Performance Assessment to Enhance Training Effectiveness

Susan M. Stevens-Adams, Justin D. Basilico, Robert G. Abbott, Charles J. Gieseler & Chris Forsythe
Sandia National Laboratories
Albuquerque, NM
smsteve@sandia.gov, jdbasil@sandia.gov, rgabbot@sandia.gov, cjgiese@sandia.gov, jcforsy@sandia.gov

ABSTRACT

Training simulators have become increasingly popular tools for instructing humans on performance in complex environments. However, the question of how to provide individualized and scenario-specific assessment and feedback to students remains largely an open question. To maximize training efficiency, new technologies are required that assist instructors in providing individually relevant instruction. Sandia National Laboratories has shown the feasibility of automated performance assessment tools, such as the Sandia-developed Automated Expert Modeling and Student Evaluation (AEMASE) software, through proof-of-concept demonstrations, a pilot study, and an experiment. In the pilot study, the AEMASE system, which automatically assesses student performance based on observed examples of good and bad performance in a given domain, achieved a high degree of agreement with a human grader (89%) in assessing tactical air engagement scenarios. In more recent work, we found that AEMASE achieved a high degree of agreement with human graders (83-99%) for three Navy E-2 domain-relevant performance metrics. The current study provides a rigorous empirical evaluation of the enhanced training effectiveness achievable with this technology. In particular, we assessed whether giving students feedback based on automated metrics would enhance training effectiveness and improve student performance. We trained two groups of employees (differentiated by type of feedback) on a Navy E-2 simulator and assessed their performance on three domain-specific performance metrics. We found that students given feedback via the AEMASE-based debrief tool performed significantly better than students given only instructor feedback on two out of three metrics. Future work will focus on extending these developments for automated assessment of teamwork.

ABOUT THE AUTHORS

Susan M. Stevens-Adams is a PhD candidate in the Department of Psychology at the University of New Mexico. She received her M.S. in Cognitive Psychology from the University of New Mexico in 2006. Her dissertation work focuses on individual differences in false memory and semantic networks. In addition, she works at Sandia National Laboratories in both the Reliability Assessment and Human System Integration Department and, most recently, the Cognitive Science and Applications group. Her research experience at Sandia Labs includes conducting usability studies and researching, designing, and implementing experiments.

Justin D. Basilico is a Senior Member of the Technical Staff in the Cognitive Science and Applications group at Sandia National Laboratories He received his B.A. in Computer Science from Pomona College in 2002 and his M.S. in Computer Science from Brown University in 2004. He is the lead designer and developer of the Cognitive Foundry, a software platform for machine learning and cognitive simulation. His research interests include machine learning, information retrieval, user modeling, personalization, statistical text analysis, and human-computer interaction.

Robert G. Abbott, PhD**,** is a Principal Member of the Technical Staff in the Cognitive Science and Applications group at Sandia National Laboratories, where his team develops software for automated behavior modeling. He holds a Ph.D. in computer science from the University of New Mexico. He has been a member of the technical staff at Sandia since 1999. His current research focuses on automating the creation of human behavior models with the objectives of reduced cost and rapid development. Applications include trainable software agents to assume the roles of friendly and opposing forces, and

automated student assessment for distributed virtual training environments. This line of research is supported primarily by the U.S. Navy and includes validation experiments with human subjects to assess the impact of new training technologies. Other research interests include distributed systems, security-related data mining, and computer vision.

Charles J. Gieseler is a software engineer supporting the research efforts of the Cognitive Science and Applications group at Sandia National Labs in Albuquerque New Mexico. He holds a Master's of Science in Computer Science from Iowa State University and Bachelor's of Science in Computer Science from Santa Clara University. Some of his interests include applications of cognitive modeling and augmented cognition, agent-based computational economics, social science simulation, and computational ecology.

Chris Forsythe, PhD, is a Distinguished Member of Technical Staff for the Cognitive Science and Applications group. Chris helped found Sandia National Laboratories' cognitive systems program in 1999 and has served as a technical lead throughout this time, including playing a pivotal role in activities leading to Sandia establishing a Grand Challenge in cognition and establishing Cognitive Science and Technology as a lab-wide Focus Area. Chris is co-inventor and patent holder for Sandia's computational framework for modeling human cognition and initiated the development of automated knowledge capture technologies. Chris has led a wide range of projects and served as Sandia interface to government sponsors for programs that include Capable Manpower, CAT-M and HPTE for the Office of Naval Research, Augmented Cognition for DARPA, and CDMTS for NavAir. Chris has over thirty publications and has edited two books in the fields of cognitive psychology, human factors and intelligent systems.

Performance Assessment to Enhance Training Effectiveness

Susan M. Stevens-Adams, Justin D. Basilico, Robert G. Abbott, Charles J. Gieseler & Chris Forsythe
Sandia National Laboratories
Albuquerque, NM
smsteve@sandia.gov, jdbasil@sandia.gov, rgabbot@sandia.gov, cjgiese@sandia.gov, jcforsy@sandia.gov

INTRODUCTION

Simulation-based training is becoming an increasingly important tool for teaching humans to perform complex tasks in novel environments. Simulation helps to reduce the costs associated with live training, for example, by training pilots on the ground without incurring the fuel and mechanical costs of operating an aircraft. Nevertheless, simulators still suffer from the high labor costs associated with providing individually relevant instruction and feedback. Intelligent tutoring systems (Murray, 1999) can mitigate these costs by automatically providing individualized feedback, but require a substantial investment of time and expertise to construct the needed knowledge base, which in turn can become quickly obsolete.

In this paper, we evaluate the Sandia-developed Debrief tool and Automated Expert Modeling and Automated Student Evaluation (AEMASE) system (Abbott, 2002; Abbott, 2006) to determine whether the system provides useful feedback to students. The system assesses student performance on complex tasks by comparing against learned models of expert behavior in similar situations, thereby reducing the cost of engineering hand-coded knowledge bases. Most research into automated student evaluation has been conducted in the context of intelligent tutoring systems. Murray (1999) provides a survey of intelligent tutoring systems, while Corbett (2001) provides a review of the empirical support for their effectiveness. Jensen, Chen, and Nolan's (2005) work on Combined Arms Command and Control Trainer Upgrade System (CACCTUS) provides one exception. This tool analyzes events from training sessions to find causal relationships among student errors and undesirable outcomes. The system then applies a set of rules to determine and highlight the correct behaviors. This work differs from AEMASE in that AEMASE attempts to learn a model for correct behaviors by observing experts, instead of relying on a crafted rule base. Relatively few efforts have been made at automatically acquiring models of correct behaviors. Anderson, Draper and Peterson (2000) used neural networks to create behavioral clones for piloting simulated aircraft, but their work focused on personal insights based on examination of neural network models of individual students. AEMASE uses its learned models to compare novice and expert behavior automatically.

In prior work, we have demonstrated the feasibility of automated performance assessment tools such as AEMASE through proof-of-concept demonstrations, a pilot study, and an experiment (Stevens, Forsythe, Abbott & Gieseler, 2009). The current study provides a more rigorous empirical evaluation of the enhanced training effectiveness achievable with this technology.

SIMULATION TRAINING

A significant cost in simulation-based training is the time demands on human instructors who monitor student actions and provide corrective feedback. The work presented here focuses on U.S. Navy training of Naval Flight Officers for the E-2-Hawkeye aircraft using a high-fidelity simulator. The three flight officers must learn to detect, track, and identify all assets, such as aircraft, and to provide communication among the commanding officers and all friendly assets. This currently requires a separate instructor to observe each student within the context of team performance and provide instruction based on observed misunderstandings, inefficient task execution, and ineffective or inappropriate actions. Such individualized instruction is labor intensive and contributes to high training costs. The purpose of this study was therefore to determine whether a group given verbal feedback from an instructor on their performance using an AEMASE-based debrief tool would outperform a group simply given verbal feedback alone. A positive result would then imply that use of automated evaluation systems such as AEMASE help to reduce overall training costs.

Establishing the validity of automated assessments requires moving beyond simple laboratory tasks to studies in a realistic training environment. Naval Flight Officers are trained and tested on several different simulators ranging from a part-task computer-based training system that runs on a single PC, to high-end systems, which faithfully replicates most aspects of

E-2 operations and requires a team of instructors and operators to conduct training. For this study, we used the E-2 Enhanced Deployable Readiness Trainer, a fielded, medium-fidelity training system that presents students with the same mission software used on the E-2 aircraft. Simulation training sessions require multiple instructors and can last hours at a time. Automated assessment of E-2 operator performance in these sessions would greatly reduce instructor workload and would increase overall efficiency.

AEMASE

The goal of AEMASE is first to let subject matter experts rapidly create and update their own models of normative behavior and then use these models to evaluate student performance automatically (Abbott, 2006). The system operates in three steps. First, the system must acquire examples of behavior in the simulated environment. Next, machine-learning techniques are used to build a model of the demonstrated tactics. The system then compares student behaviors in the same task environment to the expert model to establish a score. Afterwards, the student and instructor can review the training session by interacting with a plot of the time-dependent grade. The remainder of this section provides additional detail on these steps.

In the initial step, the system records examples of task behavior. The examples may include both good and bad behavior performed by either students or subjects matter experts. Examples may be obtained by performing exercises on the target simulator or within a relevant proxy environment. However, a subject matter expert must accurately grade the examples to provide AEMASE with points of reference in its comparisons to student behaviors during evaluation.

After acquiring graded example behaviors, the system applies machine learning algorithms to create the behavior model. An appropriate learning algorithm must be selected for each performance metric, depending on the type and amount of example data available, such that the resulting model generalizes assessments of the observed behaviors to novel student behaviors. We have implemented a suite of machine learning algorithms (e.g. neural networks, instance-based / nearest neighbor algorithms, support vector machines, linear regression, rule induction) and cross-validation tests to determine which algorithm makes the most accurate predictions for each metric.
Finally, the system uses the learned behavior model to assess student behaviors. As each student executes a simulated training scenario, his or her behavior is compared to the model for each performance metric. The model determines whether student behavior is more similar to good or bad behavior from its knowledge base, and helps to identify and target training to individual deficiencies. Initially, the knowledge base is sparse, and incorrect assessments may be common. However, the instructor may override incorrect assessments. The model learns from this interaction and improves over time.

For the research described here, we used AEMASE as a tool for after action reviews (see Figure 1), although the system could also be used to provide students with feedback throughout a training exercise. After action review is a general process for discussion of a training session to evaluate performance, diagnose problems, and build on successes. For training Naval Flight Officers, we used two basic types of AEMASE metrics.

The first type of AEMASE metric is Context Recognition, which assesses whether the student is maintaining the tactical situation within norms established by previous expert demonstrations. This is done by monitoring the values of one or more continuous metrics (e.g. positions, ranges, headings, fuel load, etc). Unexpected combinations of values indicate the student may not know what to do, or may be losing control of the situation. The Fleet Protection metric described below is a simple (one-metric) example.

The second type of AEMASE metric is Sequence Recognition, which assesses whether certain sequences of events provoke the expected sequence of responses. An example is Labeling Neutral Entities; a set of events (appearance of a radar track, detection of certain RF emissions) should lead to specific actions by the subject (labeling the track as a non-combatant). Any failure of the student to complete the sequence within a time limit (determined by modeling expert response times) is flagged for review.

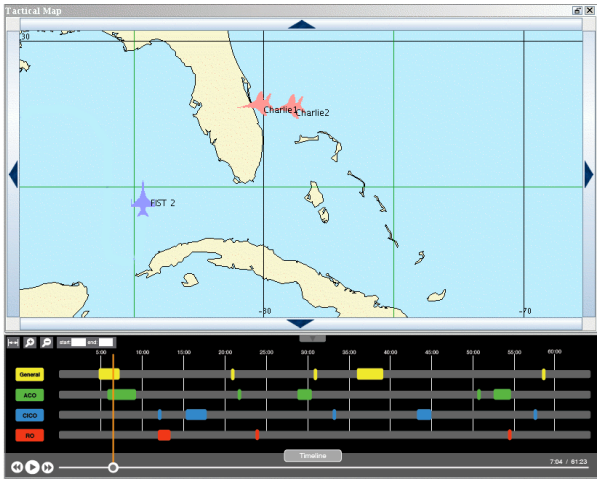

Figure 1. Debrief Tool With Automated Event Flagging. The debrief tool used in the experiment displays a video replay of the operator console (similar to this map display), and a timeline of events suggested by AEMASE for discussion during debrief. The tool also includes visualizations of entity movement over time (see Figure 3).

In an earlier study, AEMASE achieved a high degree of agreement with a human grader (89%) in assessing tactical air engagement scenarios (Abbott, 2006). However, the 68 trials assessed used only four subjects under three different training scenarios, and the range of correct behaviors was limited. In a more recent study, AEMASE achieved a high degree of agreement with human graders (83-99%) for three different E-2 metrics (Stevens, et al., 2009). However, these studies did not test whether giving students feedback based on the automated metrics would enhance training effectiveness and improve student performance. The current study takes the next step by quantifying the training benefit of instructor feedback based on automated metrics.

METHODS

The goal of this work is to determine whether students achieve higher proficiency when their instructor is assisted by the automated system. Toward this end, we compared two groups of students using the Naval Flight Officer training program. In the debrief group, the instructor used the debrief tool to detect student errors and replay them during debrief for students. In the control group, the instructor used the same amount of time for debriefs but did not use a debrief tool.

Participants

Volunteer civilian employees were recruited via advertisement. All twenty-two participants met certain required criteria for the experiment that reflected the requirements for an entry-level E-2 Hawkeye operator. The participants were both men and women and were between the ages of 20 and 28. The participants were split into two groups: a control group (N=12) and a debrief group (N=10). Two experienced E-2 Hawkeye Naval Flight Officers served as subject matter experts (SME's).

Materials

Materials included an E-2 Distributed Readiness Trainer simulator obtained from the Naval Air Systems Command's Manned Flight Simulator organization. The U.S. Government-owned Joint Semi-Automated Forces simulation software was used to create and drive the scenarios for training and testing participants. In addition, the AEMASE software and AEMASE-based debrief tool were used during the experiment. Finally, the U.S. Government-owned Common Distributed Mission Training Station software was used in the data analysis.

.

Procedure

The participants provided informed consent and were then scheduled for an initial eight-hour training session. Here, an E-2 Hawkeye Naval Flight Officer provided a tutorial on E-2 operations emphasizing the basic radar systems task that would be the subject of the experiment. Following this initial session, the participants were scheduled individually for five simulation-based training sessions. All participants were led through these sessions in the same order. After finishing the training sessions, the participants individually completed two testing sessions. Human graders assessed each of three metrics (described below) for the testing sessions. Two trained experimenters graded each participant's performance and performance was compared between the two groups.

Training Sessions

The five simulation-based training sessions were designed by an E-2 subject matter expert to teach the basic operations of the E-2 radar system on the simulator. The topics included simulator familiarization, check-in procedures, and managing air assets, managing surface assets and integration of air and surface pictures in complex tactical scenarios. For each session, the experimenters first demonstrated the proximate operation(s) on the simulator, after which the participant was asked to perform the operation(s) in scaled down, yet realistic, simulations. Since all five of these sessions were for training purposes, the

experimenters were available to answer questions. Each training session lasted approximately 1.5 hours.

For the control group, the instructor gave participants real-time, verbal feedback of their training session performance deficiencies. For the debrief group, the instructor used a debrief tool featuring graphical depictions (e.g., timeline and occupancy maps derived by AEMASE) of participants' performance in addition to real-time, verbal feedback. The instructor was given sufficient training on how to use the debrief tool before the experiment started.

Testing Session

The last two sessions were testing sessions in which the participants were assessed on their knowledge of the operations and tactics covered in the five training sessions. The participants completed these more difficult simulations without the help of the experimenters. Each testing scenario lasted about 1 hour.

METRICS

Based on guidance from the subject matter experts, we developed three metrics to grade the participants' performance in the test sessions. These metrics correspond to a subset of those used by the Navy in training Naval Flight Officers, and include fleet protection, labeling of neutral entities, and battlespace management.

Fleet Protection

Participants were instructed to prevent non-friendly entities from nearing the carrier group. Performance was assessed based on the latency to commit friendly fighters to enemy fighters as they approached the carrier group. During training, participants were given feedback regarding how quickly they committed friendly fighters to non-friendly entities entering the battlespace. For those in the debrief condition, the Debrief tool was used to playback the scenario (during training) and participants were shown their performance.

Labeling Neutral Entities

Participants were instructed to label any neutral entity that appeared on the radar scope promptly and appropriately. This required a high degree of situational awareness due to the large number of radar tracks. The complexity of a scenario also prompted a subject to fixate on a small portion of the battlespace. The accuracy and latency with which the participants labeled these entities was assessed. During training, participants were given feedback regarding how quickly and accurately they labeled neutral entities. For those in the debrief condition, the Debrief tool was used to playback the scenario in order to point out the participants' mistakes.

Battlespace Management

In one test scenario, the student was instructed to re-task fighter aircraft away from the initial combat air patrol station. Moving the fighters created a gap in air defenses, possibly allowing an incursion into protected air space as shown in Figure 2. The student was expected to notice this vulnerability and re-assign other fighter assets to fill the gap.

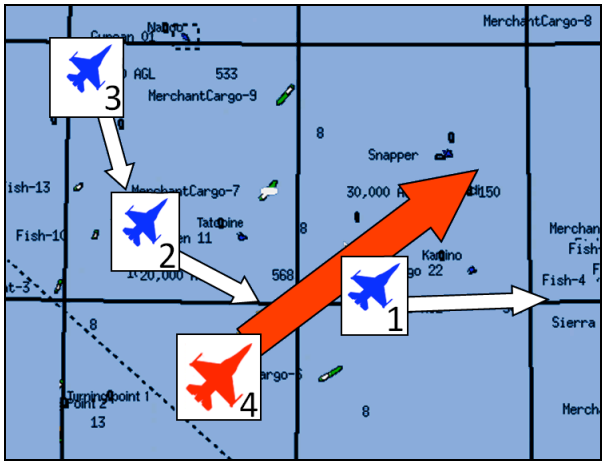

Figure 2: Battlespace Management. In this battle problem, Fighter 1 is re-assigned to the East, leaving a gap in air defenses. The student should move Fighters 2 and 3 to fill the gap; otherwise, enemy Fighter 4 may penetrate the defenses.

At this time, AEMASE could not recognize speech from radio calls, so the automated assessment was based on analysis of readily available simulation data, such as the positions of friendly and enemy fighters over the course of the scenario. One method used to represent this data was an Occupancy Grid, shown in Figure 3. The battlespace was divided into a grid and the total amount of time spent in each grid cell by friendly and enemy fighters was computed, resulting in two matrices of time-weighted values. This approach is more informative than simple "snail trails" left behind by each entity because it captures information about how much time an entity spends at a location.

During training, participants were given feedback regarding whether or not they correctly re-tasked friendly fighters. Those in the debrief group were also

shown how their AEMASE Occupancy Grid differed from an expert's Occupancy Grid (Figure 3).

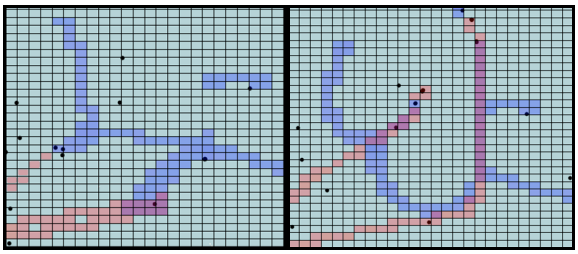

Figure 3: Occupancy Grids. Blue and red tracks show the paths of friendly and opposing forces, respectively. On the left, friendly forces were pre-positioned correctly and repelled the incursion. On the right, gaps in defenses allowed the penetration of protected airspace.

RESULTS

The two groups' performance on the three metrics were compared using the t-test.

Fleet Protection

Participants in the debrief group committed their friendly assets to a potential threat much sooner (t = 2.03, p<0.05) than did the control group (Figure 4).

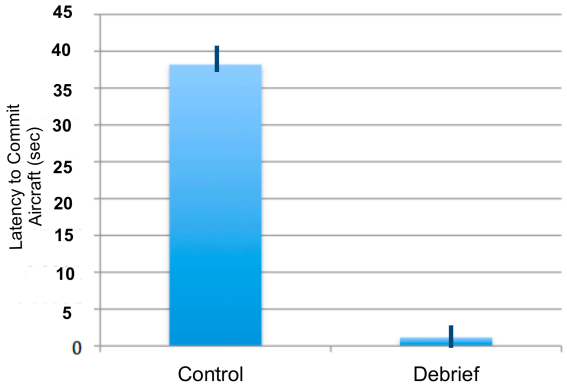

Figure 4: Fleet Protection – Response Time. When enemy aircraft approached the friendly aircraft carrier, the debrief group took significantly less time to respond (t=2.03, p<0.05).

Labeling of Neutral Entities

Participants in the debrief group labeled neutral entities both significantly more quickly (t=1.69, p<0.05) and more accurately (t=1.87, p<0.05) than did the control group (Figure 5).

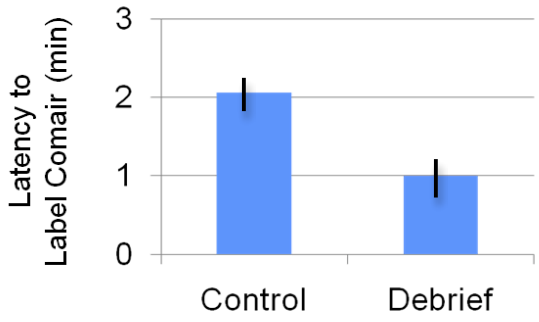

Figure 5: Labeling Neutral Aircraft – Response Time. The debrief group responded to the appearance of neutral aircraft in significantly less time (t=1.69, p<0.05).

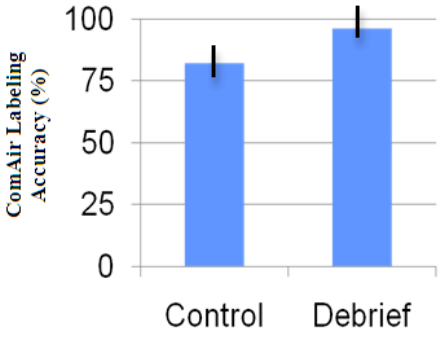

Figure 6: Labeling Neutral Aircraft – Accuracy. The debrief group correctly identified neutral aircraft more often (t=1.87, p<0.05).

Battlespace Management

In the test scenario, very few participants in either group re-positioned fighter aircraft correctly, as specified by the subject matter expert (Figure 7). There was no difference between the two groups. From this data, it is not possible to determine which group would have achieved competency more rapidly.

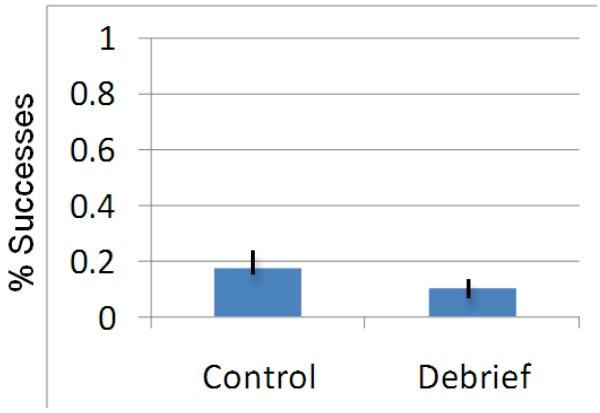

Figure 7: Battlespace Management. In a complex scenario, very few students in either group ordered their fighter aircraft to positions consistent with those specified by our subject matter experts, and there was no significant difference between the groups.

DISCUSSION

For two of three metrics, the group who received feedback via an AEMASE-based debrief tool (featuring graphical depictions of student performance) performed statistically better than the control group who simply received verbal feedback. This provides evidence that this tool facilitates training targeted at individual performance deficits. These results suggest that tools may decrease the cost of training to a fixed level of proficiency, either by increasing the student/teacher ratio, or decreasing the amount of time required. While AEMASE was the focus of this paper, these results generalize to similar systems.

Our next research objective for AEMASE is to support team training. We will identify team performance metrics consistent with the Team Dimensional Training (TDT) Paradigm (Smith, 1998), and enhance the capability of AEMASE by integrating speech recognition software to analyze communications between team members.

We have already performed preliminary speech analysis looking at novices vs. experts for a very similar experiment. In this experiment, two pairs of experts and novices performed numerous test sessions on the EDRT. The speech of both expert and novice teams were recorded. We hypothesized that the language of the teams would be useful in discriminating between experts and novices and are in the process of evaluating the speech of the novice vs. experts. This work is inspired by earlier research in which TF/IDF with Latent Semantic Analysis was highly effective in automated essay grading, despite disregarding the order of word usage (Foltz, Laham & Landauer, 1999). Our primary concern here is whether the approach will still be effective given the limited accuracy of automated speech recognition. We have achieved similar speech recognition rates across a variety of open-source and commercial speech recognition systems, ranging from 90% under near-ideal conditions to 50% for certain speakers in the presence of background noise. This limitation motivates our statistical (rather than grammar-based) approach.

Preliminary results have been achieved from a single scenario that was manually transcribed by one of the authors. All eight participants' speech was transcribed and put in a document. Each document was transformed into a *term vector* where each element represents a term and the value represents a weighting of the term. We used the popular term frequency-inverse document frequency (tfidf) weighting which incorporates the proportion of times a term was used weighted by the number of documents it appeared in. By calculating the cosine similarity between document term vectors we can evaluate how similar the speech of the teams were. Preliminary results indicate that for 7 of 8 subjects, the two most similar subjects (using cosine-similarity of term vectors) were in the same category of expertise (novice vs. expert).

Based on these results, we started an analysis of the documents using the perceptron algorithm. The goal here is to try to learn a classifier which would take as input a set of terms and output a decision on whether these terms were generated from an expert or a novice. Preliminary results indicate we can achieve low error rate on the documents, and analysis of the perceptron weight matrix has provided us with an idea of which terms indicate experts vs. novices. These results are preliminary and will be presented in later work.

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. This work was performed through a contract award from the Office of Naval Research.

REFERENCES

Abbott, R. G. (2002). Automated tactics modeling: Techniques and applications. *Dissertation Abstracts International, 68*.

Abbott, R. G. (2006). Automated expert modeling for automated student evaluation. *Intelligent Tutoring Systems, 4053*, 1-10.

Anderson, C. W., Draper, B. A. & Peterson, D. A. (2000). Behavioral cloning of student pilots with modular neural networks. *In ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning,* (pp. 25-32). San Francisco: Morgan Kaufmann.

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.

Foltz, P. W., Laham, D., Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. *Proceedings of EdMedia '99.*

Jensen, R., Nolan, M. & Chen, D. Y. (2005). Automatic causal explanation analysis for combining arms training AAR. *In Proceedings of the Industry/Interservice, Training, Simulation and Education Conference (I/ITSEC).*

Murray, T. (1999). Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education, 10*, 98 – 129.

Smith-Jentsch, K. A., Zeisig, R. L., Acton, B., McPherson, J. A. (1998). Team dimensional training: A strategy for guided team self-correction. In E. Salas and J. A. Cannon-Bowers (Eds). *Making Decisions under Stress: Implications for Individual and Team Training,* (pp. 271-297). Washington, D.C.: APA.

Stevens, S. M., Forsythe, J. C., Abbott, R. G. & Gieseler, C. J. (2009). Experimental assessment of accuracy of automated knowledge capture. *Foundations of Augmented Cognition, Neuroergonomics, and Operational Neuroscience, 5638*, 212-216.