Trusted Communication: Utilizing Speech Communication to Enhance Human-Machine Teaming Success

E. L. Meszaros*
Brown University, Providence, Rhode Island, USA

Lisa Le Vie†
NASA Langley Research Center, Hampton, Virginia, USA

B. Danette Allen‡
NASA Langley Research Center, Hampton, Virginia, USA

An area of increasing interest for the next generation of aircraft is autonomy and the integration of increasingly autonomous systems into the national airspace. Such an integration requires humans to work closely with autonomous systems, forming teams. Our hypothesis is that a team composed of both humans and autonomous systems will operate better than either entity alone. We have existing procedures for certifying pilots to operate in the national airspace and are currently working on methods for validating the function of autonomous systems, however we have no method in place for assessing the interaction of these two disparate systems. Communication is one avenue. This paper will examine the use of language as a metric for ascertaining human-machine teaming effectiveness.

A proof-of-concept of the application of two communication-based analysis techniques, Linguistic Inquiry and Word Count (LIWC) and Latent Semantic Analysis (LSA), for the prediction of success in human/chatbot teaming was conducted. By running these analyses over data from the 2014 and 2015 Loebner Prize competitions of human/chatbot teaming, numerical scores were obtained that can be associated with scores provided by human judges during the competition. Correlating their LIWC and LSA data with the scores provided by the judges, and using linear regression over this correlation, formulae were obtained that predict the score of human/chatbot interaction. These formulae were tested over the 2013 Loebner Prize transcripts, determining that, though there was strong correlation between predicted and actual scores, the predictive success of this method was not strong. However, with specialized topic spaces and lexica, as well as larger data sets, the predictive power of these metrics will improve. Given the importance of providing metrics for human-machine system team success and given the promise shown by the communication-based LIWC and LSA methods, continuing research in this area is necessary.

After examining the potential for using communication and spoken language as a metric for the success of human/autonomous system teaming, this paper then examines aspects inherent to communication systems that may contribute to unreliability and reduced trust. Modern natural language processing tools rely on deep learning algorithms to create language rules that produce accurate results, but these rules are uninterpretable. The resulting blackbox system lacks transparency necessary for full validation and complete trust. Additionally, speech-based interfaces pose other difficulties to developing coordinated teamwork between humans and autonomous systems. Human communication is infrequently limited to speech only, instead usually relying on a combination of verbal, gestural, and general body language communication. Reducing an analysis of team effectiveness to a study of spoken language alone is problematic as it leaves these other equally important forms of communication out. This paper will examine these problems and the general deficiencies in speech-based metrics for human-machine teaming.

*Aerospace Research Intern, Autonomy Incubator
†Aerospace Research Engineer, Crew Systems Aviation and Operations, MS 152, AIAA Member
‡NASA Senior Technologist for Intelligent Flight Systems, MS 233, AIAA Senior Member.

I. Introduction

An area of increasing interest for the next generation of aircraft is autonomy and the integration of increasingly autonomous systems into the national airspace. Before such systems can be implemented, however, they must first be certified for use. Given the unique interaction between human operators and autonomous components, we must be able to certify not only the operation of the autonomous system and the human user individually, but also their operation as a team. This is especially important because the intention with the implementation of these autonomous systems is for the heterogeneous team to function better holistically than a homogenous one. Existing procedures for certifying pilots exist and methods for validating the function of autonomous systems are currently being researched; however, no methods are currently in place for assessing the interaction of these two disparate systems. Numerous reports and projects have identified this as a crucial area of research, but little headway has been made in providing metrics to enable the necessary certification.

It is therefore essential to establish an overview of available information on metrics for validation, verification, and certification of human/autonomous system teams. The focus of such an overview should be on examining what metrics have been tried, whether they have proven successful, what has prevented success, and any identified barriers to success. It is important to remember during such an investigation, however, that we cannot rely on only measurements of the individual components of the team if we want to be able to accurately predict success. No matter how complex the act of verifying dynamic and non-deterministic systems becomes, it is worthwhile and necessary to measure how well they function in situ, as a teammate. Assuring the efficacy of a human/autonomous system team increases the safety of the national air space through enhancement of autonomy usage and increased reliability in human/autonomy teams.

Additionally, significant focus is being given to certifying autonomous systems through ATTRACTOR (Autonomy Teaming and TRAjectories for Complex Trusted Operational Reliability). By establishing a basis for certifying trustworthiness and trust, ATTRACTOR seeks to facilitate multi-agent teaming between humans and autonomous/increasingly autonomous systems. Interaction with such autonomous systems must be intuitive to human operators and rely on multiple communication modes. To establish metrics for evaluating interactions, methods of understanding and quantifying aspects of successful teams must be explored. To this end, previous methods both proposed and used to evaluate the success of human/autonomous system teams are initially presented here.

II. Background

Previous literature on human/autonomous system teams and existing teaming metrics has identified areas critical to successful interaction, including trust, confidence, reliance, transparency, etiquette, and frustration. Additional research from the area of human/robot interaction as well as human/human teaming can be leveraged in this initial analysis as well. How these metrics are calculated and how they have been used previously provide information on whether they are applicable to human/autonomous system teams and whether they are useful for ATTRACTOR.

A. Human/Robot Interaction and Human Performance Modeling

Numerous studies into human/robot interaction have allowed for the development of a number of metrics to assess performance, including neglect tolerance, interface efficiency, fan out, robot attention demand, and human performance moderator functions. Neglect tolerance describes how well the robot members of a team can function without input from human members. This metric usually incorporates a measure of neglect time, or how long a robot can be neglected before its productivity or functionality drops below some usability criteria. It also usually incorporates a measure of neglect impact, or how the state of a robot may change without interaction from human team members. Neglect tolerance is solely a method of analyzing robot performance, and usually not in the context of any advanced or increasingly autonomous systems. Additionally, neglect of one robot is often the result of the human team member applying their attention to different robotic team member. The application of neglect tolerance to teams with multiple human members, able to attend to multiple robots at once, is not frequently acknowledged [1] [2] [3] [4] [5] [6] [7] [8].

Interface efficiency describes a measure of how long it takes to interface with a robot, how much effort is needed to interact with a robot, and how well a robot performs a particular task as a human operator interacts with it. Unlike neglect tolerance, metrics of interface efficiency measure some aspects of the robots performance but also aspects of how well the human performs using the robot's interface. Despite seeming to take into account aspects of human teammembers, however, interface efficiency metrics really measure intuitiveness and functionality of the interface rather than human performance. This metric too, then, does not fully encompass the entire team and cannot serve, at least on its own, as an accurate teaming metric [1] [2] [3] [4] [5] [6] [7] [8].

Robot Attention Demand (RAD) outlines the amount of human operator attention required as a function of interface

efficiency and neglect tolerance. This in turn leads to the "free time" concept, or how much time the human operator may spend on non-robot related tasks [6] [7] [9]. Fan-out provides a metric for determining how many robots can be effectively operated by one human user. Related to neglect tolerance, fan-out tends to increase as robotic team members become able to operate with less human intervention, intuitively because human operators can ignore them while providing instructions to other team members. This metric is also related to RAD, increasing as the demand of each robot decreases [5] [10] [11]. While these metrics have been centered on addressing robot performance, they are frequently driven by human input and therefore implicitly model human behavior during interaction with robots.

Some of these metrics may be valuable as-is for use in certifying the human component of a human/autonomous system team. For instance, situation awareness and workload measurements can provide information on human performance. More interesting, however, would be to refocus some of these other concepts on the human operator rather than the autonomous system. What does it mean for the robot to neglect the operator? Is it a measure of human situation awareness when a robot is not communicating? This definition of human neglect tolerance would certainly have an impact on situation awareness, which is an established factor in human performance. Quantitatively measuring a robot's neglect time and the human operator's neglect time may lead to the development of a holistic system neglect time that provides an overall metric for human/autonomous system performance.

Human Performance Moderator Functions (HPMFs) provide a more explicitly human-centric method for assessing team function. HPMFs are equations developed from a combination of subjective, self-reported measurements as well as objective, biophysiological measurements of human performance in human/robot teams. These equations are then used to predict how a human teammember's performance will be impacted by aspects such as workload, fatigue, temperature, and many others. Such metrics may lay the groundwork for providing evaluative metrics for human/autonomous system interaction by using already existing HPMFs to model the human component of human/robot teams. If HPMFs provide accurate modeling of human/robot teaming, then we already have at our disposal a method for quantitatively assessing this interaction. However the subjective nature of HPMFs, dependent as they are on self-reported characteristics, makes this somewhat suspect [12] [13] [14].

Quantitative evaluation and modeling tools for evaluating human performance have been in use for a while, however. While evaluation of the human component of a human/autonomous system team will probably be necessary, these performance models may prove useful as an evaluation tool for the entire team. For instance, evaluating a human operator's workload during a task and during the same task with help from an autonomous system can provide us with a metric for determining any workload benefits associated with the introduction of the autonomous system. In a high risk area, increases to workload may be a small price for the introduction of an autonomous system, while a low risk area may be less tolerant of workload increases.

B. Trust

Another focal point of previous research has been trust, focusing especially on the identified components of trust, including confidence, reliance and reliability, transparency, etiquette, and frustration. Because trust plays such a critical role in the efficacy of a team, it is important to establish trust with human/autonomy teams. This may mean both the human trusting the autonomous system as well as the autonomous system trusting the human. Some of the literature discusses ways for the autonomous system to self-evaluate confidence levels for information reported to the human, providing a metric for evaluating confidence which, in turn, impacts trust levels in a potentially-quantifiable way [15] [16]. Finding ways to establish and enhance trust levels, preventing both over-trust and under-reliance, produces more successful teams [17] [18] [19]. This may be an area with established metrics that can be leveraged for evaluation, but if not it may still be important for the impact that it has on other crucial areas of human/autonomy interaction (communication, transparency, teaming, etc.).

Transparency and Etiquette help describe ways in which the autonomous system can better communicate with the user to increase trust and, therefore, the success of the team. The more transparent the system is and the more the user understands decisions made, the more likely the user is to trust in the system. However, transparency comes at a cost – often presenting all of this extra information takes a toll visually on the user, and misinterpretation of the presented data can be as detrimental as not presenting any data at all. A system the reports information with etiquette by checking to see if a user is already responding to a situation before alerting the user, or by not interrupting to present new alerts, often results in smoother operations. Both of these have been measured qualitatively, but developing quantitative metrics may allow for better determination of the success of the team [20] [21] [14].

Frustration within a team can work to reduce trust and thereby undermine the overall success of the team. This aspect has been measured in a number of different ways – including quantitatively – and can provide information

on how effective the human/autonomy team is. Specifically, researchers have used physiological measurements to determine when it is most likely that a user will quit an interaction. If the autonomous system had such knowledge, they could personalize settings, information presentation, etc. to help reduce frustration and improve success. Additionally, researchers have investigated various methods for reducing frustration, by prevention, fixing the problem that lead to frustration, or reducing the effects. The implementation of the most effective method could lead to improved human/autonomy teaming, and the use of frustration metrics could allow us to measure this success [22] [23] [24] [25].

C. Communication

One final aspect of human/autonomous system teaming that could contribute to the development of teaming metrics is communication. Occurring in many different forms, from spoken to written to input methodologies, communication between team members can often signal the health of the team based on tone, word choice, or even style of communication [19] [26] [27] [28] [29] [30] [31] [32] [33].

Previous research based on emails used for team communication identified linguistic style matching to correlate to overall success in teams composed of all humans, as measured in the grades received by teams [27]. Such correlation has met middling success at predicting how well teams perform, with the best performance focused on the use of future-oriented words [27]. Additional research has relied on Latent Semantic Analysis (LSA) to evaluate team cognition as teams of all humans collaborate to complete tasks such as writing a paper or flying a plane [28].

Communication analysis suggests a unique method for evaluating team success since it is not limited to only one team member but inherently measures the interaction between multiple team members. Moreover, metrics based on communication may incorporate evaluation of previously discussed metrics; changes in aspects of trust, frustration, attention demand, etc. may be realized in changing communication patterns and styles. Communication measurements may also be made over written and spoken modalities, providing an easy transition from application to human/human teams to application to human/autonomous system teams.

D. Metric Discussion

Previous research has clearly identified aspects that may help identify good teams – high levels of trust, low frustration, and a shared workload all contribute to the identification of a good team. However, measuring such subjective metrics remains problematic. Subjective questionnaires, like the NASA TLX, have been used to allow subjects to self-select levels of trust [12] [13] [14], but holistic measures of all of these, especially holistic measures applied from outside the team rather than self-rated among team members, remain more difficult.

Many of the identified metrics focused on only one aspect of the team (i.e., either the human or the autonomy). Since the goal is for the overall team to be more successful than either of the individual components alone, it is therefore critical to develop metrics of overall team success rather than component success. While additional individual metrics may still be necessary (e.g., autonomy verification, pilot certification), identifying methods for holistic evaluation of the human/autonomous system team is of significant value. Initial review of the extant metrics identified one likely candidate for a likely holistic metric technique: Communication. This area in particular encompasses data about other aspects of teaming that have been identified as critical, such as trust, situation awareness, etc., without stopping to measure these values independently. For example, any changes in situation awareness may appear as changes in communication, while differing levels of trust may similarly appear in the language used to communicate with the autonomous system. By examining team communication, it is therefore not necessary to separately measure these individual aspects. Instead, their contribution the overarching team communication can be measured.

However, existing studies on team communication have focused on teams composed only of humans. As the primary language users of this planet, such a focus makes a great deal of sense. Is there anything preventing such metrics, however, from being applied to communication between humans and autonomous systems? As voice recognition and speech-based interfaces have improved and become increasingly ubiquitous, analyzing their communication not only becomes possible but may provide insight on team performance. Recent work has identified communication as a possible method for identifying levels of many of these aspects [34]. Examining the language used between members can provide additional information on how well the team is performing, and available linguistic analysis methods provide for quantitative measures of this success [35] [36]. These tools indicate that speech and communication analysis may allow for evaluating team success.

III. Proof of Concept

To identify whether communication between human and autonomous team members can serve as such a metric, a proof-of-concept of a language-based human/autonomy team analysis has been carried out. Specifically, this project has been designed to demonstrate the application of metrics that assess communication among human teams to situations where part of the team is a non-human autonomous system.

Two communication analysis tools described in the literature were evaluated: Latent Semantic Analysis (LSA) and Linguistic Inquiry and Word Count (LIWC) [36] [35]. Each of these methods analyzes textual communication data based on differing characteristics, producing numerical results that can be used for predicting the success of a team. LIWC is a computerized text analysis tool developed by James W. Pennebaker in 1991 to quantify the relation between natural language features and psychological states [37]. The 2015 version provides 94 different variables and a robust internal dictionary associating words in the English language with these 94 variables. For 89 of these variables, LIWC provides a percentage-of-total-wordcount number after analysis, meaning that the software counts the total number of words in the body of text, then counts the number that correspond to a given category, and provides that percentage. A number of 4.6 for the variable "Pronoun" would mean that 4.6% of the document was pronouns. In addition, LIWC also provides five "summary variables" that are based on algorithms proprietary to the software [38].

LSA is a technique for comparing the semantic space of documents, or even within documents. In LSA, a document is compared to other documents within a group based on term use frequency. Documents can then be weighted and compared numerically to produce quantitative assessments of semantic similarity. For this study, a pre-fabricated LSA tool provided by the University of Colorado – Boulder was used, including their existing semantic topic spaces. LSA then easily provides numbers for the similarity between differing documents as well as the sentences within each document [35].

A. Methodology

These two communication analyses were tested by applying them to transcripts of human/autonomous system teams in the Loebner Prize competition. The Loebner Prize is an annual competition for artificially intelligent chatbots, where prizes are awarded for the most human-like bot [39]. This is essentially a Turing test for chatbots, and the gold medal and highest monetary prize has been reserved for a chatbot that is able to fool human judges into thinking that it, too, is human. To date, this medal remains un-awarded. To test how human a chatbot is, each chatbot is asked 20 questions by a human judge; while these questions change from year to year, within a particular year each chatbot is asked the same set of questions. The answers given to these questions are then evaluated by human judges in order to obtain an overall numerical score. Bots with higher scores are better able to approximate humanity and fool judges, and were therefore more successful [40].

The Loebner Prize is useful for this initial proof-of-concept because it provides transcripts of the communication between the competing chatbots and a human judge, as well as records of the final scores assigned to each chatbot [41]. Both communication analysis tools can therefore be run over these transcripts in order to obtain numerical results. By plotting the numerical results of these analyses against the judge-provided score, formulae defining the relation between these two variables can be obtained using linear regression. In order to test how well these formulae work, how effective they are at predicting team success, they can be applied to the data from a previous Loebner Prize Competition. Running the same analysis techniques over transcripts from this competition, the formulae can then be applied to the provided numerical results to get our predicted score. These predicted scores can then be compared with the actual scores assigned by the judges in order to determine the efficacy of the developed scoring metric.

Transcripts of chatbot and human judge conversations from the 2014 and 2015 Loebner Prize competitions were edited to remove the labels for each statement. Where once the transcript might have read "Judge: My name is. . . ," it was made to read only "My name is. . . " This was done to better mimic verbal communication, where these identifying labels are not provided. For LIWC analysis, these modified transcripts were used without further changes. For LSA analysis, the judge's statements were separated from the chatbot's statements into two documents for comparison. For each chatbot, data were collected for all 94 LIWC variables, and for LSA the term and document one-to-many comparison were gathered, as well as the mean sentence coherency and standard deviation for sentence coherence, using two different pre-defined corpora. Each data point was then compared with the chatbot's original, judge-assigned score to see how well they correlated. For this initial analysis, our correlation threshold was set at Pearson correlation coefficient ($r > 0.4$ and Two-tailed T-distribution ($p < 0.05$.

Table 1 Predictive algorithms, r-values, and p-values based on LIWC and LSA categories

Category	Formula	r-value	p-value
Dictionary Words	(0.0307 * Dictionary Words Value) + 3.3985	-0.50	0.0026
Emotional Tone	(0.0058 * Emotional Tone Value) + 0.1633	0.57	0.00038
Total Pronouns	(-0.0294 * Total Pronouns Value) + 1.2532	0.43	0.0117
Auxiliary Verbs	(-0.0358 * Auxiliary Verbs Value) + 1.2	-0.48	0.0039
Prepositions	(0.0369 * Prepositions Value) + 0.3597	0.43	0.0117
Regular Verbs	(-0.0222 * Regular Verbs Value) + 1.1738	-0.48	0.0039
Positive Emotion	(0.0712 * Positive Emotion Value) + 0.3354	0.49	0.0036
Negations	(-0.0479 * Negation Value) + 0.7901	-0.59	0.0003
Insight	(-0.0407 * Insight Value) + 0.7868	-0.56	0.0006
Present Focus	(-0.0238 * Present Focus Value) + 1.0849	-0.49	0.003
Term Comparisons	(2.2315 * Term Comparison Value) – 1.4767	0.55	0.0009
Document Comparison	(0.6727 * Document Comparison Value) + 0.2456	0.45	0.008

IV. Results

Ten LIWC categories showed statistically significant correlation: Total Pronouns, Prepositions, Auxiliary Verbs, Regular Verbs, Positive Emotion, Present Focus, Dictionary Words, Insight, Emotional Tone, and Negations. Of these ten categories, eight are the straight forward percentage-of-total-wordcount variables, while Emotional Tone and Dictionary Words are proprietary summary algorithms native to LIWC. In addition, both LSA Term and Document comparison, when used with the "General_Reading_up_to_1st_ year_college" corpus, proved correlated with score (see table 1). For these twelve concepts, then, formulae were obtained using linear regression in order to predict the score based on a transcript's value for that variable.

To test the usability of the acquired formulae, LIWC and LSA were applied to transcripts from the 2013 Loebner Prize competition, processed in the same manner, to obtain values for the twelve correlated variables. These formulae were then used to predict the scores for these transcripts, and compared these predicted scores with the actual scores given each chatbot.

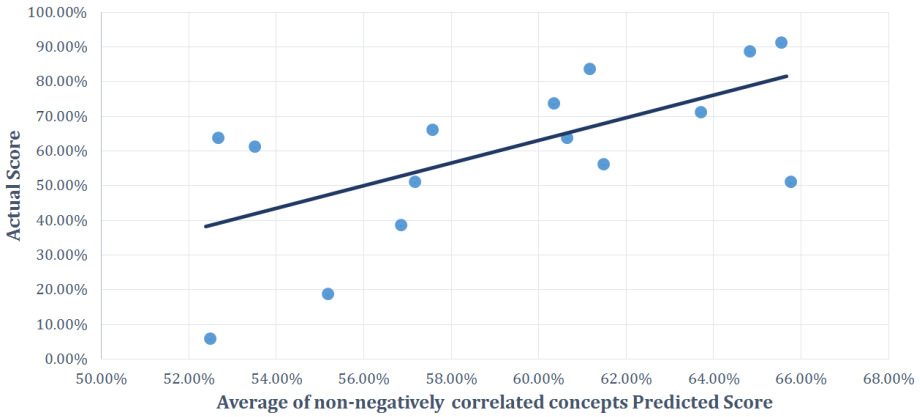

Fig. 1 Predicted score vs. actual score based on LIWC categories with positive correlation

The predicted scores for only eight of the ten LIWC categories were positively correlated. The highest correlation came from averaging the scores from the three of these variables that were positively correlated and statistically significant. An average of all LIWC categories with positive correlation predicted scores with an average difference of 16.97% from actual scores (see figure 2). Of these positively correlated categories, only three proved to be statistically significant: Pronoun, Insight, and FocusPres. Averaging these three categories provided less predictive power, with an average absolute difference of 17.20% between predicted and actual scores and an r of 0.69 and p of 0.004 (see figure 1).

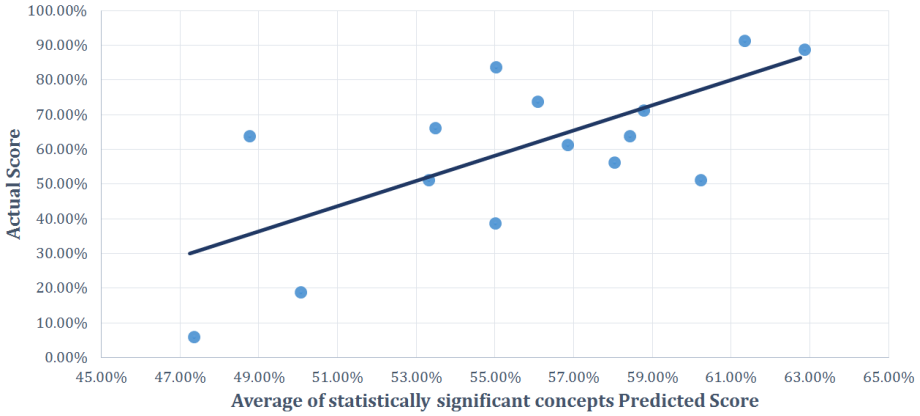

Fig. 2 Predicted score vs. actual score based on LIWC categories that were statistically significant

A smaller difference is attainable by relying only on the LIWC summary category "Dictionary Words," though this variable has a reduced correlation.

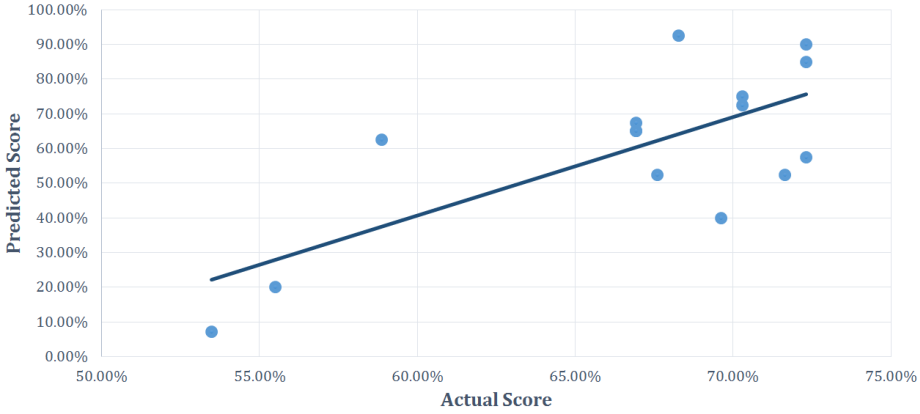

Fig. 3 Predicted score vs. actual score based on LSA document comparison

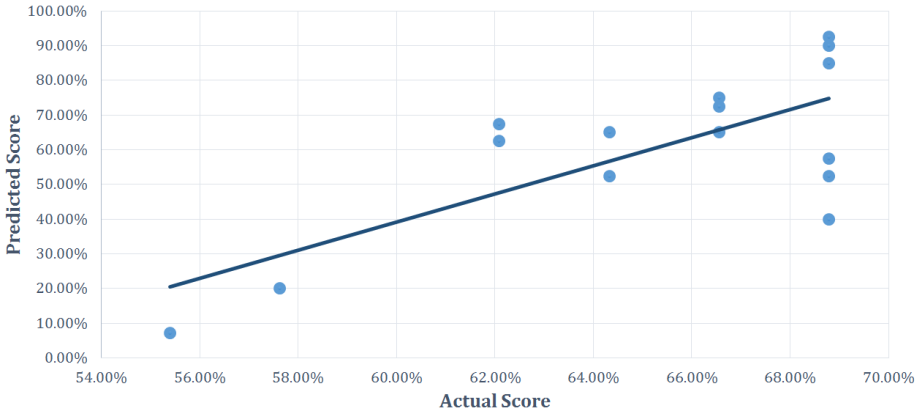

Fig. 4 Predicted score vs. actual score based on LSA term comparison

Both LSA categories provide high correlation between predicted and actual scores. Using document comparison,

LSA predicted scores that differed an average of 15.34% from the actual scores awarded to the chatbots, with an r of 0.72 and a p of 0.002 (see figure 3). Term comparison provided an average difference of 15.84% with an r of 0.73 and p of 0.002 (see figure 4). The most strongly correlated prediction method provided by LSA was based on term comparison with a correlation coefficient. Document comparison, however, provided the smallest average difference. Both term and document comparison proved more strongly correlated with smaller average differences than any LIWC category. Despite this, reliance on general purpose LSA algorithms and topic spaces produced middling predictive results.

V. Discussion and Limitations

Though there are some LIWC and LSA variables that are significantly correlated with score, their success at score prediction was not very strong. While this proof-of-concept work demonstrates the possibility of using communication as a metric of human/autonomy team success, the predictive capability must be improved before these metrics can be used. Additional and continuing research should focus on increasing this functionality in two key ways. First, the formulae we used to predict scores were generated from the 34 available transcripts from the 2014 and 2015 Loebner Prize competitions, with testing done over the 2013 competition. With more data points we can develop more accurate formulae, and even carry out multivariate analysis and generate better formulae from combinations of different variables that have better predictive power. Second, all of the formulae were developed with off-the-shelf algorithms and non-specialized dictionaries and lexica. By specializing these databases to the domain of research – either chatbot communication, for this example, or eventually to the national airspace – researchers can likely generate results that are far improved. Similarly, implementing our own LSA algorithms, rather than relying on pre-trained and non-specialized ones provided online, will likely improve our results as well.

As the chatbot case study demonstrates, evaluating the speech patterns between human and autonomous team members may provide a method for determining the efficacy of the team. However, certain aspects of speech-based communication and team behaviors could not be studied in this proof-of-concept analysis. Limitations of written language analysis and team size, as well as the benefits of non-speech data in understanding language are examined in this section.

The nature of chatbots is such that this study was necessarily conducted over written language. The most intuitive method of human communication, however, is spoken language, and most real, observable team communication will likely take this form [42] [43]. The LIWC and LSA tools used in this study require written transcripts, indicating that their application to analyzing spoken team communication requires an extra step: speech-to-text translation. Introducing a stage of translation in this analysis simultaneously introduces the potential for translation errors, leading subsequently to potential errors in team success analysis. The steady improvement of natural language processing tools suggests that such errors can be limited, but their likelihood and impact must first be understood [44].

Within each chatbot conversation, every line of dialogue from the human could be safely assumed to be addressed to the chatbot. Similarly, everything that the chatbot wrote was intended for the human. In two-member teams, the direction of the dialogue is easy to understand. This relationship is complicated in teams composed of more than two members, where directions and questions may be given toward any number of fellow teammates. There is some evidence that the larger the team, the better the performance, due to concurrent work, the ability to handle more tasks and more complex tasks, and the tendency of humans to perform even practiced tasks better when in the presence of their peers [45]. The skills necessary to navigate and interpret multi-person conversations could severely impact the conversation patterns of human/autonomous system teams and make speech monitoring more difficult or potentially less successful. The sterile enforced two-member environment of this proof-of-concept chatbot study provides no insight on larger teams.

Additionally, speech-based systems have some inherent limitations, most often understood in the context of speech processing. Suprasegmental information, including intonation, stress, and even gestural data, is left out of speech analysis, though often this information includes valuable data on the conveyed content and certainly on team dynamics [44]. Additionally, recent studies have suggested that multimodal systems that account for gestural suprasegmental data are more intuitive for users [46] [47]. Extending evaluation beyond just speech data to include additional information on team gesture, tone, and intonation may lead to better determination of team success. Such measures may be limited, however, as autonomous systems rarely have the ability to respond with gestures, intonation, or changing stress patterns.

Natural language processing tools used in communication evaluation introduce some further difficulties into the analysis. Advanced language tools are often black boxes that produce good results without providing complete understanding of how those results were arrived at [44]. Without offering a way to understand how conclusions about communication success are drawn, trust in these language analysis tools is inherently reduced. If communication

evaluation is proposed as a way to evaluate team success in order to establish levels of trust and trustworthiness in human/autonomous system teams, there must simultaneously be trust in the methods used for evaluation.

VI. Conclusion

Despite current limitations, the proof-of-concept described here seems to demonstrate that communication may be a helpful metric in evaluating human/autonomous system teams. Though limited only to analysis of text, evaluation of team communication provided methods to predict the overall success. If expanded to incorporate verbal communication, multimember teams, and suprasegmental information, this suggests the potential for a powerful method of evaluating teams.

Test and evaluation needs extend beyond the non-trivial issue of dealing with non-deterministic systems and dynamic environments. In addition to evaluating the autonomous system itself, considering its function as part of a team in cooperation with humans is critical. Evaluation metrics for determining if increasingly autonomous systems add value to a given task are likely to inform the adoption and implementation of these systems. Moreover these evaluation procedures may lead to new training techniques for human operators and collaborators. Because the goal of using increasingly autonomous system and human teams is to produce results better than homogenous teams, the next generation of metrics must account for team success.

As ATTRACTOR works to establish a baseline for certifying trust and trustworthiness in such teams, a focus on communication between these team members may provide necessary and valuable information. Speech analysis has been identified as an area of interest for evaluating heterogeneous human/autonomous system team success due in no small part to the intuitive nature of speech-based communication in humans. This study suggests that use of Latent Semantic Analysis and the Linguistic Inquiry and Word Count tools may provide a method for assessing the success of team speech patterns to determine their efficacy, providing the necessary tools to aid in certifying safety within the national airspace. While current natural language processing tools, and therefore speech analysis techniques, bear some significant difficulties and express some inherent limitations, the intuitiveness and ubiquity of speech identifies it as still worthy of consideration.

Acknowledgments

The authors would like to thank the Autonomy Incubator and Crew Systems and Aviation Operations Branch at NASA Langley for support in carrying out this research. Additional guidance came from Dr. Karin Knorr Cetina at the University of Chicago.

References

[1] Elara, M. R., Calderon, C. A. A., Zhou, C., and Wijesoma, W. S., "False alarm demand: A new metric for measuring robot performance in human robot teams," *Autonomous Robots and Agents, 2009. ICARA 2009. 4th International Conference on*, IEEE, 2009, pp. 436–441.

[2] Crandall, J. W., and Goodrich, M. A., "Measuring the intelligence of a robot and its interface," *Proc. of PERMIS*, 2003.

[3] Goodrich, M. A., and Olsen, D. R., "Seven principles of efficient human robot interaction," *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, Vol. 4, IEEE, 2003, pp. 3942–3948.

[4] Wang, J., Wang, H., Lewis, M., Scerri, P., Velagapudi, P., and Sycara, K., "Experiments in coordination demand for multirobot systems," *Proceedings of IEEE international conference on distributed human-machine systems*, 2008.

[5] Olsen, D. R., and Goodrich, M. A., "Metrics for evaluating human-robot interactions," *Proceedings of PERMIS*, Vol. 4, 2003.

[6] Crandall, J. W., Cummings, M. L., Della Penna, M., and de Jong, P. M., "Computing the effects of operator attention allocation in human control of multiple robots," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, Vol. 41, No. 3, 2011, pp. 385–397.

[7] Crandall, J. W., and Cummings, M. L., "Identifying predictive metrics for supervisory control of multiple robots," *IEEE Transactions on Robotics*, Vol. 23, No. 5, 2007, pp. 942–951.

[8] Crandall, J. W., Cummings, M. L., and Nehme, C. E., "Predictive model for human-unmanned vehicle systems," *Journal of Aerospace Computing, Information, and Communication*, Vol. 6, No. 11, 2009, pp. 585–603.

[9] Savla, K., Nehme, C., Temple, T., and Frazzoli, E., "On efficient cooperative strategies between UAVs and humans in a dynamic environment," *AIAA Guidance, Navigation and Control Conference and Exhibit*, 2008, p. 6841.

[10] Goodrich, M. A., and Schultz, A. C., "Human-robot interaction: a survey," *Foundations and trends in human-computer interaction*, Vol. 1, No. 3, 2007, pp. 203–275.

[11] Crum, V., Homan, D., and Bortner, R., "Certification challenges for autonomous flight control systems," *AIAA Guidance, Navigation, and Control Conference and Exhibit*, 2004, p. 5257.

[12] Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., and Goodrich, M., "Common metrics for human-robot interaction," *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, ACM, 2006, pp. 33–40.

[13] Harriott, C. E., Zhang, T., and Adams, J. A., "Evaluating the applicability of current models of workload to peer-based human-robot teams," *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*, IEEE, 2011, pp. 45–52.

[14] Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., and Barnes, M., "Situation awareness-based agent transparency," Tech. rep., Army Research Lab Aberdeen Proving Ground, 2014.

[15] Hunt, S., Martin, L., and Mercer, J., "Adapting a Human–Automation Trust Scale to an Air Traffic Management Environment," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 58, SAGE Publications Sage CA: Los Angeles, CA, 2014, pp. 26–30.

[16] Hutchins, A. R., Cummings, M., Draper, M., and Hughes, T., "Representing Autonomous Systems' Self-Confidence through Competency Boundaries," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 59, SAGE Publications Sage CA: Los Angeles, CA, 2015, pp. 279–283.

[17] Lee, J. D., and See, K. A., "Trust in automation: Designing for appropriate reliance," *Human factors*, Vol. 46, No. 1, 2004, pp. 50–80.

[18] MacLeod, I., and Hendford, W., "Certification of the EC/Aircrew Team-a Cognitive Control Loop," *The Human-Electronic Crew: The Right Stuff*, Vol. 26, 1997, pp. 155–162.

[19] Parasuraman, R., and Miller, C. A., "Trust and etiquette in high-criticality automated systems," *Communications of the ACM*, Vol. 47, No. 4, 2004, pp. 51–55.

[20] Lyons, J. B., "Being transparent about transparency," *AAAI Spring Symposium*, 2013.

[21] Kim, T., and Hinds, P., "Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction," *Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on*, IEEE, 2006, pp. 80–85.

[22] Klein, J., Moon, Y., and Picard, R. W., "This computer responds to user frustration: Theory, design, and results," *Interacting with computers*, Vol. 14, No. 2, 2002, pp. 119–140.

[23] Chang, S. H.-H., "HCI Seminar Final Report: User Frustration," 2007.

[24] Mower, E., Feil-Seifer, D. J., Mataric, M. J., and Narayanan, S., "Investigating implicit cues for user state estimation in human-robot interaction using physiological measurements," *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on*, IEEE, 2007, pp. 1125–1130.

[25] Scholtz, J., and Bahrami, S., "Human-robot interaction: development of an evaluation methodology for the bystander role of interaction," *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, Vol. 4, IEEE, 2003, pp. 3212–3217.

[26] Miller, C. A., "Trust in adaptive automation: the role of etiquette in tuning trust via analogic and affective methods," *Proceedings of the 1st international conference on augmented cognition*, Citeseer, 2005, pp. 22–27.

[27] Munson, S. A., Kervin, K., and Robert Jr, L. P., "Monitoring email to indicate project team performance and mutual attraction," *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, ACM, 2014, pp. 542–549.

[28] Kiekel, P. A., Cooke, N. J., Foltz, P. W., and Shope, S. M., "Automating measurement of team cognition through analysis of communication data," *Usability evaluation and interface design*, 2001, pp. 1382–1386.

[29] Kiekel, P. A., Cooke, N. J., Foltz, P. W., Gorman, J. C., and Martin, M. J., "Some promising results of communication-based automatic measures of team cognition," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 46, SAGE Publications Sage CA: Los Angeles, CA, 2002, pp. 298–302.

[30] Foltz, P., LaVoie, N., Oberbreckling, R., Chatham, R., and Psotka, J., "DARCAAT: DARPA competence assessment and alarms for teams," *Proceedings of the 2008 Interservice/Industry Training, Simulation & Education Conference*, 2008.

[31] Fischer, U., McDonnell, L., and Orasanu, J., "Linguistic correlates of team performance: Toward a tool for monitoring team functioning during space missions," *Aviation, space, and environmental medicine*, Vol. 78, No. 5, 2007, pp. B86–B95.

[32] Fischer, U., Mosier, K., Orasanu, J., Fischer, U., Morrow, D., Miller, C., Mosier, K., Veinott, B., and Orasanu, J., "Exploring communication in remote teams: Issues and methods," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 57, SAGE Publications Sage CA: Los Angeles, CA, 2013, pp. 309–313.

[33] Tausczik, Y. R., and Pennebaker, J. W., "Improving teamwork using real-time language feedback," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2013, pp. 459–468.

[34] Meszaros, E. L., "Improving Human/Autonomous System Teaming Through Linguistic Analysis," *Midwest Computational Linguistics Colloquium*, 2016.

[35] Laham, D., and Steinhart, D., "Latent Semantic Analysis at CU Boulder," `http://lsa.colorado.edu/`, 1998.

[36] Pennebaker, J., Booth, R., Boyd, R., and Francis, M., "Linguistic Inquiry and Word Count," `www.LIWC.net`, 2015.

[37] Short, J. C., McKenny, A. F., and Reid, S. W., "More Than Words? Computer-Aided Text Analysis in Organizational Behavior and Psychology Research," *Annual Review of Organizational Psychology and Organizational Behavior*, Vol. 5, No. 1, 2018.

[38] Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K., "The development and psychometric properties of LIWC2015," Tech. rep., 2015.

[39] Bradeško, L., and Mladenić, D., "A survey of chatbot systems through a loebner prize competition," *Proceedings of Slovenian Language Technologies Society Eighth Conference of Language Technologies*, 2012, pp. 34–37.

[40] Floridi, L., Taddeo, M., and Turilli, M., "Turing's imitation game: still an impossible challenge for all machines and some judges—-an evaluation of the 2008 Loebner contest," *Minds and Machines*, Vol. 19, No. 1, 2009, pp. 145–150.

[41] al Rifaie, M. M., "Loebner Prize at the Society for the Study of Artificial Intelligence and Simulation of Behavior," `http://aisb.org.uk/events/loebner-prize`, 2018.

[42] Novoa, J., Wuth, J., Escudero, J. P., Fredes, J., Mahu, R., and Yoma, N. B., "DNN-HMM based Automatic Speech Recognition for HRI Scenarios," *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ACM, 2018, pp. 150–159.

[43] Meszaros, E. L., Chandarana, M., Trujillo, A., and Allen, B. D., "Speech-based natural language interface for UAV trajectory generation," *Unmanned Aircraft Systems (ICUAS), 2017 International Conference on*, IEEE, 2017, pp. 46–55.

[44] Meszaros, E. L., Chandarana, M., Trujillo, A., and Allen, B. D., "Compensating for Limitations in Speech-Based Natural Language Processing with Multimodal Interfaces in UAV Operation," *International Conference on Applied Human Factors and Ergonomics*, Springer, 2017, pp. 183–194.

[45] McComb, C., Cagan, J., and Kotovsky, K., "Optimizing design teams based on problem properties: computational team simulations and an applied empirical test," *Journal of Mechanical Design*, Vol. 139, No. 4, 2017, p. 041101.

[46] Chandarana, M., Meszaros, E. L., Trujillo, A., and Allen, B. D., "'Fly Like This': Natural Language Interface for UAV Mission Planning," 2017.

[47] Chandarana, M., Meszaros, E. L., Trujillo, A., and Danette Allen, B., "Natural Language Based Multimodal Interface for UAV Mission Planning," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 61, SAGE Publications Sage CA: Los Angeles, CA, 2017, pp. 68–72.