DARCAAT: DARPA Competence Assessment and Alarms for Teams

Peter Foltz, Mark Rosenstein	Noelle LaVoie	Rob Oberbreckling
Pearson Knowledge Technologies	Parallel Consulting	Perceptive Research
Boulder, Colorado	Longmont, Colorado	Boulder, Colorado
pfoltz@pearsonkt.com, mbr@pearsonkt.com	lavoie@parallel-consulting.com	rob.oberbreckling @gmail.com

Ralph Chatham	Joseph Psotka
ARPA Consulting	U.S. Army Research Institute
Falls Church, Virginia	Arlington, VA
ralph.chatham@verizon.net	Joseph.psotka@hqda.army.mil

ABSTRACT

Assessing teams in complex military environments requires effective tracking of individual and team performance. Indeed, performance measures must be both accurate and timely in order to provide effective real-time alarms. However, current methods of monitoring team and group performance often rely on delayed outcomes or global metrics that are insufficiently detailed to detect failures until recovery is impossible, and are often unable to reveal the causes of failures. An untapped source of timely and diagnostic information lies in the communications among team members. The DARCAAT program developed and tested a toolset for automating team assessment and near real-time alarms. The toolset uses Automated Speech Recognition and Statistical Natural Language-based techniques for embedding automatic, continuous, and cumulative analysis of team communication in training and operational environments. The techniques include measures of the content, patterns, and style of team members' communications. These measures were combined using machine learning techniques to develop performance models based on Subject Matter Expert (SME) ratings of teams.

Focusing on the domain of convoy training, we collected team performance and communication data from the Fort Lewis DARWARS Ambush! convoy training virtual environment and from the National Training Center's live convoy STX lane training. Tests of the performance models and critical incident detection capabilities showed that the technology agreed significantly with SMEs' ratings of teams, and could identify a majority of the team critical incidents. In this paper we discuss the implications for modeling team performance based on communication, describe the development of the technology, and demonstrate how it can process communication to detect critical incidents and to generate team performance metrics. Finally we describe how this technology can be integrated into training systems for automatic team assessment. These systems can provide automated feedback and can alert teams and commanders of potential problems before they occur.

ABOUT THE AUTHORS

Peter W Foltz, Ph.D. is founder and Vice President for Research at Pearson Knowledge Technologies and Senior Research Associate at the University of Colorado, Institute of Cognitive Science. He was previously a professor at New Mexico State University. His research has focused on computational modeling of knowledge, team research, and technologies for automated training assessment. He has published a range of articles on Team assessment, information retrieval, natural language processing, training technology, clinical diagnosis, and cognitive modeling. Peter has served as principle investigator for research for the Army, Air Force, Navy, DARPA, National Science Foundation, and Intelligence Agencies. Contact information: Pearson Knowledge Technologies. 4940 Pearl East Circle, Suite 200, Boulder, CO, 80305. pfoltz@pearsonkt.com

Mark Rosenstein is a Senior Member of Technical Staff at Pearson Knowledge Technologies applying machine learning and natural language processing techniques to problems involving understanding and assessing language and the activities connected with the use of language. Contact information: Pearson Knowledge Technologies. 4940 Pearl East Circle, Suite 200, Boulder, CO, 80305, mbrmbr@acm.org.

Noelle LaVoie is a founder of Parallel Consulting, LLC where she acts as the lead Cognitive Psychologist. Parallel Consulting specializes in combining qualitative and quantitative methodologies in conducting applied social science research. Previously Noelle held the position of Senior Member of Technical Staff at Pearson Knowledge Technologies, where she focused on developing innovative applications of Latent Semantic Analysis (LSA) and other machine learning technologies. These included military applications involving tacit knowledge based assessment of military leadership, online collaborative learning, visualization tools to support multinational collaboration and design of interactive electronic manuals. Noelle received her Ph.D. in Cognitive Psychology from the University of Colorado, Boulder, in 2001. Contact information: Parallel Consulting, 806 Bowen Street, Longmont, CO 80501, lavoie@parallel-consulting.com.

Rob Oberbreckling is a founder of Perceptive Research Inc. His interests include applying software systems to problems in cognitive science, natural language processing, machine learning, audio signal processing, and automated human performance measurement, modeling, and assessment. Robert previously was a Senior Member of Technical Staff at Pearson Knowledge Technologies where he led team communication data collection efforts in the field as well as created predictive systems for individual and team performance for commercial and military applications. Contact information: Perceptive Research Inc., 3050 24th St., Boulder, CO 80304, rob.oberbreckling@gmail.com

Ralph Chatham is a physicist, storyteller, all-purpose curmudgeon, and lately program manager for the Defense Advanced Research Projects Agency. He is currently a private insultant, delivering advice on technology development, and training in the Defense Department. He has been a submarine officer, laser builder and chairman of two task forces of the Defense Science Board, herding DoD elephants to explore the issues of training superiority and training surprise. He has managed, either inside or outside the government, contract research on: putting lasers in space to talk to submarines patrolling under water and clouds; synthetic aperture sonar; real science applied to detecting deception; and digital tools, games and simulations for training such things as language and information technology troubleshooting. He created and managed from afar the research program discussed in this paper. In addition to the Defense Superior Service Medal, the Secretary of Defense Medal for Exceptional Public Service and other DoD award, Ralph and his wife jointly received a 2003 National Storytelling Network Oracle Award. Contact Coordinates: 2631 Kirklyn Street, Falls Church, VA 22043; 703 698 5456; ralph.chatham@verizon.net.

Joseph Psotka is a Program Manager for basic and applied research in behavioral and social sciences at the Army Research Institute. He earned a Ph.D. degree in cognitive psychology from Yale University in 1975. He taught at several colleges and universities, including Southern Connecticut State College and the University of Waterloo, before becoming Director of Research at NPSRI in Alexandria, Va. in 1978. He was made a Resident Scholar of the National Institute of Education (NIE) in 1981. Dr. Psotka joined the Army Research Institute in 1982 as a team chief within the Training Laboratory, where he has remained. In 1988 his edited volume on Intelligent Tutoring Systems: Lessons Learned was published. His research now focuses on social network analysis, LSA and automated text understanding, leadership, communities of practice, unobtrusive measurement technologies, automated tutoring by intelligent agents, simulation technologies, and higher order thinking. Contact information: Joseph Psotka, US ARI, 2511 Jefferson Davis Highway, Arlington, VA 22202-3926. joseph.psotka@hqda.army.mil

DARCAAT: DARPA Competence Assessment and Alarms for Teams

Peter Foltz, Mark Rosenstein	Noelle LaVoie	Rob Oberbreckling
Pearson Knowledge Technologies	Parallel Consulting	Perceptive Research
Boulder, Colorado	Longmont, Colorado	Boulder, Colorado
pfoltz@pearsonkt.com mbrmbr@acm.org	lavoie@parallel-consulting.com	rob.oberbreckling @gmail.com

Ralph Chatham	Joseph Psotka
ARPA Consulting	U.S. Army Research Institute
Falls Church, Virginia	Arlington, VA
ralph.chatham@verizon.net	joseph.psotka@hqda.army.mil

TEAM COMMUNICATION AND PERFORMANCE

Monitoring teams of decision-makers in complex military environments requires effective tracking of individual and team performance. However there are numerous challenges to effectively identify, track, analyze, assess, and report on team performance in real-time in complex operational environments. For example, current methods of assessing team and group performance often must rely on temporally delayed outcomes or global metrics. These metrics often lack information rich enough to diagnose failures, detect critical incidents, or suggest improvements for the teams for use in performing their tasks. An untapped source of more timely and diagnostic information lies in the ongoing communications among team members.

Team members who communicate with each other provide a rich source of information about their performance. The communication data includes information both about the structure of the social network and the content and quality of information flowing through the network. The structure and communication patterns of the network can provide indicators of team member roles, paths of information flow and levels of connectedness within and across teams. The content of the information communicated provides detailed indicators of the information team members know, what they tell others, and their current situation.

Additionally, communication data can provide information about team cognitive states, knowledge, errors, information sharing, coordination, leadership, stress, workload, intent, and situational status. Indeed, within the distributed training community, trainers and subject matter experts typically rely on listening to a team's communication in order to assess that team's performance. A number of studies have shown that communication provides valuable indicators of team performance. For instance, Achille, Schulze and Schmidt-Nielsen (1995) found that the use of military terms, acknowledgments, and identification statements increased with experience. Similarly, Jentsch, Sellin-Wolters, Bowers and Salas (1995) found that teams that identified typical flight problems faster made more leadership statements and more observations about the environment than slower teams. Coding of communications has shown that team performance is significantly associated with the frequency, sequences and types of task-related communications as well as the appropriate use of social markers such as acknowledgements (see, Bowers, Braun & Kline, 1994; Bowers, Jentsch, Salas, & Braun, 1998; Fischer, McDonnell & Orsanu, 2007; Kiekel et al., 2002; Kiekel et al., 2004).

However, to effectively exploit the information inherent in communication data, technologies are needed that can assess both the content and patterns of the verbal information flowing in the network and convert the analyses into straightforward metrics that are usable by teams and commanders. With the advent of improved natural language processing, computational semantics, automated speech recognition and machine learning techniques, it is feasible to develop automated techniques to analyze team communication and predict performance. Applying such techniques to training would permit the development of low-cost tools that could automatically and unobtrusively monitor, assess, and provide feedback to team members and trainers.

Objectives

The primary objective of the DARPA Automated Competence Assessment and Alarms for Teams (DARCAAT) program was to develop and validate a

toolset for embedding automatic, continuous, and cumulative analysis and assessment of verbal interactions in team training and operational environments. The goal was to create the toolset and implement it as a real-time team performance alarm system using natural language, statistical, and other analyses of team communications, and then test it with convoy training data collected from the DARWARS Ambush! simulation at Fort Lewis and the National Training Center (NTC). Once built, the objectives were to test the performance of the system on the collected datasets and to evaluate the feasibility of developing a toolset for low-cost automated performance assessment and alarms.

Automated Communication Analysis

Automated verbal communication analysis involves applying a set of computational modeling approaches to networked communication in order to characterize the verbal communication as useful assessments of performance. These characterizations could include metrics of team performance, feedback to commanders, or alerts about critical incidents related to performance. This type of analysis has three prerequisites. The first is the availability of sources of clear verbal communication. Second, there must be performance measures which can be used to categorize, rank or quantify the communication in terms of actual team performance. Finally, these prerequisites can be combined with a set of computational approaches applied to the communication in order to perform the analysis. These computational approaches include computational linguistics methods to analyze communication, machine-learning techniques to associate communication to performance measures, and finally cognitive and task modeling techniques.

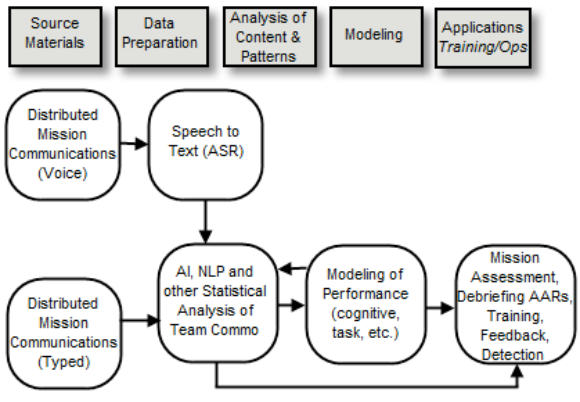

Figure 1. The Communication Analysis Pipeline

By applying several computational approaches to the communication, we have a complete communication analysis pipeline as represented in Figure 1. Proceeding through the tools in the pipeline, spoken and written communication are converted directly into performance metrics which can then be incorporated into reports and visualization tools. This analysis makes possible applications that can support commanders and Soldiers such as near-real-time alerts of critical incidents, timely feedback to commanders of poorly performing teams, graphic representations of the type and quality of information flowing within a team and automatically augmented AARs and debriefings. This paper outlines the overall approach, reports on results from tests of its application to automatically convert team communication into effective and accurate performance metrics, and discusses how it can be integrated into training and monitoring applications.

MODELING APPROACH

In order to process communication, technology is needed that can "understand" the meaning of what is being conveyed in the communication. The primary underlying technology used in this analysis is a method for mimicking human understanding of the meaning of natural language called Latent Semantic Analysis (LSA), (see Landauer, Foltz & Laham, 1998 for an overview of the technology, and Foltz, 2005 for its application to team communication analysis). LSA is automatically trained on a body of text containing knowledge of a domain, for example a set of training manuals and/or domain relevant verbal communication. After such training, LSA is able to measure the degree of similarity of meaning between two communication utterances in a way that closely mimics human judgments. This capability can be used to understand the verbal interactions in much the same way that a subject matter expert can compare the performance of one team or individual to others. The results from the LSA analysis are combined with other computational language technologies which include techniques to measure syntactic complexity, patterns of interaction and coherence among team members, audio features, and statistical features of individual and team language (see Jurafsky & Martin, 2008 for an overview of approaches to language analysis). These features include measures that examine how semantically similar a team transcript is to other transcripts of known quality, measures of the semantic coherence of one team member's utterance to the next, the overall cohesiveness of the dialogue, characterizations of the quantity and quality of information provided by team members, and measures of the types of words chosen by the team members.

The computational representation of the team language and the team performance ratings are then combined with machine-learning technology to predict team performance metrics. Machine learning techniques including hill-climbing methods such as stepwise regression, discriminant analysis, and Support Vector Machines (SVMs) are then used to determine the language features that best model the performance metrics without overfitting the data. In a sense, these methods learn which features of team communication are associated with different metrics of team performance and then, predict team performance metrics for new sets of communication data.

PRIOR WORK

Individual components of the communication analysis pipeline have been previously researched and tested. Over a series of studies, computational language-based communications methods have been evaluated favorably in terms of their ability to predict team performance. For instance, they are successfully able to predict team performance scores in simulated task environments based only on communication transcripts (Foltz, Lavoie, Rosenstein, & Oberbreckling, 2007; Foltz, 2005; Foltz, Martin, Abdelali, Rosenstein & Oberbreckling, 2006; Gorman, Foltz, Kiekel, Martin & Cooke, 2003; Kiekel, Cooke, Foltz, Gorman & Martin, 2002; Kiekel, Gorman & Cooke, 2004). Using human and ASR transcripts of team missions the methods predicted both objective team performance scores and SME ratings of performance at very high levels of reliability in a UAV environment, in simulators of F-16 missions, and in Navy TADMUS exercises.

Overall, the results from prior research indicate that this technology can provide a robust approach to the development of a system for automated analysis of team verbal communication. While the prior work shows that individual components can succeed, the present work sought to build the combination of the technologies into a single modeling and performance prediction toolset running on data collected from live and virtual military events.

DATA COLLECTION

In order to develop and test the tools, convoy lane training was chosen as a domain because it is a critical component of effective operations in Iraq and Afghanistan. Convoy lane training involves interacting teams of 5 to 40 persons, with disparate pieces of knowledge, and includes command and control and situation awareness issues (see Kuhn, 2004). The goal was to observe convoy training and collect communication data along with other indications of the performance of the teams, including videos and event logs.

Team audio data collection

Two communication datasets were collected and analyzed during this effort. In collaboration with the Fort Lewis Mission Support Training Facility, the project team collected audio, video and meta-data from the DARWARS Ambush! virtual environment convoy training activities. DARWARS Ambush! is a widely used game-based training system that has been integrated into training for many brigades prior to deployment in Iraq (Diller, Roberts, Blankenship & Nielson, 2004; Diller, Roberts & Wilmuth, 2005). In Ambush!, up to 50 Soldiers jointly practice battle drills and leadership during simulated convoy operations. A second data set consisting of data from live mounted convoy STX lane training was collected at the National Training Center (NTC), Fort Irwin. In collaboration with the NTC Observer/Controllers (O/Cs) we collected performance assessments of the datasets and recorded AARs and hot washes from the live training exercises. Both data collection efforts concentrated on platoon and squad-level teams performing convoy operations.

We collected over 250 DARWARS Ambush! training missions on of approximately a half hour apiece including VOIP audio communication, and video and event logs in some cases. At the NTC, we collected voice activated recordings of SINCGARS FM communications during STX lane training. Data was collected during rotations from January through June of 2007. We recorded a total of 105 STX lane training missions, of which we selected 57 recordings that had acceptable quality audio, and training events of interest. Combined, this resulted in approximately 300 convoy training missions.

SME performance rating collection

Providing feedback on team performance requires the toolset to automatically associate performance metrics with communication streams. Thus, the system typically requires one or more metrics of team performance, which can include objective measures of performance, such as threat eliminations or mission objectives completed, or subjective measures of performance, such as Subject Matter Experts' (SME) ratings of aspects of performance including command and control and situation awareness. In both the Ambush! and NTC convoy training contexts, evaluation occurred as part of the AAR process, so it was important that the performance metrics were drawn from the same task context, and developed in

conjunction with SMEs with extensive experience working with convoys.

We developed five scales that captured the important dimensions of performance in this domain based on a mission essential task list (METL): command and control (C2), situation understanding (SA), adherence to standard operating procedures (SOP), battle drills (CA) and general team performance (TEAM). Seven SMEs rated the audio collected from Fort Lewis and NTC on these scales using a rating tool developed for the project that presented the audio in a visual interface to allow SMEs to select segments of audio and complete their ratings. The SMEs were also asked to distinguish between critical events, defined as events that change the scope of battle, the commander's plan or disrupt the operational tempo, and other training events in the communication. Finally, SMEs conducted AARs for every mission they rated, providing sustains, improves and ratings on each scale for the entire mission.

Before using SME ratings as a performance measure, it is important to assess how well the SMEs agreed with each other. All SMEs were asked to rate a pair of missions selected for the purpose of collecting data to compute reliability and agreement. Intraclass correlations among the SMEs ranged from .76 to .85 ($p<.001$) for average items suggesting excellent reliability. The intraclass correlations for single items ranged from .38 to .66 ($p<.001$). Exact agreement (two SMEs agree on the exact score) was calculated between every pair of SMEs, and average exact agreement ranged from 24% to 50%. Average adjacent agreement (SMEs agree within one score point) ranged from 74% to 96%. Two SMEs had extremely high agreement, with their adjacent agreement ranging from 93% to 100%, and exact agreement ranging from 51% to 86%. The agreement among SMEs was impressive and indicates that the SME ratings are appropriate for computational modeling. It also provides support that SMEs are able to accurately detect performance from communication.

ANALYSES AND MODELING RESULTS

The overall goal was to develop modeling techniques that could convert the speech stream into text and then accurately predict the SME performance ratings including both their rating scales and their indications of critical events during the training. A majority of the data modeling was conducted on a set of 72 training missions which included communication data, speech analysis variables, and SME-selected critical events and ratings of performance.

The entire suite of techniques were integrated into a single, unified toolset. Below we describe the development and testing of each component of the toolset. The components include:

1) Automated Speech Recognition to convert the communication into text for processing
2) Speech signal feature analysis to assess stress
3) Modeling to predict SME ratings of individual events
4) Modeling to predict overall team performance
5) Modeling to predict SME rated critical events

Automated Speech Recognition

The automatic speech recognition (ASR) component to converted the audio into text and to extracted some of the audio features using BBN Technologies' AVOKE STX speech-to-text software system. AVOKE transforms the raw, digitally recorded audio into a machine-readable text transcript for analysis. The software itself is language and domain independent and can be configured to run on different types of data.

Training the ASR system

Many ASR systems, including AVOKE, require preliminary training in the domain and acoustic environment of interest to produce reasonable recognition accuracy rates. The ASR system used here is trained from accurately transcribed audio recordings sampled from the earliest set of collected mission audio. The system inductively "learns" associations between features in the audio signal and the transcribed words that humans interpreted when they listened to and transcribed the audio signal. This process of learned association results in a trained language model. When recognizing speech from new, unheard audio, the ASR software consults the language model to determine which words should be associated with, or recognized from, the audio features found in the new audio.

The DARCAAT training data consisted of over 16 hours of recorded communications from the Fort Lewis Ambush! environment collected prior to data collected at the NTC. Of this, 2 hours were randomly selected and set aside for testing and optimization of the language model. The training set was recorded on the same hardware used during similar task scenarios as those analyzed by the toolset. Humans transcribed training audio by hand and then an ASR model was built and tested

ASR Accuracy and Evaluation

In order to test the ASR recognition performance, the system was trained on 16 hours of command net utterances. In Ambush! both the command net and intra-vehicle nets are recorded. While valuable

information is contained on the vehicle net, for this analysis only the command net audio was used. A set of 802 utterances were held out from the ASR training set and this set was then run through the trained ASR system and compared against the human transcribed transcript. Word error rate was calculated as the sum of the insertions, deletions and substitution errors made by the ASR system divided by the total number of words.

Overall, we found a word error rate of 33.7%. These error rates are consistent with results found for Speech In Noisy Environments (SPINE) evaluation (see Schmidt-Nielsen et al., 2001). Prior modeling work suggests that this range of error rate may decrease system prediction accuracy by about 10% from verbatim transcripts, which can still provide acceptable performance predictions (see Foltz, Laham & Derr, 2003). Thus, the resulting ASR system's performance is well within the range of what cold be expected from an ASR system in this domain.

There are further steps that can be taken to improve ASR performance. Post-processing can enhance performance on ASR, including techniques such as weighting performance prediction scores by ASR confidence, recalculating error rates ignoring function word errors that do not affect measurement of the team performance context, and doing automated re-insertion of words based on LSA-based predicted context. In addition, training could be done with greater amounts of command net data in order to improve ASR performance. Thus, the resulting ASR system's performance is within the range of what could be expected of an ASR system in this domain. So while the ASR performance is certainly functional in the DARCAAT application, the above issues have reasonable solutions that could boost performance. The text-based output of the ASR system was then used in the subsequent performance modeling.

Speech Feature Analysis

Voice stress analysis examines the physiological changes that a person's stress level causes including micro-muscle tremors (MMT) in the vocal tract muscles. These MMTs can affect the energy and frequency of the speech signal, (see Lippold, 1971; Hanson et al., 2002). Voice Stress analysis has been tested for deception detection with moderate success (see Haddad et al., 2002; Hopkins, Benincasa, Ratley, & Grieco, 2005), but not for predicting performance in teams. While in deception detection a speaker is trying to hide effects of stress, in a team communication situation, stress does not always need to be hidden, and indeed may help to convey urgency, failures, or degree of criticality in a situation. Thus, with appropriate analyses it may be possible to detect stress features in team communication.

In our approach we used a number of statistical transformations of the speech signal in order to detect how likely it is that stress was present in a segment of team communication. Based on the analysis of the speech samples Hidden Markov Models (HMM) were used to categorize speech as excited or neutral. The primary features that were used in the models were measures of power, pitch, change over time, frequency components (FFT/MFCC), rate, duration and frequency of speech and their changes over time.

Overall, the results show that an excitement classification algorithm can work with 87% accuracy for female voices and 81% accuracy for male voices. Of course, just being able to detect excitement in an utterance does not fully determine whether there is a critical event, or whether a team is performing poorly or well on a particular team performance metric. However, these results suggest that the method can provide accurate predictions that can be incorporated with the other methods described below to help detect critical events and help tune the performance models.

Predicting Performance Metrics for Events

During training, it is important to be able to determine how a team is doing at any point in time or in any particular event they are completing. During the rating process, SMEs identified spans of time as "events" and then provided ratings on one or more of the metrics for that event. Typical events ranged from a minute to five minutes in duration. Using a version of the final dataset divided up by events, we developed automated prediction models in which we trained the system on the communication of 80% of the events randomly chosen and then tested predictions on the remaining 20% of the events. In this testing, we used both text-based variables (those based on analysis of the ASR transcripts, using semantic, syntactic and statistical text analyses), and variables based on the speech analysis (those analyzing the audio features of the communication during that event). For the speech analysis variables, the variables represented either means or deviations in the speech variables across the whole event. Stepwise regression was then used to select the best variables for each model.

Table 1. shows the correlation between the model's predicted rating and the SMEs ratings of the events using a model that combined text and speech variables. The correlations are all significant, and ranged from .36 to .41. These are slightly lower than the .38 to .66 found in the intraclass correlations for single items for

the SMEs. Nevertheless, they do show that the model can provide fairly accurate predictions of a team's performance at the event level. The speech and text-based variables were also modeled separately. Generally, combining the speech and text variables improves these correlations by about .03. This indicates that there is considerable colinearity between the speech and text variables, with similar information about team performance carried in both. Overall, the results show that given a short segment of communication (e.g., a few minutes), the system can automatically generate a prediction of how that team is performing.

Table 1. Correlation Between SME Ratings and Model Predictions Using Text and Speech-Based Variables Per Event

Metric	R	N	p value
CA	.37	572	<.001
CC	.41	838	<.001
SA	.41	833	<.001
SOP	.43	886	<.001
TEAM	.36	799	<.001

Overall Team Performance Prediction

Team performance was also modeled for entire missions, not just the separate training events, based on the ratings of the two SMEs with the highest agreement. In this approach the models predicted the SMEs' overall ratings of the teams for each mission using the text-based variables.

Because the unit of analysis for this model was the entire mission, and the agreement results for the SMEs were reported using events as the level of analysis, additional agreement measures were calculated based on the team performance ratings for entire missions rated by both of the SMEs. Table 2 shows that the model's predictions correlated well with the SME ratings, with correlations ranging from .70 to .81 across the five scales, only slightly lower than the correlations between the two SMEs. Adjacent agreement between the SMEs and the model was also quite high, strongly supporting the use of the model in the toolset for assessing a team's performance.

These performance numbers match quite well those generated from the inter-rater reliability analysis. The exact inter-rater reliability range for SMEs was 33.3% to 66.6% which is quite similar to the combined range in Table 2 of 43.1% to 60.8% and the adjacent range from was 66.7% to 100.0% which is bettered here ranging from 86.3% to 100%. The SME correlations were slightly better than the model's with 3 in the .8 range, one at .93 and the last at .59, while the correlations here are clustered mostly in the .7s. Nevertheless, these correlations are quite strong, showing that this analysis technique can account for 50 to 69 percent of the variance of the overall team performance. The results further show that there is more communication to analyze (e.g. whole missions), the techniques are able to provide more accurate predictions than for smaller amounts of communication, such as individual events. This result holds true for the SMEs as well.

Table 2. Prediction Performance between the model and SMEs for overall team performance

Metric	Correlation	Exact agreement	Adjacent agreement
C2	0.71	43.1	86.3
CA	0.74	47.1	90.2
SA	0.83	54.9	100.0
SOP	0.73	47.1	94.1
TEAM	0.78	60.8	94.1

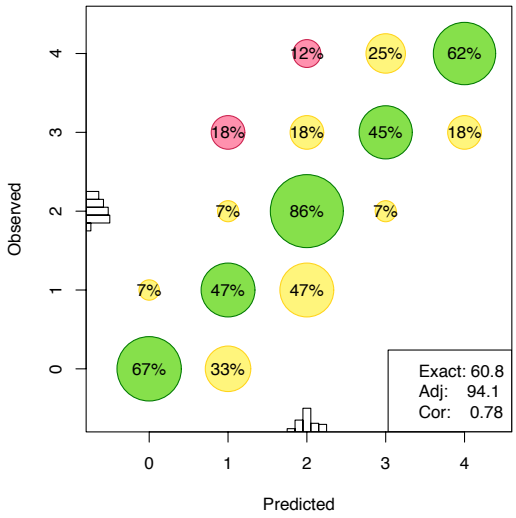

Figure 2. Comparison of SME Rating versus Predicted Rating Overall Team Performance-Combined Fort Lewis and NTC missions

Figure 2 shows an example of the fit between the SME ratings and the prediction model for the overall Team score. The SME ratings are on the y-axis and the model predicted values on the x-axis. The area of the green circles along the diagonal represents exact agreement, the area of the yellow off-diagonal elements represents adjacent agreement and the area of the red circles

represents non-adjacent agreement. The main point that the plots reveal is that the models predict quite well with most of the ratings at the exact and adjacent levels, and only a few small red circles of non-adjacency. The overall implications of the results are that, based on the communication from a mission, the model can accurately determine how well the team performed.

Critical Event Detection

A critical event is anything that changes the scope of battle, the commander's plan or disrupts the operational tempo. Such changes are important in training since teams and/or commanders may not notice the change or may not respond appropriately to the change. Thus, it is important to be able to identify critical events so as to be able to assess performance during that period of time. One can then be able to later play back the events that lead up to the critical event to provide feedback in an AAR.

A spectrum method was used to analyze the team communication data to predict SME-rated critical events. A spectrum method uses windows or time slices as the unit of analysis; all utterances within a window are considered one instance or data point. The windows were created by segmenting the communication by a constant interval of time. Analyses are then conducted moving across the windows of data. In our experiments, overlapping windows were used to model the transitions in communications during a mission.

The critical event detection model was trained on the text-based communication features found in critical events and then tested on a held-out communication data set. A model was developed that would detect whether any of the windows across an event were predicted to be critical events. Over all the Fort Lewis and NTC missions, 81% of critical events were detected with a 22% false positive rate (ROC area under the curve was 95.6%).

The approach shows that a significant number of the critical events can be automatically detected. In addition, the sensitivity can be adjusted so that more critical events could be detected, although with higher levels of false alarms. Varying the sensitivity may be useful in cases when a commander needs to be alerted to any kind of team anomaly as well as cases where sensitivity could be reduced so that teams or commanders are alerted only if the system is highly confident that a critical event is occurring.

USING THE DARCAAT TOOLKIT FOR AARS

As a demonstration of the application of the DARCAAT toolset, a prototype After Action Review application was developed that could be integrated into a training program to allow Observer/Controllers (O/Cs) and commanders to monitor teams and receive feedback on the team's performance. The application processes the incoming communication data from a team and then allows an O/C or commander to load any mission and provides immediate access to several critical pieces of information. Through a series of graphical representations of events in the mission, it enables efficient automatic augmentation of AARs by assisting the O/Cs in selecting the most appropriate segments of missions to illustrate training points.

Using the automated models described in this paper, the application automatically rates a unit for each detected event on the five scales: command and control, situation understanding, use of standard operating procedures and battle drills. For each rating scale, the application selects appropriate training events that reflect the units' range of performance from untrained, through practiced, to trained. The application's interface makes it easy to spot performance weaknesses at a glance and then to drill down to understand these weaknesses by listening to the relevant radio communication. The application also enables commanders to create a custom AAR by selecting events of interest and the associated radio communication and then adding their own comments.

The application was designed to provide additional support to the AAR process by essentially extending an O/C's reach automatically. Two SMEs reviewed the AAR application in order to provide us with feedback about its usefulness in supporting AARs, suggested improvements, and other possible applications. Both SMEs thought the AAR application was valuable and would reduce the time required to prepare for an AAR, as well as increasing the scope of events that could be discussed. The SMEs also believed that the application could easily be extended to provide an O/C or commander support beyond a typical training mission AAR in order to track team improvement longitudinally over time and detect performance trends. Overall, application should allow O/Cs to be more efficient at locating training issues and spend more time interacting and monitoring trainees. Additional details of the AAR application can be found in LaVoie et al., (submitted).

CONCLUSIONS

The content and patterns of a team's communication provide a window into performance and cognitive states of the individuals and the team as a whole. By applying computational analyses to the communication stream, we can automatically derive team performance metrics. The DARCAAT program demonstrated the feasibility of using this approach for automatically detecting critical incidents, identifying performance changes, and evaluating team performance in both live and virtual training environments.

The system uses a Statistical Natural Language-based intelligent software methodology for embedding automatic, continuous, and cumulative analysis of spoken interactions for individual and teams in both training and operational environments. Starting with an incoming stream of free-form verbal communication, commercial grade ASR is applied, generating transcribed text and speech characteristics, such as stress, which can, in near real-time (within seconds), be analyzed using previously trained natural language models resulting in detailed measures of team characteristics and performance. This process provides a complete communications analysis pipeline, automatically converting team verbal communications into quantifiable performance measures.

The toolkit allows the analysis and modeling of both objective and subjective performance metrics and is able to work with large amounts of communication data. Indeed, because of its machine-learning foundation, it works more accurately with more data. The toolkit can automatically extract measures of performance by modeling how subject matter experts have rated similar communication in similar situations as well as modeling objective performance measures. Because the technology uses automated machine-learning and natural language approaches, it does not require the time and resources of large amounts of previously hand-coded language analysis or task analysis. This permits rapid and more cost-effective development and application of the technology for novel tasks and situations.

Because this toolset permits low-cost development of team assessment systems, it can be integrated into training for teams, for assessing team and system performance, and for alerting teams and commanders of indications of potential problems before they occur. The tools can provide a range of alerts and feedback including:

- near real-time assessment (within seconds) of individual and team performance;
- indications of situation awareness, knowledge gaps and workload;
- detection of critical events;
- performance alarms;
- generating automated After Action Reviews (AARs).

Potential Applications and New Directions

Computational communication analysis can be easily adapted to other military and commercial applications requiring monitoring and assessment of teams. It allows almost instantaneous analysis and modeling of objective and subjective metrics of team performance for real, complex communication data in networked teams. Because the models are automatically derived, the approach does not require large up front task analyses and instead simply models team performance in the same manner as SMEs.

The toolset can be easily be integrated into systems to monitor and provide feedback for teams, in both training and operational venues. Such systems can include applications to monitor teams, give feedback, visualize team information flow, alert commanders to potential problems before they occur, track leadership, as well as being integrated into adaptable training systems which can adjust training based on performance of the team.

For example, there is potential to use the toolkit as a leadership training and development tool. A strong leader should be able to evaluate the communication and provide feedback to their unit along many of the same dimensions captured by the toolkit. Using the toolkit, leaders can practice these skills and refine their own abilities to detect critical events and training moments to share with their units. They can also build up more experience in assessing their unit's strengths and weakness, improving their ability to conduct effective AARs and increasing the benefits of training.

The overall approach to communication analysis and performance measurement further aids in understanding the role of communication in complex human and system networks. Results from applying the toolset to teams in real-world situations can help clarify how communication affects team performance, how performance is reflected through communication, and how to employ technology to monitor and improve teams.

ACKNOWLEDGEMENTS

We would like to acknowledge the contributions of the DARCAAT project team including Marita Franzke,

Kyle Habermehl. Brent Halsey, Chuck Pannacione Manju Putcha, Jim Parker, Boulder Labs, David Diller, Laura Leets, Fred Flynn, Cyle Fena, Don Scott, Paul Asuncion, Reginald King, Brian Oman, Len Dannhaus, Nick Hatchel, Rick Travis, David Leyden, and Jamison Winchell. We would also like to thank the following organizations for participating in this work: DARPA DSO, ARI, Fort Lewis, NTC, and Fort Carson. This work was sponsored by the Defense Adanced Research Projects Agency (DARPA) and the U.S. Army Research Institute (ARI).

REFERENCES

Achille, L. B., Schulze, K. G., & Schmidt-Nielsen, A. (1995). An analysis of communication and the use of military terms in Navy team training, *Military Psychology, 7*, 95-107.

Bowers, C. A., Braun, C. C., & Kline, P. B. (1994). Communication and team situational awareness. In R.D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situational Awareness in Complex Systems* (pp. 305-311). Daytona Beach, FL: Embry Riddle Aeronautical University Press.

Bowers, C. A., Jentsch, F., Salas, E., & Braun, C. C. (1998). Analyzing communication sequences for team training needs assessment. *Human Factors, 40,* 672-679.

Diller, D. E., Roberts, B., Blankenship, S. & Nielsen, D. (2004). DARWARS Ambush! – Authoring lessons learned in a training game. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference.* Orlando, FL: I/ITSEC.

Diller, D. E., Roberts, B. & Willmuth, T. (2005). DARWARS Ambush! A case study in the adoption and evolution of a game-based convoy trainer with the U.S. Army. Presented at the *Simulation Interoperability Standards Organization,* 18-23 September.

Fischer, U., McDonnell, L. & Orsanu, J. (2007). Linguistic correlates of team performance: Toward a tool for monitoring team functioning during space missions. *Aviation, Space and Environmental Medicine, 78*, 5, (pp., B86-B95).

Foltz, P. W. (2005). Tools for Enhancing Team Performance through Automated Modeling of the Content of Team Discourse. In *Proceedings of HCI International*, 2005.

Foltz, P. W., Laham, R. D. & Derr, M. (2003). Automated Speech Recognition for Modeling Team Performance. In *Proceedings of the 47th Annual Human Factors and Ergonomic Society Meeting.*

Foltz, P. W., Lavoie, N., Oberbreckling, R. & Rosenstein, M. (2007). Tools for automated analysis of networked verbal communication. *Network Science Report*, Volume 1, 1 pp 19-24, United States Military Academy Network Science Center.

Foltz, P. W., Martin, M. A., Abdelali, A., Rosenstein, M. B. & Oberbreckling, R. J. (2006). Automated Team Discourse Modeling: Test of Performance and Generalization. In *Proceedings of the 28th Annual Cognitive Science Conference*.

Gorman, J. C., Foltz, P. W., Kiekel, P. A., Martin, M. A. & Cooke, N. J. (2003). Evaluation of Latent Semantic Analysis-based measures of team communications content. In *Proceedings of the 47th Annual Human Factors and Ergonomic Society Meeting.*

Haddad, D., Ratley, R., Walter, S. & Smith, M. (2002). Investigation and Evaluation of Voice Stress Analysis Technology. *Final Report for National Institute of Justice, Interagency Agreement* 98-LB-R-013. Washington, DC, 2002. NCJRS, NCJ 193832.

Hansen, J. H. L. et. al., (2002). Methods for Voice Stress Analysis and Classification, an appendix to Investigation and Evaluation of Voice Stress Analysis Technology, *Final Report for National Institute of Justice, Interagency Agreement* 98-LB-R-013. Washington, DC, 2002. NCJRS, NCJ 193832.

Hopkins, C. S., Benincasa, D. S. Ratley, R. J. & Grieco, J. J. (2005). Evaluation of voice stress analysis technology. *Proceedings of the 38th Hawaii International Conference on Systems Science.*

Jurafsky, J. & Martin, J. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition,* New York, Prentice Hall.

Jentsch, F., Sellin-Wolters, S., Bowers, C. & Salas, E. (1995). Crew coordination behaviors as predictors of problem detection and decision making times. *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting.*

Kiekel, P. A., Cooke, N. J., Foltz, P. W., Gorman, J & Martin, M. J. (2002). Some promising results of communication-based automatic measures of team cognition. *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting.*

Kiekel, P., Gorman, J., & Cooke, N. (2004). Measuring Speech Flow of Co-located and Distributed Command and Control Teams During a Communication Channel Glitch. *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting,* 683-687.

Kuhn, C. (2004). National Training Center: Force-on-force convoy STX lane. *Engineer: The*

professional bulletin for Army Engineers, April-June.

Laham, D., Bennett, W., & Derr, M. (2002). Latent Semantic Analysis for career field analysis and information operations. *Paper presented at Interservice/Industry, Simulation and Education Conference (I/ITSEC)*, December 2-5, 2002. Orlando, FL.

Landauer, T. K, Foltz, P. W. & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25*(2&3), 259-284.

Lippold, O. (1971). Physiological Tremor, *Scientific American*, Volume 224 (3), March.

Schmidt-Nielsen, A., Marsh, E., Tardelli,, J., Gatewood, P., Kreamer, E., Tremain, T., Cieri, C., Strassel, S. Martey, N., Graff, D. & Tofan, C. (2001). *Speech in Noisy Environments (SPINE2) Part 1 Audio*. Linguistics Data Consortium catalog: LDC2001S04.