a coordinated team effort, there was a certain artificiality in having students conduct missions individually.

A third study was conducted in which subjects participated as two-person teams. In this study, there were 8 subjects, divided into 4 two-person teams. Two of these teams consisted of subjects from the second study that had received training using the AEMASE after-action debrief tool, and were considered to be novices. The other two teams consisted of reservist E-2 NFOs and were considered to be experts. In addition to the data collected in previous studies, voice communications and dense-array EEG was recorded.

Analysis of both voice communications and EEG allowed expert and novice teams to be distinguished. With the EDRT, to activate the radio, students must depress a foot pedal and continue pressing the foot pedal for the duration of the radio call. Initial analysis of voice communications considered the duration of these pedal presses. As shown in Figure 1, across scenarios, expert teams generally pressed the pedal for shorter periods of time, indicating a greater degree of brevity in their radio communications. This is consistent with observations that a key facet of expert NFO performance involves the efficient use radio channels.

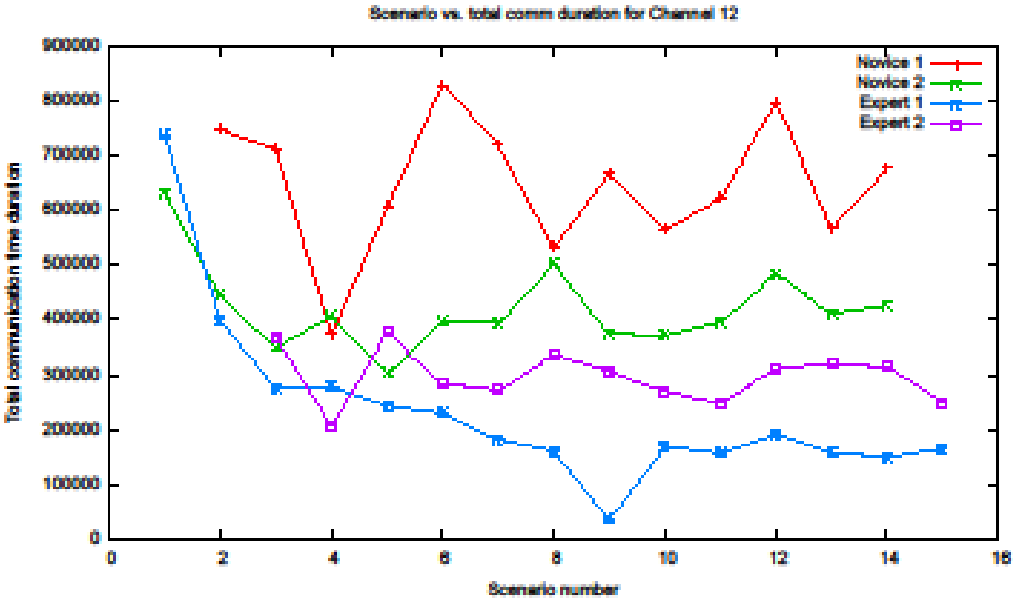

Figure 1. Generally, the duration of radio communications for novice teams was longer than for expert teams.

It was noted that novice subjects often seemed tentative in their radio communications, whereas expert teams tended to be deliberative and concise. Using speech-to-text transcription, the contents of radio communications was assessed. This tentativeness was evident in the use of filler words (e.g. ur, ah), as shown in Table 1. Overall, for five common filler words, their use occurred substantially less with expert, than with novice teams.

Table 1. Use of filler words for an illustrative scenario.

Filler Words	Experts	Novices
ah	1	6
er	4	8
like	5	9
uh	112	307
um	5	28
Total	127	358

Additional analysis of radio communications considered the semantic content of radio communications. For this analysis, transcriptions were indexed using a term frequency-inverse document frequency approach. Each subject's transcript was treated as a separate document and based on cosine similarity, each subject was compared to each of the other subjects to determine who their speech content most closely resembled. As depicted in Figure 2, for three of the four experts, their communications most resembled another expert. Furthermore, all four novices most closely resembled another novice.

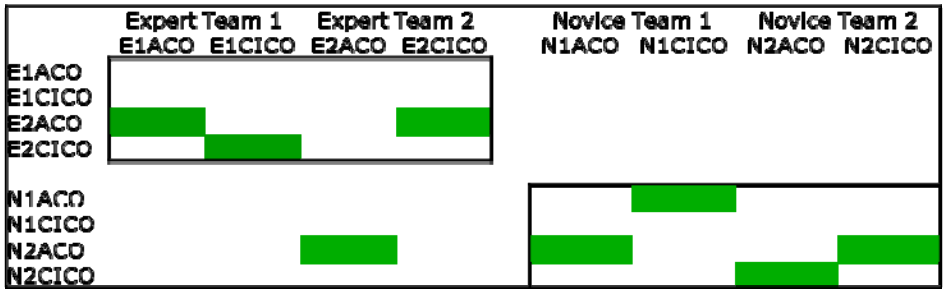

Figure 2. Semantic comparison of radio communications for expert and novice subjects. Green shaded cells indicate the subject each subject most closely resembled.

Initial analysis of EEG data considered the relative levels of activity in the theta (4-7 Hz) and beta (13-30 Hz) bandwidths. Figure 3 depicts the moment-to-moment transitions of activity for individual electrodes for a sample of data. It was observed that experts exhibited greater variability in the beta bandwidth, whereas novices showed greater variability in the theta bandwidth. This would suggest that the neural processes being engaged by experts were somewhat distinct from those being exercised by novices. Further analysis reported by Dodel et al (in preparation), found higher power correlations and reduced dimensionality with the expert teams, suggesting that the coordination of team performance involves some degree of coordinated activation of neural processes.

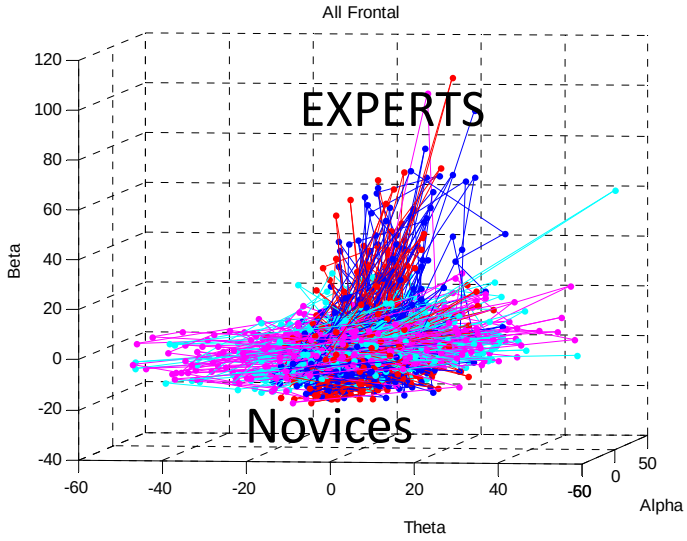

Figure 3. Expert teams showed greater variability in beta (13-20 Hz) bandwidth, while novices showed greater variability in theta (4-7 HZ) bandwidth

Conclusion

Automated performance assessment, as illustrated by the AEMASE approach, provides an opportunity to improve the effectiveness of training, while reducing cost by lessening the workload on instructors and streamlining the development process. Given that the AEMASE approach is based upon expert demonstration of desired performance, concerns may arise regarding the generality from scenarios used to train the expert models to other scenarios that differ in their contents and complexity. However, this same concern exists with expert models developed using traditional approaches based on knowledge engineering, in that resulting models are only valid for the contingencies identified and addressed during domain analysis and model development. A useful conceptualization identifies the key parameters underlying performance and represents those parameters within a multi-dimensional space. Any given scenario, or perhaps mission segment, may be depicted as a point within this parameter space. Ideally, an expert model should generalize to the entire parameter space. However, actual generality will be a function of the extent and care with which the parameter space is sampled in selecting the scenarios utilized to construct the expert model.

It may be noted that in the work summarized in this paper, automated performance assessments were based on comparing student performance to the predictions of an expert model. Often this may not be the most appropriate comparison, and the most appropriate comparison may be to compare a student to a model reflecting performance that is intermediate between the

student and an actual expert. This could be readily accomplished by obtaining data reflecting a range of performance such that intermediate levels of performance are represented within the model against which a student's performance is compared. In fact, an important distinction of the AEMASE approach is that the performance of each expert whose data contributes to the model is reflected within the model (i.e. there does not have to be a single correct solution). This accommodates situations in which there are multiple acceptable solutions to a given problem. Thus, in the same way that AEMASE accommodates variation across experts with respect to their performance, varying levels of expertise may similarly be accommodated.

For the most part, development and experimental assessment of AEMASE implementations have only utilized behavioral performance data. The integration of behavioral performance data with biometrics data offers a mechanism by which these capabilities may be extended to provide more thorough assessments of student performance. As illustrated with voice communications, a novice may say all the right words, but do it in a manner that is ineffective (e.g. use of excessive filler words). Similarly, a novice may complete a mission and if their performance is assessed on a behavioral level, they accomplish all their objectives. However, their cognitive and physiological resources may have been taxed nearly to the breaking point, whereas an expert routinely accomplishes the same objectives with ease. This discrepancy may not be readily apparent and the student allowed to progress, despite their capabilities being on the margins, and likely quite brittle if placed in a stressful situation. Biometric measurement should provide a mechanism to not only assess student's behavioral performance, but to additionally assess the levels of mental and physiological exertion required to obtain performance objectives.

Acknowledgements

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. Work described in this paper was performed through contract awards from the Office of Naval Research and DARPA.

References

Abbott, R. G., (2006). Automated Expert Modeling for Automated Student Evaluation. *Intelligent Tutoring Systems*, 1-10.

Corbett, A. T. (2001). Cognitive computer tutors: Solving the two-sigma problem. In UM '01: Proceedings of the 8th International Conference on User Modeling, (pp. 137 – 147). London: Springer-Verlag.

Dodel, S., Jirsa, V. Mersmann, J., Forsythe, C., Wheeler, T.L. & Cohn, J. (in preparation). Expert teams show reduced variability and enhanced coordination of source generators on a complex decision making task. *Proceedings of Human Computer Interaction International*, Orlando, FL.

Stevens, S., Basilico, J., Forsythe, C., Abbott, R. & Gieseler, C. (2010) Experimental assessment of automated knowledge capture. Proceedings of the I/ITSEC, Orlando FL.

Stevens, S., Basilico, J., Forsythe, C., Abbott, R. & Gieseler, C. (2010) Using After-Action Review Based on Automated Performance Assessment to Enhance Training Effectiveness Proceedings of the Human Factors and Ergonomics Society, San Francisco, CA.