Estimating Trust in Conversational Agent with Lexical and Acoustic Features

Mengyao Li[1], Isabel M Erickson[1], Ernest V. Cross[2], John D. Lee[1]
[1]Department of Industrial and Systems Engineering,
University of Wisconsin – Madison, Madison, Wisconsin, USA
[2]TRACLabs, Webster, Texas, USA

As NASA moves to long-duration space exploration operations, there is an increasing need for human-agent cooperation that requires real-time trust estimation by virtual agents. Our objective was to estimate trust using conversational data, including lexical and acoustic features, with machine learning. A 2 (reliability) × 2 (cycles) × 3 (events) within-subject study was designed to provoke various levels of trust. Participants had trust-related conversations with a conversational agent at the end of each event. To estimate trust, subjective trust ratings were predicted using machine learning models trained on three types of conversational features (i.e., lexical, acoustic, and combined). Results showed that a random forest model, trained on the combined lexical and acoustic features, best predicts trust in the conversational agent ($R^2_{adj} = 0.67$). Comparing models, we showed that trust is not only reflected in lexical cues but also acoustic cues. These results show the possibility of using conversational data to measure trust unobtrusively and dynamically.

INTRODUCTION

Measuring trust obtrusively and dynamically has become increasingly essential for human-AI (Artificial Intelligence) teaming in various domains. Particularly for space missions, as the National Aeronautics and Space Administration (NASA) moves to long-duration missions, longer time delays of communication between crews and ground control will require more interdependent cooperation between the crews and the onboard conversational agent. Trust affects human-agent cooperation, especially under uncertain and risky situations such as space missions (Chiou & Lee, 2016, 2021; Endsley et al., 2021; Johnson & Vera, 2019). While self-reported trust is often used as a gold standard for measuring trust, it is unable to satisfy the need for unobtrusively monitoring trust dynamics, especially for space missions where the human-AI team must work independently from a support structure (Li, Alsaid, Noejovich, Cross, & Lee, 2020; Yang, Christopher, & Christine, 2021). While several attempts have been developed to measure trust using behavioral and physiological data (Hu, Akash, Jain, & Reid, 2016; Kohn, de Visser, Wiese, Lee, & Shaw, 2021), we propose that measuring trust using conversational data is a promising yet under-explored approach. To measure trust, trust indicators in conversations need to be identified. In this paper, we identified trust indicators from conversation using the machine learning approach.

Conversational Indicators of Trust

Humans naturally express and recognize emotions via voices, facial expressions, and gestures. Indeed, as can be seen in the well-known adage: “It’s not what you say, it’s how you say it.” Therefore, for the same set of words, people with different tones or volumes in their voices express different messages. If we only consider word choice, we may miss important aspects of an utterance and may even completely misunderstand the meaning of the message (Sebe, Cohen, & Huang, 2005). Thus, it is important to not only consider lexical cues (e.g., words used in conversations) but also consider acoustic or paralinguistic cues (e.g., pitch, formants) (Elkins & Derrick, 2013). Acoustic cues capture speech signals such as pitch that can be extracted and interpreted to provide information about the speaker, or about the underlying message of the words. We propose that trust can be measured from conversations using lexical and acoustic cues.

Understanding these cues in communication is essential to measuring and managing human trust to promote cooperation with virtual agents. There is increasing effort to use conversational cues to monitor trust dynamics. Elkins and Derrick (2013) showed that vocal pitch is inversely related to trust, especially during the early stages of interactions. Waber and colleagues showed a significant correlation between non-linguistic features of speech and initial trust in technical communication in hospital settings (Waber, Williams, Carroll, & Pentland, 2015). Gauder and colleagues showed that acoustic features in speech can classify the trust state (i.e., trust or distrust) with up to 76% accuracy (Gauder et al., 2021). However, there is limited research unifying lexical and acoustic features in conversation to predict trust. The combination of lexical and acoustic cues can provide more a complete set of conversational indicators of trust. Additionally, most research considers trust as a binary trust state, rather than a continuous variable (Chen, Levitan, Levine, Mandic, & Hirschberg, 2020; Waber et al., 2015). Our study used a combination of lexical and acoustic cues to estimate the continuous level of trust in a conversational agent.

Machine Learning for Trust Estimation

Successfully estimating trust using Machine Learning (ML) requires crafting a situation that produces variations in trust and generates repeated measures of trust. First, variation in a ground truth trust measure need to be established by inducing changes in trust. We used the well-validated variable, reliability, which has shown a strong causal relationship with

trust, to induce changes in trust (Desai et al., 2012; Lee & See, 2004; Li, Holthausen, Stuck, & Walker, 2019). For example, a high-reliability scenario, meaning the virtual agent performs well, will lead to higher levels of trust. Second, a well-labeled target response (i.e., trust) is needed for supervised ML models predicting a continuous outcome (i.e., regression models). Thus, we collected subjective trust ratings along with conversational data. Third, trust changes as a dynamic process that varies across the interactions. A single point estimate of trust is insufficient (Yang et al., 2021). In our experiment, we designed multiple check-in points after every interaction with the automated system to capture multiple measures of trust and multiple samples of conversational data.

METHOD

The study was a 2 (reliability) × 2 (cycles) × 3 (events) within-subject study (see Figure 1). Participants performed 12 decision-making tasks associated with managing a system of a simulated space station: the Habitat's Carbon Dioxide Removal System (CDRS). Participants were assisted by a conversational agent with 2 levels of reliability (i.e., high, and low). Each level of reliability had 2 repeated cycles of the CDRS tasks, each including 3 events (i.e., startup, venting, shutdown). To induce substantial changes in trust, the high-reliability conversational agent provided 100% correct recommendations whereas the low-reliability agent provided 20% correct recommendations. The 12 total events were designed to elicit various levels of trust through manipulation of the agent's reliability. At the end of each event, the agent initiated a conversation by asking six trust-related questions (Li et al., 2020). The questions asked about participants experiences and feelings in the last interaction centered around conversational agent's performance. For example, "How would you describe your experience selecting the procedure?". Once the participant finished the conversation, they then rated their trust in the agent (Jian, Bisantz, & Drury, 2000). In total, each participant had the opportunity for at least 72 conversational turns with the agent.

Participants

A total of 24 participants (18 female, 6 male) were recruited from the Madison, WI area (M = 23.7, SD = 3.6). Recruitment required participants to be comfortable using a computer and a touch screen interface as well as have some technical background (e.g., completion of engineering or science courses). Due to the safety concerns of COVID-19, the study took place online. It was a two-session, two-day study with each session lasting up to two hours. In total, the study lasted approximately four hours for each participant. Participants received $30 per hour for up to $120 for four hours of participation.

Apparatus

The experimental task used the Procedure Integrated Development Environment (PRIDE), which is an electronic procedure software, to maintain the simulated International Space Station habitat (Izygon, Kortenkamp, & Molin, 2008; Schreckenghost, Milam, & Billman, 2014). Participants were tasked with removing carbon dioxide using the Carbon Dioxide Removal System (CDRS). PRIDE had pre-defined procedures for activating or deactivating the system, each with a unique order of steps. The procedures were complex, involving sub-procedures and diagrams. PRIDE also automated certain tasks after confirmation from the participant.

A conversational agent, named Bucky, was programmed with procedure protocols to provide recommendations and overall guidance to help participants maintain the habitat task in PRIDE. Google Dialogflow, a Natural Language Understanding (NLU) platform was used to design and integrate the user interface. Participants interacted with Bucky throughout the experiment via a webpage. Participants were asked to speak to the conversational agent using their microphone while answering the questions for conversational measures of trust. Participants also had the option of using keyboard and response buttons. Both audio and speech-to-text data were collected as conversational measures.

Procedures

After signing the consent form, participants completed a two-part training: the first part provided a study overview, trained the participants on how to use PRIDE, and trained them on the CDRS system, while the second part included an interactive demonstration of working with Bucky on decision making in PRIDE. During the study, participants had 25-minutes to use the CDRS system to remove CO_2 from the Habitat's environment by running the CDRS through three events (startup, venting, and shutdown) before their crew experienced CO_2 poisoning. For each event, the participant made two essential decisions with Bucky's aid. The first was selecting a procedure to run to remove the CO_2. Bucky recommended a procedure. The participant could either accept Bucky's recommendation or reject it and choose a different procedure.

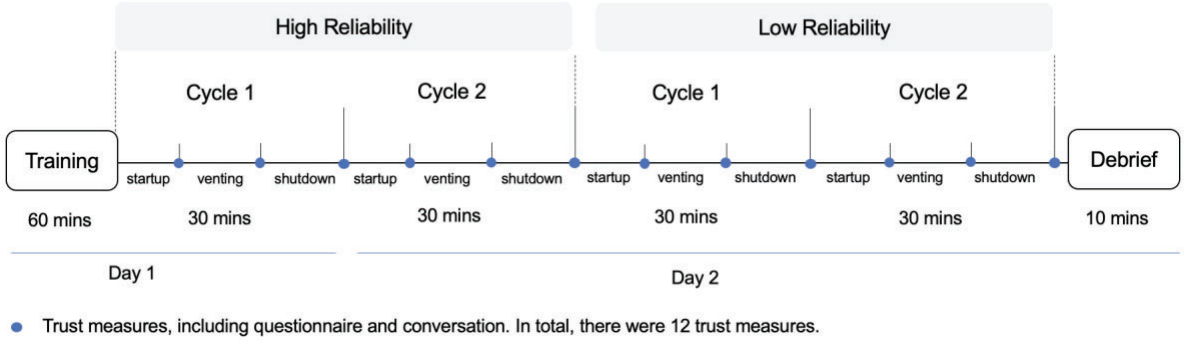

Figure 1. Study design with 12 points of trust measurement.

The second decision was deciding whether to rerun the procedure selected. As part of this decision participants would be advised by Bucky if the state of the CDRS was incorrect and if a different procedure should be run. The participants could either accept Bucky's recommendation or reject Bucky's recommendation and run a different procedure. The participants made their decisions either based on their knowledge from their training session and by relying on Bucky's recommendation. Once the procedure was selected, PRIDE automated the procedure execution. While the procedure was running, participants engaged in a secondary task where they checked and reported the statuses of CDRS components to Bucky. If the participant selected the incorrect procedure, an error occurred. The participant then had to manually stop the procedure and reselect a procedure. The participant finished the event by confirming the procedure ran correctly, then Bucky administered six trust-related conversational questions. For the administered questions, there were some variations in question framing to avoid appearing repetitive. After the conversational questions, participants completed the trust ratings. The total time for each cycle, including conversations with Bucky and the survey, was approximately 30 minutes (see Figure 1). At the end of the study, participants were debriefed and compensated.

RESULTS

The audio data contained 1806 short audio clips with a mean length of 8.17s (SD = 10.88). The text data contained 810 lines of utterances, with a mean length of 38.25 characters (SD = 26.49). Note that the text data only included utterances with elements of sentiment. The combined dataset was inner joined by matching the common audio identifiers, leaving the final dataset with 810 lines. The mean trust score for the high-reliability condition was 5.78 (SD = 0.86) whereas, for the low condition, the mean trust score was 4.37 (SD = 1.44). The overall trust score was 5.15 (SD = 1.35).

Machine learning Pipeline

We adapted a previously described machine learning pipeline (McDonald, Ade, Peres, Ferris, & Wiener, 2020) for trust estimation, which is shown in Figure 2. The conversations were first separated into audio, text, and combined data analysis streams. The audio and text features were extracted using speech signal processing and text analysis respectively. Then, the features were reduced and standardized. The processed features were then used to fit the machine learning models. We considered three main types of machine learning models (i.e., gradient descent-based, distance-based, and tree-based). The best-performing model was selected based on the evaluation metrics: root mean squared error (RMSE) and adjusted R-squared ($R^2{}_{\text{adj}}$). RMSE indicates the absolute fit of the model in the units of the response variable and R^2_{adj} indicates the variance in the response variable that can be explained by the predictor variables with a penalizing factor for adding independent variables. The dataset was processed and analyzed in R (R Development Core Team, 2011).

Feature Engineering

A total of 22 features were identified and extracted, including 13 for audio and 9 for text. Using the Boruta algorithm and the variance inflation factor (VIF) score for multicollinearity, 3 features were removed since the VIF scores were higher than 10 indicating a strong multicollinearity. A z-score standardization was conducted on all features sets and the response variable. The 19 reduced features are summarized in Table 1.

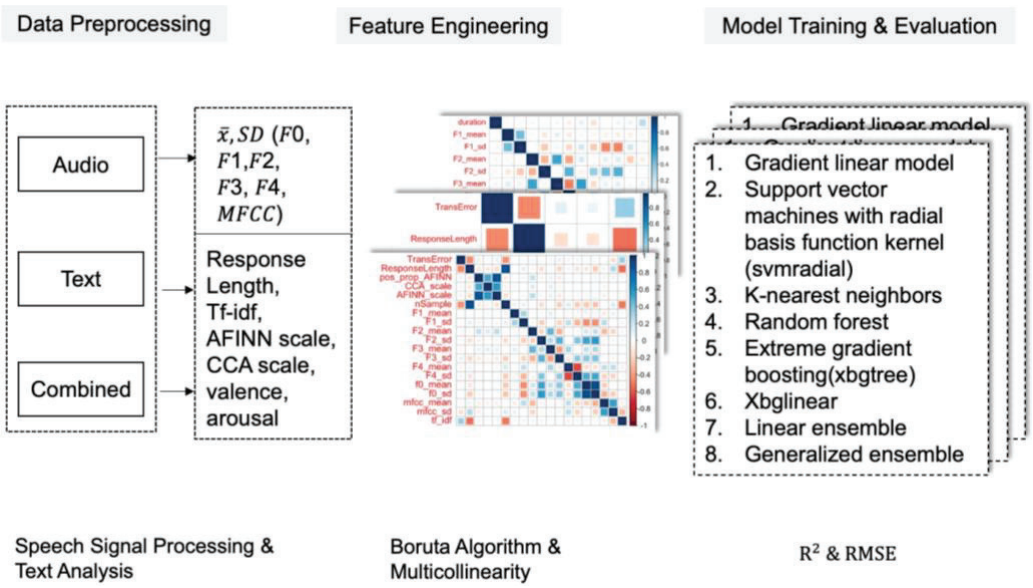

Figure 2. Machine learning pipeline to predict trust.

Trust Estimation

Table 2 shows the machine learning model performance across text, audio, and the combined features. Using only audio features, k-nearest neighbors (KNN) outperformed other models in terms of R^2_{adj} and RMSE values. For text-only features and the combined text and audio feature sets, both metrics agreed that random forest outperformed other models by having the lowest RMSE and the highest R^2. Trust score prediction had an RMSE score of 0.59 which represents the difference between the predicted and the actual trust score. Compared to the linear baseline model, the best-performing model's R^2 improved from 0.34 to 0.67. The result is notable because human behaviors, especially trust, are inherently difficult to predict.

Table 1. Reduced 19 Features for Model Training.

Category	Feature	Description
Audio	nSample	Total number of records/samples in the sound.
	$\bar{x}, SD$ ($F0$)	Mean and standard deviation of fundamental frequency (F_0).
	$\bar{x}, SD$ ($F1$)	Mean and standard deviation of the first formant in vowels (F_1), which is inversely related to vowel height. The higher the F_1, the lower the vowel height.
	$\bar{x}, SD$ ($F2$)	Mean and standard deviation of the second formant in vowels (F_2), which is related to the degree of backness. The higher the F_2, the more front the vowel.
	$\bar{x}, SD$ ($F3$)	Mean and standard deviation of the third formant in vowels(F_3), which is related to the degree of roundness. The lower the F_3, the rounder shape of the lip.
	$\bar{x}, SD$ ($F4$)	Mean and standard deviation of the fourth formant in vowels (F_4), which is related to the degree of resonance/larynx. The higher the F_4, the higher the larynx.
	$\bar{x}, SD$ ($MFCC$)	Mean and standard deviation of Mel-frequency cepstral coefficients.
Text	Response length	Number of words in text response before any text cleaning (e.g., removing stop words, tokenization, stemming, etc.).
	TF-IDF	Term Frequency-Inverse Document Frequency evaluates how relevant a word is to a document in a collection of documents.
	AFINN	The overall sentiment of the utterance using AFINN lexicon(Nielsen, 2011), divided by the square root of total terms with the sentiment, was scaled to -5 to 5.
	Positive AFINN	The proportion of positive sentiment, divided by the square root of total terms and the overall AFINN score.
	Context sentiment	Sentiment score considering the context for the utterance (window size of 4 words before and 2 words after), scaled to -5 to 5.
	Translation error	A binary indication on whether Bucky made a mistake in the speech-to-text process.

Table 2. Machine Learning Models Evaluation Using RMSE and R^2.

		Linear Model	kNN	svmRadial	RF	XbgTree	XgbLinear	Linear Ensemble	Generalized Ensemble
Text	RMSE	0.90	0.90	0.91	**0.78**	0.82	0.79	1.26	0.94

	R^2_{adj}	0.15	0.15	0.16	**0.34**	0.29	0.37	0.04	0.11
Audio	RMSE	0.93	**0.78**	0.88	0.84	0.95	0.87	2.66	2.54
	R^2_{adj}	0.16	**0.41**	0.27	0.32	0.20	0.29	0.25	0.25
Combined	RMSE	0.83	0.71	0.76	**0.59**	0.65	0.64	0.92	0.87
	R^2_{adj}	0.30	0.49	0.44	**0.67**	0.57	0.59	0.59	0.60

DISCUSSION AND CONCLUSION

To improve human-AI teaming, the AI needs to monitor and manage the trust in real-time. Conversational data provides a novel approach to monitor real-time trust. This study trained machine learning models on lexical, acoustic, and combined conversational features as predictors of self-reported trust. The random forest algorithm trained on the combined features explained 67% of the variation of trust, which outperformed the lexical or acoustic features alone. These results show the importance of including both audio and text features when measuring trust in a dynamic situation.

To manage trust through conversation, our next step is to investigate the feature importance. The most important lexical and acoustic features will help identify how conversations reveal trust. Such features might also serve as "conversational affordances" that can be used by the AI to actively measure trust by inserting these affordances into the AI's response to the person and monitoring how the person responds. These and other conversational affordances, such as mirroring and pauses, could also be used to actively manage trust. The conversation might temper trust when trust is too high and repair it when it is too low (Chiou & Lee, 2021).

ACKNOWLEDGMENTS

This work was supported by NASA Human Research Program No.80NSSC19K0654. We would like to thank Sofia Noejovich and Dong Fang for their assistance on the conversational agent development. We also thank members of the University of Wisconsin-Madison Cognitive Systems Laboratory for their insightful discussions and comments.

REFERENCES

Chen, X., Levitan, S. I., Levine, M., Mandic, M., & Hirschberg, J. (2020). Acoustic-prosodic and lexical cues to deception and trust: Deciphering how people detect lies. *Transactions of the Association for Computational Linguistics*, *8*, 199–214.

Chiou, E. K., & Lee, J. D. (2016). Cooperation in human-agent systems to support resilience: A microworld experiment. *Human Factors*, *58*(6), 846–863.

Chiou, E. K., & Lee, J. D. (2021). Trusting automation: Designing for responsivity and resilience. *Human Factors*, *00*(0), 1–29.

Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfeld, A., et al. (2012). Effects of changing reliability on trust of robot systems. *HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 73–80).

Elkins, A. C., & Derrick, D. C. (2013). The sound of trust: Voice as a measurement of trust during interactions with embodied conversational agents. *Group Decision and Negotiation*, *22*(5), 897–913.

Endsley, M. R., Caldwell, B., Chiou, K. E., Cooke, J, N., Cummings, L. M., Gonzalez, C., Lee, D. J., et al. (2021). *Human-AI Teaming : State-of-the-Art and Research Needs*. Washington, DC: The National Academies Press.

Gauder, L., Pepino, L., Riera, P., Brussino, S., Vidal, J., Gravano, A., & Ferrer, L. (2021). A Study on the manifestation of trust in speech. *arXiv preprint* (pp. 1–31).

Hu, W. L., Akash, K., Jain, N., & Reid, T. (2016). Real-time sensing of trust in human-machine interactions. *IFAC-PapersOnLine*, *49*(32), 48–53. Elsevier B.V.

Izygon, M., Kortenkamp, D., & Molin, A. (2008). A procedure integrated development environment for future spacecraft and habitats. *In Proceedings of the Space Technology and Applications International Forum (STAIF 2008)* (p. 969).

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71.

Johnson, M., & Vera, A. H. (2019). No AI is an island: The case for teaming intelligence. *AI Magazine*, *40*(1), 16–28.

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y. C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, *12*(October), 1–23.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Li, M., Alsaid, A., Noejovich, S. I., Cross, E. V., & Lee, J. D. (2020). Towards a conversational measure of trust. *AAAI Fall Symposium FSS-20 / SSS-20* (pp. 1–6).

Li, M., Holthausen, B. E., Stuck, R. E., & Walker, B. N. (2019). No risk no trust: Investigating perceived risk in highly automated driving. *Proceedings - 11th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2019* (pp. 177–185).

McDonald, A. D., Ade, N., & Peres, S. C. (2020). Predicting procedure step performance from operator and text features: A critical first step toward machine learning-driven procedure design. *Human Factors*, *00*(0), 1–17.

McDonald, A. D., Ferris, T. K., & Wiener, T. A. (2020). Classification of driver distraction: A comprehensive analysis of feature generation, machine learning, and input measures. *Human Factors*, *62*(6), 1019–1035.

Nielsen, F. Å. (2011). A new evaluation of a word list for sentiment analysis in microblogs. *ESWC2011 Workshop on "Making Sense of Microposts": Big things come in small packages* (pp. 93–98). Retrieved from http://arxiv.org/abs/1103.2903

R Development Core Team, R. (2011). *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing*.

Schreckenghost, D., Milam, T., & Billman, D. (2014). Human performance with procedure automation to manage spacecraft systems. *in Proceedings of the 35th International Conference for Aerospace Experts, Academics, Military Personnel, and Industry Leaders* (pp. 1–16).

Sebe, N., Cohen, I., & Huang, T. S. (2005). Multimodal emotion recognition. *In Handbook of Pattern Recognition and Computer Vision* (pp. 387–409).

Waber, B., Williams, M., Carroll, J., & Pentland, A. (2015). A voice is worth a thousand words: The implications of the micro-coding of social signals in speech for trust research. *Handbook of Research Methods on Trust: Second Edition* (pp. 302–312).

Yang, X. J., Christopher, S., & Christine, S. (2021). Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Human Factors*, *00*(0), 1–17.