AUTOMATED SPEECH RECOGNITION FOR MODELING TEAM PERFORMANCE

Peter W. Foltz [1,2], Darrell Laham[1] Marcia Derr[1]

1 Knowledge Analysis Technologies
Boulder, Colorado

2 New Mexico State University
Las Cruces, New Mexico

While team tasks provide a wealth of data on individual and team performance, techniques for modeling team communication can be quite effortful and time-consuming. Automated techniques of analyzing team discourse provide the promise of quickly judging team performance and permitting feedback to teams both in training and in operations. In previous research, techniques using Latent Semantic Analysis (LSA) have proven successful for analyzing team transcripts. However, converting the audio discourse into transcripts often requires hand transcription. In this work, we describe applying automated speech recognition (ASR) to team transcripts and using the output of the ASR to predict overall team performance. Results indicate that ASR can be used in conjunction with semantic methods of modeling team communication to provide accurate predictions of performance. The work has potential for assisting operators in the performance of their tasks because it can "listen" and in real-time evaluate free-form verbal communication from a variety of sources.

INTRODUCTION

Teams play an increasingly critical role in complex military operations in which technological and information demands require a multioperator environment (Salas, Cannon-Bowers, Church-Payne, & Smith-Jentsch, 1998). These environments generate large amounts of communication data as operators perform their tasks. Thus, to study and assess team performance, this communication stream can be monitored and evaluated. Nevertheless, assessment of teams has been hindered by the fact that the richest source of data, the verbal communication among team members, is difficult to collect and analyze. Prior attempts at coding the content of communication have relied on tedious hand-coded techniques or have used limited data such as frequencies and durations of communications. With the advent of artificial intelligence techniques that can measure the semantic content of communication discourse, novel methods for the analysis of communication can be applied.

In prior research (Kiekel, Cooke, Foltz, Gorman & Martin, 2002, Kiekel, Cooke, Foltz, & Shope, 2001), have successfully developed and tested methods for automatically analyzing discourse from team tasks. Using Latent Semantic Analysis (LSA) as the basis for the analyses of the content of the discourse the research has shown that automated analyses of the semantic content of team communication can provide effective prediction and modeling of team performance. However, a limitation of that research was that it relied on typed transcripts of team discourse. This limits analyses to being done only well after the discourse is generated, rather than in real-time. The purpose of the current research is to investigate the efficacy of combining automated speech recognition (ASR) with LSA-based modeling of team discourse to predict overall team performance. Success in this endeavor would show that it is possible to systematically parse and evaluate verbal communication to identify critical information and content required of many of today's multi-operator environments.. It would have the potential for assisting operators in the performance of their tasks because it can "listen" and in almost real-time evaluate free-form verbal communication from a variety of sources

LATENT SEMANTIC ANALYSIS (LSA)

LSA is a method for automatically extracting and representing knowledge in massive databases of relevant electronic text (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). It was developed through ten years of basic and applied research supported by Bell Communications Research, DARPA, ONR, ARI, NASA, AFRL, the McDonnell Foundation and others. LSA has been extensively validated in both controlled experiments and field tests (Landauer & Dumais, 1997; Landauer, Foltz, and Laham, 1998; Landauer, 1998).

Automated Analysis of Meaning

As a psychological theory of the acquisition, induction, and representation of knowledge, LSA research has provided new insights on how people learn the meanings of words. LSA is instantiated as a mathematical system for computational modeling of cognitive processes. As a tool, LSA is used as an artificial intelligence (machine learning) system useful in various educational and industrial applications.

LSA provides a method for determining the similarity of meaning of words and passages by analysis of large text corpora such as domain knowledge libraries, writing samples, e-mail files, course materials, and job and training historical records. After processing a large sample of machine-readable language, LSA represents the words used in it, and any set of

these words—such as a sentence, paragraph, or essay—either taken from the original corpus or new, as points in a very high (e.g. 300) dimensional "semantic space". LSA is closely related to neural net models, but is based on singular value decomposition, a mathematical matrix decomposition technique closely akin to factor analysis that is applicable to text corpora approaching the volume of relevant language experienced by people.

Word and passage meaning representations derived by LSA have been found capable of simulating a variety of human cognitive phenomena, ranging from developmental acquisition of recognition vocabulary to word-categorization, sentence-word semantic priming, discourse comprehension, and judgments of essay quality. In many applications LSA judgments of similarity agree well with human judgments (Landauer, Foltz, and Laham, 1998).

MISSION COMMUNICATIONS ANALYSIS

Introduction

The goal here is to develop and implement an LSA-based "Automated Communications Analysis" pipeline for performance assessment of mission communications applicable to both simulated and live Distributed Mission Training. The analysis of communications will be used to inform instructors and students for feedback both during mission performance and in related After Action Briefings. (Figure 1).

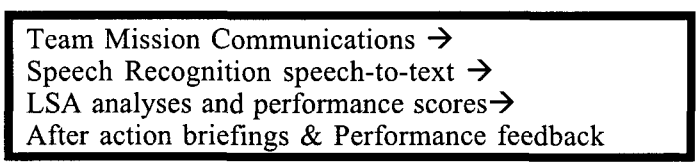

Figure 1. Automated Communications Analysis Pipeline

As a proof of concept, LSA was successfully able to predict team performance in a simulated UAV task environment (Kiekel, Cooke, Foltz, Gorman, and Martin, 2002) based only on communications transcripts. Using human transcriptions of 67 team missions in the UAV environment, LSA predicts objective team performance scores at a very high level of reliability (LSA alone, r = 0.74; LSA combined with additional text analysis measures, r = 0.85) The Team Performance Score used as the criterion measure is a composite of objective measures including the amount of fuel and film used, the number and type of photographic errors, route deviations, time spent in warning and alarm states, unvisited waypoints and violations in route rules. In this analysis, LSA compares the content of a mission transcript of unknown performance quality to those of known performance quality to generate the LSA Performance scores. A weighted average of the objective scores of the most semantically similar transcripts is calculated as the LSA score (this procedure is called the LSA *k-near* score). The strong performance of this automated technique, also validated by KAT in its Intelligent Essay Assessor software, suggests that it could be a very valuable tool for both summative assessment of performance and for feedback—similarity of a new transcript to known performance deficits could be used to provide the most applicable feedback to team members, individually or as a group.

The Speech Recognition problem

For use in the proposed Analysis Pipeline, either in near-real time or in an After Action Briefing, human typed transcription of the speech to text is not possible, therefore the speech-to-text transcription must be produced automatically. Output produced by commercial Speech Recognition (SR) systems is known to contain errors, even under the best of conditions. The question we want to answer is how robust is LSA in the presence of such noise? In particular how well does LSA correlate with human assessment of performance as errors are introduced into mission communications transcripts?

Synthesizing Noisy Data. Because transcripts produced by a SR system were not yet available to us, we evaluated the robustness of LSA using synthetic SR output. We developed a program to add noise to human-created transcripts of the UAV mission communications. Noise is defined by three types of errors:

- **Insertion**. Insert a word from an LSA space. Inserted words are limited to no more than m characters. In this study we used m=8.
- **Deletion**. Delete a word in the original transcript.
- **Substitution**. Substitute an original word with a word from an LSA space. Substituted words have two constraints. The first p characters must match the original word and the length must be within l characters of the original word. In this study we used p=2 and l=4.

In this study the LSA semantic space from which insertion and substitution words were selected was created from a corpus of mission communications transcripts, UAV training material, and transcripts of interviews with subject matter experts, which contained 6103 unique terms. Words were chosen randomly, subject to the constraints described above. The constrained space of terms mimics those vocabularies seen in military applications of SR systems. The error rates were systematically varied, with the ratio of the frequencies of Insertions, Deletions and Substitutions following the speech in noisy environments evaluations from Schmidt-Nielsen et al. (2001).

Each noise or degradation level was defined in terms of an overall per-word error rate and component insertion, deletion and substitution rates. In this study we used twelve different degradation levels. The first four levels represent "best" and "typical" error rates for two speech recognition algorithms, Linear Predictive Coding (LPC) and Mixed Excitation Linear Prediction (MELP). The remaining eight levels were created using the insertion, deletion and substitution rates of "typical LPC" and "typical MELP" and 57%, 71%, 85% and 99% for the overall error rates from Schmidt-Nielsen et al. (2001). Samples of the original text and the effects of degraded texts are shown in Figure 2. The percent of error rates for the twelve degradation levels are shown in Table 1.

Original	Sample 1	Sample 2
this is Intelligence to AVO	this is Intelligence to AVO	this is198 Intelligence to AVO
Intelligence this is the AVO	Intelligence this is thank AVO	Intelligence this the AVO
how many targets have you taken so far?	houses many targets have you tactful edges so?	how many tazsar have you taken so far?
we've taken two pictures we are on the third one.	maneuver we've taken twice pictures we are the third one.	wearing taken two we on thirty one.
thank you.	thank you.	bunched thank yoda.
I haven't taken the picture yet hold on.	I happen taken the pieces yet hold on.	I haven't taken the picture yeas hold on.
go ahead AVO	go ahead AVO	go ahead aviator
okay DEMPC my question is what is my effective time for change over to MSTE over	okay DEMPC my question is what is my effective time for change over to MSTE over	okay DEMPC myself question issues what is my effective time for checkpoint over
as soon as she take the picture you can switch over to MSTE	asks soon as she tasks the piece you can switch over to mst	kicks to MSTE over asap soon asked she take the
there's no effective on them.	thick effective systems on thirty.	picture you can over to MSTE
this is an effective radius of 5.0	is an effective racks of 55	there's no effective on them.
picture taken let's go	pitch taken	this is an radius of 5.0
let's change over we are a little off course but we'll get back on track.	let's change over we are secrets a little off course but we'll generic back radio on.	picture taken let's go let's change over we are lit wph off course but we'll generic back on 140

Figure 2: Sample SR Degraded Transcripts

Table 1. Degradation levels as defined by error rates (percent).

Degradation Level	Overall Error Rate	Insertion Rate	Deletion Rate	Substitution Rate
Best LPC	29	10	28	62
Best MELP	29	14	31	55
Typical LPC	44	11	18	71
Typical MELP	42	14	17	69
57% LPC	57	11	18	71
57% MELP	57	14	17	69
71% LPC	71	11	18	71
71% MELP	71	14	17	69
85% LPC	85	11	18	71
85% MELP	85	14	17	69
99% LPC	99	11	18	71
99% MELP	99	14	17	69

Performance Assessment of Synthetic SR missions

The evaluation corpus consists of 67 simulated mission communication transcripts, produced by human listeners. This evaluation corpus is termed the *verbatim* corpus and is assumed to have an error rate of 0%. The verbatim transcripts were evaluated by LSA to produce a set of *text* and *comparison* measures. Text measures are based on properties of each transcript. Comparison measures are obtained by comparing a transcript to its *k-nearest* neighbors in the LSA space. From these measures, two LSA scores were produced for each transcript. The LSA score is the single LSA *k-near* measure that has the highest correlation with human scores. The LSA+ score was produced using stepwise linear regression to build a model from the additional computational linguistic measures which measure syntactic and semantic properties of the transcripts. These were used to predict the team scores for each transcript. The reliability of Verbatim LSA+ with human scores is 0.85, while the reliability of Verbatim LSA is 0.74 (see Figure 3). (Note, these scores do not adjust for within subject variance due to multiple missions, which reduces results by about 10%).

For each of the twelve degradation levels, five samples of the corpus were generated using the program described earlier. Each sample was then evaluated by LSA to produce a set of text and comparison measures.

Stepwise regressions and correlations were performed to obtain LSA+ and LSA scores for each sample and to compute reliability with human scores. The reliability measures were averaged over the five samples for each of the twelve degradation levels. Reliability measures, along with standard error bars, are presented in Figure 3. The points connected by the two lines show reliability for LSA+ scores on LPC and MELP samples. The points connected by the bottom two lines show for LSA scores on LPC and MELP samples. Table 2 provides the average reliability scores for the models.

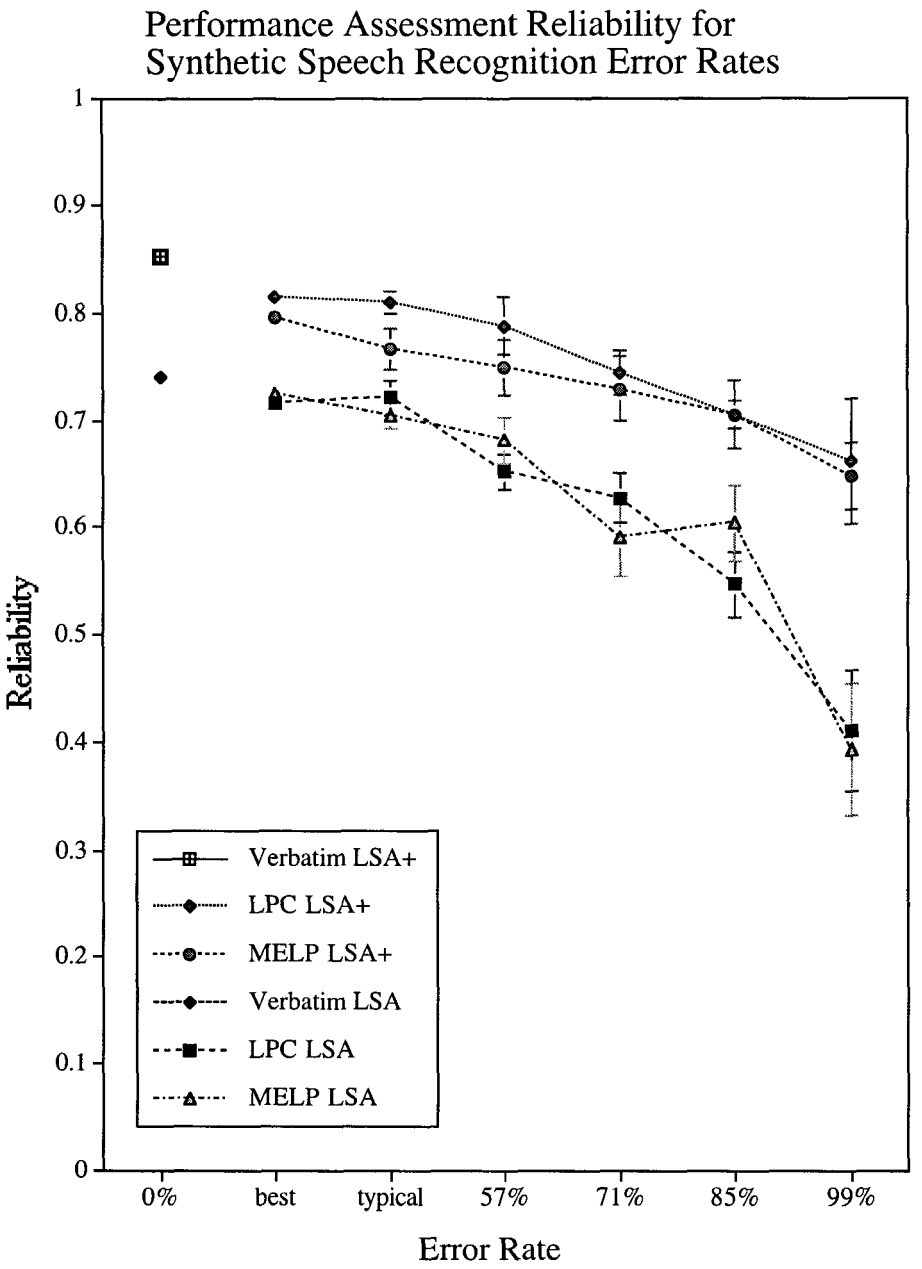

Figure 3. Performance of LSA using SR degraded transcript

Table 2. Reliability for models

		Best		Typ		57 %	
	Y	L	M	L	M	L	M
LSA	**.74**	.72	.73	.72	.71	.68	.68
LSA+	**.85**	.82	.80	.81	.77	.79	.75

		71 %		85 %		99 %	
	Y	L	M	L	M	L	M
LSA	**.74**	.64	.60	.55	.60	.41	.40
LSA+	**.85**	.74	.73	.71	.71	.66	.65

Y – Verbatim text (human transcriptions)
L – Average LPC SR text
M – Average MELP SR text

A second analysis was conducted which compared the original transcripts to their SR counterparts in LSA space. A Cosine Similarity judgment was made between each Original and its SR transcripts. A Cosine Similarity of 1.0 indicates perfect agreement. The average agreements decreased steadily as more and more noise was introduced, but as can be seen in Figure 4, both the best and typical error rates from commercial SR systems are judged very similar by LSA with scores above 0.90.

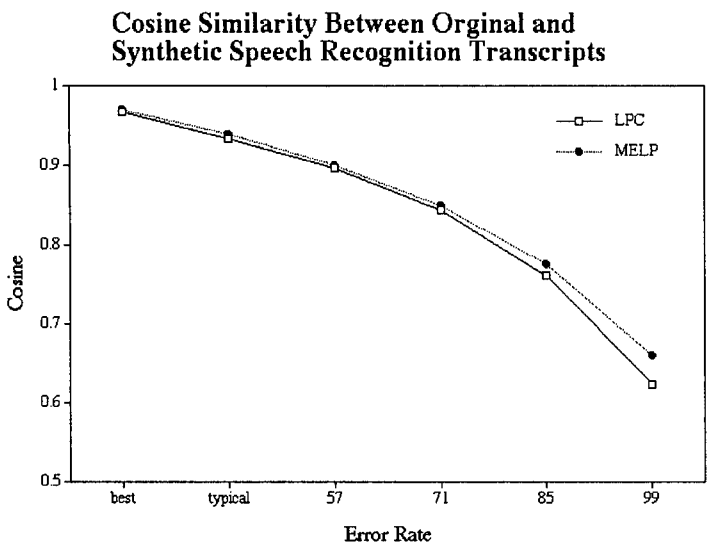

Figure 4 Average similarity judgments

CONCLUSIONS

Overall, the results of the study show that even with typical or worse speech recognition error rates, LSA is still able to accurately predict the performance of a team based on the transcript. With speech recognition error rates at 57%, LSA's predictive performance only degrades by about 10%. Thus, LSA appears to be highly robust to the typical types of errors that are encountered in ASR systems in noisy environments. An additional study is currently under the way using state of the art ASR technology to verify the error rates for the UAV corpus and evaluate performance under true ASR conditions.

The research suggests that LSA is an effective analysis tool even in conditions where the text to be analyzed has been significantly degraded. The noise introduced by SR systems is essentially random—enough of the original signal survives to be effectively analyzed—even at today's less than optimal SR error rates. The results suggest a range of potential applications. One application of this technology being explored with the Air Force Research Laboratory is tracking and scoring the tactical communications that occur between the members of a four-ship air combat flight and their weapons director to identify areas of training need and as an additional tool for assessing the efficacy of Distributed Mission Training (DMT) scenarios and missions. Similarly, we envision the combined technologies being useful in providing an embedded assistant to help track and evaluate incoming communication and to highlight or otherwise "flag" pertinent information and changes in content that may be of importance to operators and other personnel.

The capabilities suggested by these studies—to automatically and in real-time predict levels of team performance based on their communications and to identify and diagnose common error patterns should provide near future DMT systems with an enormous instructional advantage over current systems. These early success of Latent Semantic Analysis based tools are indicators of continuing improvement in simulator systems which will ultimately lead to better and more cost effective training.

REFERENCES

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

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing By Latent Semantic Analysis. Journal of the American Society For Information Science, 41(6), 391-407.

Kiekel, P. A., Cooke, N. J., Foltz, P. W., Gorman, J. & Martin, M. (2002). Some Promising Results of Communication-Based Automatic Measures of Team Cognition. Proceedings of Human Factors 2002.

Kiekel, P. A., Cooke, N. J., Foltz, P. W., & Shope, S. M. (2001). Automating measurement of team cognition through analysis of communication data. In M. J. Smith, G. Salvendy, D. Harris, and R. J. Koubek (Eds.), *Usability Evaluation and Interface Design,* pp. 1382-1386, Mahwah, NJ: Lawrence Erlbaum Associates.

Landauer, T. K. (1998). Learning and representing verbal meaning: The Latent Semantic Analysis theory. Current Directions in Psychological Science, 7, 161-164.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.

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

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.

ACKNOWLEDGEMENTS.

This work was supported by the Air Force Research Laboratory and by the Office of Naval Research. The authors are grateful to Nancy Cooke and Steven Shope for assistance in obtaining the transcripts and audio tapes. The authors may be contacted by email at pfoltz@k-a-t.com, dlaham@k-a-t.com, mad@k-a-t.com. or at Knowledge Analysis Technologies, 4940 Pearl East Circle, Suite 200, Boulder, CO, 80301