Non-Invasive Measurement of Trust in Group Interactions

Lee A. Spitzley  
*University at Albany, SUNY*

Xinran Wang  
*University of Arizona*

Xunyu Chen  
*University of Arizona*

Steven J. Pentland  
*Boise State University*

Jay F. Nunamaker  
*University of Arizona*

*See next page for additional authors*
Non-Invasive Measurement of Trust in Group Interactions

Lee A. Spitzley, Xinran Wang, Xunyu Chen, Steven J. Pentland, Jay F. Nunamaker, Judee K. Burgoon, Norah E. Dunbar

Abstract— Trust between group members has many implications for how well a group performs. In this study, we predict perceived trustworthiness of group members when there are subversive group members. We collected multimodal verbal and nonverbal data from a group interaction experiment. During the interaction, we periodically surveyed the group members about their perceptions of trustworthiness of other group members. We used this data to model the relationship between observable behavior and perceptions of trustworthiness. We report the most predictive features and describe them in the context of existing literature on verbal and nonverbal correlates of trust. This research advances the study of behavioral measurement in groups and the role of behavior on perceived trustworthiness.

Index Terms— H.3.1.d Linguistic processing, L.2.0.u Multimodal systems, O.1.4 Multi-modal recognition, O.1.1 Nonverbal signals, O.1.2.c Paralanguage analysis, O.1.5 Recognition of group emotion, O.1.2.b Speech analysis

1 INTRODUCTION

GROUP interactions in organizational settings are crucial for achieving meaningful changes. An essential aspect of group success is trust among members. For example, greater levels of trust in groups improves cooperation, motivation, and performance. Further, trust in a group reduces the likelihood of task conflict and relationship conflicts [1]. Although task conflict is thought to be productive it is often correlated with relationship conflict. Trust in a group helps to reduce the likelihood of task conflict, which in turn leads to a reduction in relationship conflicts [1]. In situations where groups are formed rapidly (e.g. committees, task forces and special project teams), the swift formation of trust can also mitigate uncertainty from a lack of shared history between group members [2].

Trust is defined as a basic emotion [3], but has received much less attention from the affective computing community compared to other emotions (e.g. joy, happiness, fear, anger, surprise). Barbalet [4] argues, “… trust is the emotional basis of cooperation” (p. 77). That is, for humans to cooperate, there must be an underlying affective reliance on each other. An ability to detect cues of trust computationally has implications for assessing human-human interactions, such as during negotiation tasks, as well as for assessing human-computer interactions where systems can adapt or intervene in order to build trust.

Various behavioral elements affect the establishment of trust. These include verbal and nonverbal behavior of group members. For example, vocal pitch [5] and affective language [6] correlate with perceptions of trust. Additionally, and perhaps more importantly, a history of actions also affects trust. Each of these elements must be considered while assessing trust. By better understanding behavioral antecedents of trust within group settings, we can design better collaborative and affective systems to optimize trust during interactions.

Research has evaluated the role of trust in dyadic interactions [6], [7]. However, group interactions are not necessarily the same as dyadic interactions [8]. One major barrier in face-to-face group research is the difficulty in capturing information during a group interaction. The study of human behavior, such as trust, in face-to-face group interactions normally requires intrusive data collection tactics, like surveys, to measure relational dynamics and intragroup attitudes. Other studies have focused on the non-invasive, technical methods of collecting observable behavioral data in group settings [9]–[11]. However, research that aligns intragroup attitudes and observable behavioral measures is lacking. A better understanding of this will help with the development of more natural trust evaluation strategies where group members do not need to be interrupted to self-report their attitudes towards others. For example, relief personnel arriving at a scene of a disaster may need assistance in understanding relationships...
among local residents. A goal of this study is to take a step beyond the largely intrusive task of collecting attitudinal data from groups by automatically inferring viewpoints between group members using verbal and nonverbal behavior.

This research investigates the automated prediction of trust in group settings and the relative importance of behavioral measurements at inferring intra-group trust. We propose a set of affective and behavioral measurements for analyzing the evolution of trust during extended group interactions. To this end, an adversarial group dataset is used where opposing objectives lead to trust and distrust between members. Verbal and nonverbal behavior of group members are tracked using noninvasive instruments, and a history of actions and intragroup attitudes are logged using individual tablet computers. The relative importance of each feature is then assessed.

This research extends work in affective computing to shed light on how trust changes based on verbal and nonverbal behavior during group interactions. This approach gives researchers the potential to measure behavior and related outcomes in ways that complement or supersede those of traditional data collection and analysis techniques. With this research, we expand upon affective computing literature around dyadic conflicts and develop a method for understanding trust and duplicity within group interactions.

In the remainder of this paper, we cover the relevant literature on group communication and the role of trust. We review the literature on the relationships between language, voice, and trust. Following that, we describe the dataset. We then conduct a multimodal analysis using verbal and nonverbal behaviors and a history of actions to predict trust. The paper concludes with a discussion and potential future research.

2 Background

In this section, we provide information relevant to trust in groups and the differences between groups and dyads.

2.1 Measuring Intragroup Communication

Intragroup trust has often been a topic for management researchers and sociologists [12]. However, the complexity of analyzing group member behavior and intragroup dynamics limits the advancement of this research stream. A socio-technical solution is needed for efficient and cost-effective data capture and analysis.

Despite current research achievements on monadic and dyadic communication, these results cannot be simply applied to analyzing communication of larger groups. While current methods are very manual, for example, tagging individual speakers and calculating each person’s turn-at-talk, new technologies may alleviate some of this issue.

While there is some debate on whether dyads are groups, we take the view of Levine and Moreland [8] and Dunbar et al. [13] that groups are different from dyads. Moreland [14] pointed out that compared with dyads, groups form and dissolve more slowly. Meanwhile, group members are more loosely connected, and group communication differs dramatically from dyadic communication with respect to emotional intensity and complexity [14]. Whereas turn exchange in dyads is reciprocal (e.g., Speaker A is followed by Speaker B, who is followed by Speaker A), turns in groups can be highly unevenly distributed across group members and cliques and coalitions may be formed.

Lastly, certain social relations, such as majority and diversity can only be applied to groups. Factors such as emotional contagion and interactional synchrony influence the behaviors of groups. These factors may also influence the appearance and reliability of behavioral indicators of trust. Although such factors also apply to dyads, they may operate differently in groups, because groups are more complex [14]. Our research extends current research on dyadic trust to groups and aims to infer perceived trustworthiness of group members.

2.2 Measurement of Perceived Trustworthiness

A majority of existing literature on behavioral indicators of perceived trustworthiness applies traditional statistical analysis methods, such as ANOVA [19, 20, 21], MANOVA [17, 18, 22] and regression [15, 16], for hypothesis testing. Within this stream of literature, some studies examine causal relationships by conducting experiments in which researchers manipulate behavioral indicators (e.g., pitch, linguistic understandability) and investigate how their manipulation impacts perceived trustworthiness [17–22], and other researchers study behavioral correlates of perceived trustworthiness [15, 16]. These studies typically use survey instruments [15, 17–22] or objective trusting behavior [16] as a proxy to measure participants’ perception. Although these studies provide invaluable insights into the behavioral factors that are associated with perceived trustworthiness, they generally study only a few such factors despite the complexities within the broad set of human behaviors. Moreover, these studies mainly focus on hypothesis testing and do not aim to predict the level of perceived trustworthiness during human interaction, although such prediction would help design better collaborative and affective systems to understand the dynamics of human interaction. Overall, this stream of literature does not address the need to predict perceived trustworthiness with a comprehensive collection of multimodal human behaviors.

To bridge this research gap, a nascent stream of literature has begun to demonstrate the feasibility of building computational models to automatically measure perceived trustworthiness using human behaviors. Table 1 summarizes the related papers. Two studies [6, 7] involve dyadic interaction between the individual who evaluates others’ trustworthiness (i.e., the trustee), while the other three studies [23–25] do not involve interaction other than annotation.

Despite the prevalence of groups in our daily lives and the significant role of intragroup trust in group success [1], behavioral indicators of perceived trustworthiness in a group context remain unexplored. Additionally, the two studies [6, 7] with dyadic interaction recruit participants from a single location. To make our findings more robust and generalizable, we collect a diverse sample from multiple locations around the world.
is negatively correlated with perceived trustworthiness [6]. In dyadic face-to-face negotiation experiments, word count is found to positively correlate with perceived trustworthiness, while Wenker et al. [22] showed that disfluent speakers were not perceived less favorably than fluent speakers. Disfluency is expected to decrease speakers’ clarity and indicate their lack of confidence.

Affective language may also affect perceptions of trust. Previous research has shown that verbal positivity is positively correlated with perceived trustworthiness in leadership [17]. High positivity signals intimacy [28], [29], which is correlated with affect-based trust [30]. Verbal caring such as encouragement and praise [18] also conveys speakers’ benevolence, and benevolence is also an important antecedent to trust [31]. Therefore, positive affect is expected to be positively correlated with perceived trustworthiness.

Linguistic complexity also relates to trust. Existing research suggests that cognitively based trust is significantly correlated with comprehensibility of language transmits signals of the senders’ intentions and competence [19]. For example, complex language in business contexts is likely to be interpreted as intentional obfuscation to delay market reactions to unfavorable news [33]. In an educational evaluation setting, sources of recommendations presented with jargon were rated significantly lower in believability and logicality than sources of jargon-free recommendations [34].

Disfluency consists of filled pauses and repeated phrases. Existing studies show inconsistent results about the impact of disfluency on the perception of speakers. For example, Rosenberg and Hirschberg [35] found that the ratio of disfluencies in a speech negatively correlates with ratings of speakers’ charisma, while Wenker et al. [22] showed that disfluent speakers were not perceived less favorably than fluent speakers. Disfluency is expected to decrease speakers’ clarity and indicate their lack of confi-

### Table 1: Studies that predict perceived trustworthiness using behavioral indicators

<table>
<thead>
<tr>
<th>Study</th>
<th>Interaction Type</th>
<th>Participants Recruitment</th>
<th>Scenario</th>
<th>Features</th>
<th>Analytic Models</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>[6]</td>
<td>Dyadic</td>
<td>Recruited from Craigslist and invited to a lab</td>
<td>A multi-issue bargaining negotiation</td>
<td>Facial expressions, head pose/gestures, gaze patterns and conversational features</td>
<td>Naïve Bayes classifier</td>
<td>Overall positivity (+), controlled smiles (-), contempt (-), head pose up/down variability (+), word count (-), gaze left/right avg (-)</td>
</tr>
<tr>
<td>[7]</td>
<td>Dyadic</td>
<td>Recruited undergraduates in Boston</td>
<td>A give-some game</td>
<td>Nonverbal behaviors including face touching, arms crossed, leaning backward, and hand touching</td>
<td>SVMs</td>
<td>Features engineered based on domain knowledge and sequence-based temporal features outperformed baseline models</td>
</tr>
<tr>
<td>[23]</td>
<td>None</td>
<td>Crowdsourced on Amazon Mechanical Turk</td>
<td>Rate Columbia X-Cultural Corpus</td>
<td>Linguistic and prosodic features</td>
<td>Logistic regression</td>
<td>Prosody and disfluency features are the strongest predictors of perceived trustworthiness</td>
</tr>
<tr>
<td>[24]</td>
<td>None</td>
<td>Crowdsourced on ranker.com</td>
<td>Rate videos of public interviews and speeches</td>
<td>Acoustic features</td>
<td>Neural network</td>
<td>Almost all the acoustic features from eGeMAPS were found to be relevant</td>
</tr>
<tr>
<td>[25]</td>
<td>None</td>
<td>Recruited from Amazon Mechanical Turk</td>
<td>Rate Airbnb host profiles (text only)</td>
<td>Linguistic features</td>
<td>Logistic regression</td>
<td>Profile length (+), LIWC categories “we” and “social” (+)</td>
</tr>
</tbody>
</table>

Note: (+) and (-) indicate positive and negative correlation with perceived trustworthiness, respectively.

## 3 Behavioral Measurement

Human behaviors of trustworthiness can be observed through a variety of verbal and nonverbal channels. However, for this current research, signals are limited to those detectable by microphone, specifically, language and voice features. Microphones provide simplicity and flexibility in detectable by microphone, specifically, language and voice features. Microphones provide simplicity and flexibility in group environments, while also producing rich feature sets. Prior research informs the selection of language and voice features as correlates of trust.

### 3.1 Language Measures

The verbal content of one’s messages likely influences how others perceive that individual. Prior work on the relationship between verbal content and perceived trustworthiness has provided some insight into this phenomenon. One of the most basic measures is word count. Consistent with uncertainty reduction theory [26], extra information reduces uncertainty and increases perceived likeability and trustworthiness of others. In line with this expectation, word count is found to positively correlate with perceived trustworthiness in many online contexts, such as online dating profiles [15], self-descriptions of Airbnb hosts [27], and online peer-to-peer lending loan requests [16]. In contrast, in dyadic face-to-face negotiation experiments, word count is negatively correlated with perceived trustworthiness [6], [7], [23].

Affective language may also affect perceptions of trust.
idence [35]. Like disfluency, a high level of hedges also reduces certainty and strength of arguments and leads to less authoritative and more negative perceptions [36], [37].

The use of pronouns may affect perceived trustworthiness through social categorization of an ingroup and an outgroup [38], [39]. Specifically, the use of first-person plural pronouns shows inclusiveness, while the use of second- and third-person pronouns suggests putting conversational partners in the outgroup [15]. Moreover, the use of first-person singular pronouns typically indicates self-orientation and individuality, but under marital problem-solving contexts that entail more negativity, self-referencing is beneficial [40].

3.2 Voice Measures

In many instances, what we say is not as important as how we say it. The human voice has evolved as an essential social function necessary to communicate thoughts, feelings, and attitudes [47]. The characteristics of the voice are also essential for communicating trustworthiness. Lower pitch is correlated with perceived trustworthiness and competence in general [5]. Voters for instance, trust lower-pitched political leaders more because these leaders are perceived as stronger, more competent and more experienced [48]. Syed et al. [24] reports significant correlation between acoustic features and public trust. Highly trusted individuals usually show a greater concentration of energy at the lower spectrum of their voice, which makes their voice sound deeper. However, the directionality of the pitch and trustworthiness relationship can vary based on demographics and circumstances. For example, women tend to trust high-pitched males when dating [21] and dividing up financial gains [20]. Vocal variety relates to nonverbal immediacy and enhances closeness, while a monotone voice indicates a low level of immediacy [18]. Because closeness and intimacy are correlated with affective-based trust [32], we can expect vocal variety to be associated with perceived trustworthiness. Evidence of this is also seen in pitch contours where varied intonation is viewed as more trustworthy compared to flat intonation [49]. Speech rate is found to enhance persuasion [50] and perceptions of competence [51]. Overall, faster rates of speech are more positively viewed. For example, a faster rate of speech correlates with a job applicants hireability score [52]. We can expect the positive characteristics associated with speech rate to increase perceptions of trust.

4 DATA COLLECTION

To collect data for this study, we conducted an experiment using a scenario modified from the popular party games of Mafia and Resistance [53], [54]. A complete description of the data collection can be found in [55].

Although no game or experiment task can replicate real-world scenarios perfectly, Mafia/Resistance was selected as the basis of the data collection since it includes a variety of aspects that mirror real-world group interactions. These aspects include: open discussion, differing roles, competing interests, potential deception, longer engagement over multiple episodes, and zero-acquaintance problem solving. Similar characteristics can be seen in business negotiation settings, distributed collaboration (e.g., varying business units working on a joint project) or soldier-citizen interactions during war. Like real-world settings, Mafia/Resistance allows participants to build trust over the course of interactions. A violation or perceived violation of this trust will change how individuals interact and how outcomes are formulated.

Unlike classic versions of the game, our scenario did not eliminate a player each round, giving each participant an equal amount of time for participation. A major goal of the experiment was to create an environment for collecting verbal and nonverbal cues that signal trust-distrust relationships between people participating in group communication. Groups of five to eight participants were seated around a portable assembly of tables equipped with laptop computers.

Each participant faced a laptop which had custom game facilitation software installed. The software allowed players to perform tasks (e.g., view their assigned roles, vote) during the game and answer survey questions between game rounds. The built-in cameras on the laptops, along with three other video cameras, captured video and audio data. Figure 1 illustrates the experimental setup, and Figure 2 shows the key components of the data collection procedure and an overview of the subsequent workflow.

The experiment started with an icebreaker activity which required each participant to introduce themselves briefly and answer a follow-up question from another participant sitting across the room. Next, we randomly assigned two to three players as a “Spy” (deceivers) or a “Villager” (truth-tellers) depending on the number of participants in the session. Only spies knew who the other spies were. Spies were to hide their identities and sabotage hypothetical missions, and they won a point every time a mission failed. In contrast, villagers were to discover the spies and protect the missions by excluding the spies from mission teams, and they won a point every time a mission succeeded. The games lasted up to one hour or eight rounds.

In each game round, the group first selected a mission...
leader by voting on the game app. Once the group successfully chose a leader, the leader would select a few other players to join himself / herself on a mission team. The group then voted on the proposed mission team. Once the team was approved by a majority of group members, those on the team would vote secretly for the mission to “succeed” or “fail”. The final winners were the side with more points, and they would receive extra monetary rewards in addition to the rewards for participation. Villagers always chose to succeed a mission, and spies must choose whether to succeed or fail a mission depending on their strategies to fail the missions as many times as possible. A facilitator used a script for experimental control and to encourage group discussion without favoring either side.

After the icebreaker, and then every two rounds during the game, participants completed brief surveys and rated other members on trustworthiness. These measures constitute the perceptions of trustworthiness in our analysis.

4.1 Sample Statistics

We played 95 games in eight universities around the world (United States (3 separate sites), Israel, Singapore, Fiji, Hong Kong, & Zambia). Automated speech recognition (ASR) services were employed to transcribe these games first. Given the difficulty of transcribing group interactions in various cultures using ASR services (mentioned below), we recruited research assistants to manually correct the machine-generated transcripts to ensure the quality of transcripts used in our analysis. Because of the associated high costs, we obtained manually corrected transcripts of 40 games. Due to technological issues in the capture of these games, we were able to use a subset of 28 games with 190 players (114 females and 76 males) for our analysis. Specifically, this sample included 5 games (35 players) from the Eastern US, 6 games (39 players) from the Southwestern US, 6 games (40 players) from the Western US, 2 games (15 players) from Hong Kong, 7 games (49 players) from Singapore, and 2 games (12 players) from Israel. The high diversity is an important characteristic of this sample and facilitates learning of cues to perceived trustworthiness that are generalizable across distinct regions.

184 participants reported their age, ethnicities and native languages. The mean and standard deviation of age were $M = 21.90$ and $sd = 2.76$ years. Additionally, 54.9% were Asian, 28.8% were white, while Hispanic, multiracial, and other individuals constituted 8.2%, 5.4% and 2.7%, respectively; 60.9% were native English speakers.

There were 114 participants assigned to be villagers, while 74 were spies; 67.4% had played a game like our experiment scenario before. Each participant participated in the experiment one time. The average number of rounds in a game was 5.8, with each round averaging 7.11 minutes each. Villagers won 16 games; spies won 12.

4.2 Behavior Ratings

After every two rounds of gameplay, each player rated on a five-point Likert scale how trustworthy they perceived the other players to be. Since the spies were aware of the true roles of the other players, their ratings of trustworthiness may be biased to trust other spies. Therefore, we only used the perceptions of the villagers in this analysis. Specifically, for every two rounds, we averaged the ratings from villagers to form a trustworthiness score for each player. Self-ratings were excluded. In total, the 28 games in our sample included 553 instances of trustworthiness scores, and the mean and standard deviation of these scores were $M = 3.27$ and $sd = 0.93$. We use the raw trustworthiness score (i.e., average ratings from villagers) as the response variable and formulate the prediction task as a supervised regression problem.

Trust ratings on a continuous scale are useful to identify behaviors and actions that move scores of perceived trustworthiness. However, relatively high or low trustworthiness ratings per round are important because decisions within a round are based on the available information and the relative levels of perceived trustworthiness at that time.
The nature of the game required the participants to make frequent decisions on who to trust. That is, while participants rated trust on scale, they were ultimately required to choose those they trusted to participate as a leader or team member in each game round. Studying the relative level of perceived trustworthiness among group members mirrors real-world scenarios such as business negotiations. In these scenarios, understanding who is perceived as relatively more trustworthy than other group members (given a specific interactive progress and information available at the time of decision making) potentially informs negotiation strategies.

For classification purposes, we dichotomized the trustworthiness scores to construct the predicted variable. Given the variance between games and locations, we did not use one simple cutoff point for all the trustworthiness scores. Instead, we labeled a trustworthiness score to be high if it was higher than or equal to median score of all the players in the same game and the same round. Otherwise, we labeled it as low. Among the 553 instances, 58.6% of them were labeled as high. We used this binary variable of high or low trustworthiness as the predicted variable in the following analysis.

4.3 Data Preparation

Data preparation was aimed at two major goals, namely, to obtain accurate transcripts and to segment audio into turn-at-talk clips. To obtain accurate transcripts, we conducted several data processing tasks: synchronization, round segmentation, audio extraction, audio transcription, manual correction, and speaker identification. To facilitate synchronization, the game facilitator played a pre-recorded “ding” sound through a speaker at the beginning of each round. This sound was captured by microphones on all the devices, and it was used to synchronize the video signals from all the sources. We segmented the front-facing videos of each player from an entire game into clips for each round based on the “ding” timestamps.

For the current study, we focused on turn-at-talk analysis. A turn-at-talk can be defined as a segment of time where a participant speaks to the group. Although video data capturing facial displays were available, we focused specifically on voice and language data for two reasons: 1) video data, specifically facial display data, exists beyond turns-at-talk. Participants may produce reactionary facial displays as they listen to other group members. Combining continuous facial data with sporadic turn-at-talk data offers methodological challenges that are better examined in a separate study. 2) Facial data during turns-at-talk largely reflects the articulation of the mouth and face during speech. Therefore, face behaviors during speech are likely already reflected in language and voice behaviors.

Next, we extracted the audio stream from video files of every player in each round. For audio transcription, we used IBM Watson Speech-to-Text [56] to produce player-level transcripts. These machine-generated, game-level transcripts had high word error rates (WER) and could not label individual speakers. Research assistants manually corrected these rough transcriptions. The research assistants also tagged speakers in the final transcripts.

Additionally, we performed audio extraction, speech detection, audio alignment, loudness measurement, and speaker identification to segment audio into turn-at-talk clips. We first extracted audio from the game-level videos of each player and detected speech in the audio. Then, we aligned audio of utterances using dynamic time warping [57]. Next, we measured the loudness of utterances in the aligned audio of different players and identified speakers assuming the loudest signal came from the microphone that was the closest to the speaker. Lastly, we segmented audio clips of turns-at-talk for each player. To improve the accuracy of an audio segment (not just noise or cross-talk), audio segments shorter than two seconds were not included.

4.4 Feature Extraction

After obtaining the complete and accurate transcripts for each game, we concatenated the transcripts of two rounds at a time for each player. For example, if a game had six rounds, we would have three text instances for each player. In this way, we could match the preceding linguistic features of each player with the perceived trustworthiness ratings at the end of every two rounds. We quantified the text using SPLICE, a software program that extracts linguistic features, such as readability and ratio of first-person singular words [58]. We excluded a feature from further analysis if its standard deviation was zero or it equaled zero in more than 70% of the instances. We also normalized the count-based features (e.g., the number of modal verbs that suggest hedging) by the number of words. In total, 72 linguistic features were retained for subsequent analysis.

Furthermore, we processed the turns-at-talk audio segments with a tool for audio signal analysis named openSMILE [59] and extracted 106 vocalic features, such as mean and standard deviation of loudness and pitch. Lastly, players had perceived behaviors of other players in the same game session before they rated each other’s trustworthiness at the same time. Therefore, perceptions of trustworthiness should largely rely on the relative level of behavioral features (in addition to game-specific variables as mentioned below). To account for the within-group comparison of behavioral features, both vocalic and linguistic features were standardized within the same game session and the same round.

4.5 Controls and Game-Specific Variables

We collected data on player characteristics as control variables, including age, sex, prior game experience, and whether the person was an English native speaker. We also accounted for the information revealed throughout the games that could influence trustworthiness perceptions. These game-specific variables included the number of times a person was the leader of successful and failed missions and the number of times a person was a member of successful or failed mission teams. Similar to the behavioral features, the control and game-specific variables were standardized within the same game session and the same round.
5 TRUST PREDICTION

5.1 Classification Analysis

To understand the predictive ability of each modality, we fit models with various subsets of the behavioral features, control variables, and game data. These configurations are shown in Table 2.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Unimodal or Multimodal</th>
<th>Variables Considered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic Model</td>
<td>Single-modal</td>
<td>Linguistic variables</td>
</tr>
<tr>
<td>Vocalic Model</td>
<td>Single-modal</td>
<td>Vocalic variables</td>
</tr>
<tr>
<td>Ling + voc Model</td>
<td>Multimodal</td>
<td>Linguistic + vocalic variables</td>
</tr>
<tr>
<td>Control Model</td>
<td>Single-modal</td>
<td>Control variables</td>
</tr>
<tr>
<td>Game Model</td>
<td>Single-modal</td>
<td>Game-specific variables</td>
</tr>
<tr>
<td>Multimodal Model</td>
<td>Multimodal</td>
<td>Linguistic + Vocalic + Control + Game-specific variables</td>
</tr>
</tbody>
</table>

The high dimensionality of the behavioral features relative to the sample size creates risk of overfitting. To reduce the likelihood of this occurring, we designed a variable selection algorithm based on the classic stepwise variable selection [60] that first reduces the feature set and then iteratively increases model size. Figure 3 describes the modified stepwise variable selection (MSVS) algorithm used in this study.

The algorithm starts with an empty model and then tests the addition of each variable in the feature set \( F \). In the first step of variable selection, we selected the 60 most predictive variables \( P \) based on prediction accuracy. Following step one, all combinations of these 60 variables were tested to create the first best model set \( Q \) that has a size of 40. MSVS kept the same \( P \) and updated \( Q \) in each following step. Specifically, every variable in \( P \) was added to each model in \( Q \) to create larger models. These new models were then evaluated, and the best 40 of them replaced the old models in \( Q \). In each step, a model (combination of a feature subset and a classifier) was tested with 100 random train/test splits, which is a similar practice to cross validation. To prevent information leakage between the train and test sets, we split the 28 games into a 21-game training set and a 7-game testing set, ensuring players in the training set appeared exclusively in the training set, and the same was true for players in the testing set. Prediction accuracies from each split were averaged to offer a more balanced evaluation. Aside from prediction accuracy, we also reported another valuation criteria, F1 score. To reduce the computational requirements and increase the parsimony of the models, we set the maximum model size \( M \) at 20.

We selected six different classifiers based on their fundamentally different fitting processes and their familiarity in the machine learning community: random forest, logistic regression, support vector machines, naïve Bayes, bagging, and boosting. For the models containing behavioral features, MSVS was run on each train/test split. The control and game-specific variables had relatively few features and were fit without using variable selection. This process was repeated for each of the classification algorithms.

MSVS required heavy computation to execute. To address this, we parallelized the computation on a supercomputer. The efficient parallelization algorithm completed over 2000 single-threaded computing hours in 10 hours, which significantly reduced the time of variable selection. Generally, model performance stabilized when models contained between 5 and 10 variables. Next, we will report the results of our classification models.

5.2 Classification Results

The fitting process resulted in 100 models for each combination of feature sets and classifiers. The performance of a combination was evaluated by averaging the accuracy/F1 score realized by the 100 random splits. The performance from each combination is reported in Table 3 (Accuracy) and Table 4 (F1).

<table>
<thead>
<tr>
<th>Model label</th>
<th>RF</th>
<th>LR</th>
<th>SVM</th>
<th>NB</th>
<th>BAG</th>
<th>XBG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistics</td>
<td>0.604</td>
<td>0.61</td>
<td>0.61</td>
<td>0.614</td>
<td>0.611</td>
<td>0.602</td>
</tr>
<tr>
<td>Vocales</td>
<td>0.62</td>
<td>0.646</td>
<td>0.646</td>
<td>0.657</td>
<td>0.657</td>
<td>0.622</td>
</tr>
<tr>
<td>Control</td>
<td>0.559</td>
<td><strong>0.582</strong></td>
<td>0.582</td>
<td>0.572</td>
<td>0.568</td>
<td>0.567</td>
</tr>
<tr>
<td>Game</td>
<td>0.684</td>
<td>0.692</td>
<td>0.692</td>
<td>0.689</td>
<td>0.686</td>
<td>0.674</td>
</tr>
<tr>
<td>Ling + Voc</td>
<td>0.622</td>
<td>0.667</td>
<td>0.668</td>
<td><strong>0.672</strong></td>
<td>0.661</td>
<td>0.622</td>
</tr>
<tr>
<td>Multimodal</td>
<td><strong>0.712</strong></td>
<td>0.736</td>
<td><strong>0.743</strong></td>
<td>0.735</td>
<td>0.731</td>
<td>0.719</td>
</tr>
</tbody>
</table>

Since the trustworthiness scores were created by splitting at the median, the trust labels were 58.6% “high trust” and the remainder were labeled as “low trust” Therefore, accuracy above this level is considered an improvement. We also used F1 to evaluate the balance between precision and recall.

<table>
<thead>
<tr>
<th>Model label</th>
<th>RF</th>
<th>LR</th>
<th>SVM</th>
<th>NB</th>
<th>BAG</th>
<th>XBG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistics</td>
<td>0.723</td>
<td><strong>0.732</strong></td>
<td>0.732</td>
<td>0.732</td>
<td>0.724</td>
<td>0.692</td>
</tr>
<tr>
<td>Vocales</td>
<td>0.72</td>
<td>0.74</td>
<td>0.741</td>
<td><strong>0.743</strong></td>
<td>0.741</td>
<td>0.708</td>
</tr>
<tr>
<td>Control</td>
<td>0.701</td>
<td><strong>0.734</strong></td>
<td><strong>0.734</strong></td>
<td>0.724</td>
<td>0.718</td>
<td>0.68</td>
</tr>
<tr>
<td>Game</td>
<td>0.762</td>
<td>0.764</td>
<td>0.765</td>
<td>0.762</td>
<td><strong>0.772</strong></td>
<td>0.75</td>
</tr>
<tr>
<td>Ling + Voc</td>
<td>0.735</td>
<td>0.75</td>
<td><strong>0.751</strong></td>
<td>0.746</td>
<td>0.741</td>
<td>0.708</td>
</tr>
</tbody>
</table>
Multimodal 0.78 0.792 0.793 0.779 0.783 0.775

Figure 4 reports the best accuracy and F1 score that each model family achieved in our experiment. Perhaps unsurprisingly, the control model was the least predictive model with accuracy below 60%, which is slightly better than a random guess. The poor prediction implies that the establishment of trust does not rely much on demographics.

The best linguistic model outperformed the control model by a small margin (5.5%). Meanwhile, the vocalic model achieved 65.7% accuracy on predicting perceived trustworthiness, which is a 7% improvement over the linguistic model, and a 12.9% improvement over the control model. When we consider both the linguistic and vocalic variables, the “Ling + Voc” model achieved 67.2% accuracy, which is a 15.4% improvement over the control baseline.

However, the behavioral models still cannot beat the game model built with game-specific variables only, which indicates that the interpersonal trust is largely driven by group outcomes. Though the behavioral models were outperformed by the game model, adding behavioral variables to the game model led to higher accuracy. The all-inclusive multimodal model achieved 74.3% accuracy on predicting the trust score, which is a 7.4% improvement over the game model.

Figure 5 conveys a similar message. The ROC curves plot the true positive rate against the false positive rate calculated at various thresholds, and a larger AUC score corresponds to a larger area under the ROC curve and a better predictive performance. The multimodal model achieved the highest AUC score, followed by the game model, the behavioral model (“Ling + voc” model) and the control model. Generally, the ROC curve of the multimodal model dominates the curves of the other models.

5.3 Continuous Prediction
The classification analysis demonstrated that it is possible to predict whether someone is trusted relative to other group members. To deepen our understanding of trustworthiness in groups in the context of prior literature, it is also of interest to understand which features were predictive of the original trustworthiness score on a Likert scale from 1-5.

We conduct this analysis using five continuous prediction models: random forest, SVM, multiple linear regression, bagging and boosting. Feature importance measures produced by these models are used to reflect the predictive power of a behavioral variable. In this case, we are interested in interpretable models in addition to predictive performance. As such, we simply select the (n-1) features that have the highest feature importance scores among the n predictors. The process stops when there are no predictors left. Root Mean Squared Error (RMSE) was used to evaluate regression performance using 21 games to train and 7 games to test, with game-level randomization occurring for each model size, such that the games in the train and test sets change with each iteration. Like the original classification analysis, there are 72 linguistic features and 106 vocalic features in our regression data set.

5.4 Continuous Prediction Results
Figure 6 shows the relationships between model size and out-of-sample performance. When the model size was relatively small, the out-of-sample RMSE was also relatively low. When a model contained too many behavioral features with low prediction power, its out-of-sample RMSE increased. Boosting usually produced the lowest RMSE among the five machine-learning regressors. It was also among the least affected models when including low-predictive features, i.e., the RMSE curve for boosting is flatter than most of the others. The linguistic models produced lower RMSE than the vocalic models for three out of the five machine learning algorithms. Meanwhile, the multimodal approach (Linguistic + Vocalic) was superior in RMSE to both unimodal models. The best RMSEs achieved by each family of models are reported in Table 5. The multimodal model is also the most predictive model we found in the variable selection process.
Table 5: RMSE by Model Configuration

<table>
<thead>
<tr>
<th>Model Configuration</th>
<th>Bagging</th>
<th>Boosting</th>
<th>MLR</th>
<th>Random Forest</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic Model</td>
<td>0.868</td>
<td>0.848</td>
<td>0.829</td>
<td>0.823</td>
<td>0.845</td>
</tr>
<tr>
<td>Vocalic Model</td>
<td>0.883</td>
<td>0.832</td>
<td>0.826</td>
<td>0.841</td>
<td>0.855</td>
</tr>
<tr>
<td>Ling + voc Model</td>
<td>0.838</td>
<td>0.806</td>
<td>0.804</td>
<td>0.797</td>
<td>0.839</td>
</tr>
<tr>
<td>Multimodal Models</td>
<td>0.599</td>
<td>0.539</td>
<td>0.653</td>
<td>0.563</td>
<td>0.661</td>
</tr>
</tbody>
</table>

5.5 Important Features

The best-performing continuous models contained a mix of vocalic, linguistic, and game-related features. These features are listed in Table 6. Because the multimodal feature set achieved the highest classification accuracy and the lowest RMSE, we focus on interpreting the top 10 most important features of this model. This list includes three game-specific variables, namely, the number of times the individual joins a successful mission, the number of times the individual joins a failed mission, and the number of times the individual is the leader of a successful mission.

The seven behavioral features in Table 6 were repeatedly selected as important by the linguistic model, the vocalic model, and the combined model with both linguistic and vocalic inputs. Repeat phrases ratio, present tense ratio, third person singular ratio and third person plural ratio are the most important linguistic features. Repeat phrases ratio has a marginal positive correlation with perceived trustworthiness \((r=0.074, p=0.081)\) in our dataset (correlations reported in this section are all Pearson correlations). This finding conflicts with the negative correlation identified in a setting without interaction between the trustor and trustee [23]. However, this finding is similar to Wenker et al., [22], which finds that disfluent speakers are perceived as more trustworthy than fluent speakers and that disfluent speakers are perceived as more trustworthy in live presentations than in audiotaped presentations. Therefore, the inconsistency between our findings and [23] could be attributed to live versus recorded conditions.

Additionally, present tense ratio has a marginal negative correlation with perceived trustworthiness \((r=-0.072, p=0.090)\) in our dataset. Similar to our findings, present tense verbs have been found to contribute to lower perceived trustworthiness, moderated by length of Airbnb host profiles [25]. While [15] indicates significant and negative correlation between second-person pronouns and perceived trustworthiness of online dating profiles, our dataset yields insignificant correlation between the same constructs \((r=-0.041, p=0.335)\). Instead, we find third person (both singular and plural) pronoun ratio to be important in our group conversational context. Third person singular pronouns \((r=0.113, p=0.008)\) and third person plural pronouns \((r=0.068, p=0.108)\) are both positively associated with perceived trustworthiness in our dataset.

Moreover, although word count is found to negatively correlate with perceived trustworthiness in dyadic interaction [6], we find the correlation to be positive \((r=0.180, \ldots)\).
p<0.001). Meanwhile, word count is not among the most important features of the multimodal model. The different directions in correlation could be attributed to different expectations of the amount of speech in dyadic and group interaction. Group members are less accountable to speak [61], so those who make the effort and contribute to group discussions are perceived favorably, while the pattern does not hold in dyadic interactions in which turn-taking is assumed.

A few features involving mel-frequency cepstral coefficients (MFCCs) and line spectral pair frequencies (LSPs) are found to be important in the multimodal model. Syed et al. [24] finds audio recordings with high and low perceived trustworthiness differ in similar complex frequency-based features. One of the most studied vocalic features, the average pitch [20], [21], [23], is not among the most important features of the multimodal model. There is no correlation between the average pitch and perceived trustworthiness (r=-0.003, p=0.949) in our dataset. Instead, more complex frequency-based features that reflect dynamic changes in pitch and intonation lead to differences in perceived trustworthiness [49].

| Table 6: Full Regression Model Ordered by Importance (L+V+Control+Game) |
|----------------|-----------------|-----------------|-----------------|-----------------|
| Mean of the line spectral pair frequencies computed from the 6th LPC coefficient | Standard deviation of the 1st order differential of mel-frequency cepstral coefficients (MFCCs) 2 | The number of times the individual joins a successful mission | The number of times the individual joins a failed mission |
| The number of times the individual is the leader of a successful mission | Repeat phrases ratio | Present tense ratio | Third person singular ratio |
| Third person plural ratio | Mean of MFCCs 6 |

### 6 Contributions and Implications

Trust among collaborators is essential to the successful completion of tasks. It reflects the affective overtone of group interactions. The importance of trust is amplified in group settings where several competing interests can derail progress towards a common good. Group members must believe that others are working towards shared interests and are putting aside selfish motivations. When emotions are running high in intragroup conflict, trust is one of the first casualties. Understanding factors that affect trust in a group setting is important from a social science perspective but is also critical to the development of collaborative and affective systems. To advance scholarly understanding of factors that influence trust in group settings, we implemented a semi-autonomous methodology that captures critical components of group interactions including group members’ behavior and sentiment, as well as the history of group member actions. The results demonstrate the relative importance of behaviors that influence trust.

Past research on group interactions largely assumes shared motivations between group outcomes. However, in real world situations, we cannot assume that goals of the group are shared. Individuals or subgroups may have their own motivations that are counter to defined group objectives. The design used in the present study injects suspicion and distrust into interactions, which allows us to better understand how trust is created and sustained.

This research contributes to affective computing in several ways: First, this research demonstrates that it is possible to measure interpersonal trust in group interactions with minimal intervention from researchers. The method implements a series of recording devices to capture the verbal and nonverbal behaviors of group members. Data captured from these devices were processed to extract linguistic and vocalic features. Although manual speech transcription was utilized due to poor automated speech recognition (ASR) performance, we view this as a function of current ASR limitations that will improve with time. A critical component of the methodology was the use of individual tablet computers to collect perception of trust periodically. The customized software solution allowed group members to record their sentiment as well as allowed members to record their actions when critical decisions were made. This allowed trust sentiment data to be collected without intervention from researchers which would interrupt group interactions. Since trust can grow or deteriorate over time, this systematic approach to tracking group interactions also allowed for the analysis of historic actions.

Second, the results of our analysis inform the design of systems where building trust is a critical component of task success. In our analysis, we considered three primary categories that affect trust: historic actions, language content, and vocal behavior. At the individual category level, historic actions of group members were most predictive of trust. For instance, using historical game actions alone as input to an SVM classifier resulted in an accuracy rate of 69.2% and an F1 score of 76.5%. This intuitively makes sense. If a group member was involved with a failed task, then that failure was remembered by the other players, and subsequently their trustworthiness decreased. When developing systems, this finding emphasizes the importance of tracking actions throughout interaction and including temporal activities into decision engines. Our findings also show that when historic actions are not available, a combination of verbal and nonverbal data can be used to approximate trust perceptions.

Third, our study reinforces the importance of considering multiple behavioral channels when developing affective systems. None of the individual categories of behaviors that we assessed were as predictive as the combination of behaviors. Trust prediction performance was highest when control variables, historic actions, language content, and vocal behaviors were fused together and accomplished an accuracy rate of 74.3% and an F1 score of 79.3%. Behavioral channels interact to provide greater insight into a message and underlying motivations. For instance, [18] finds significant interaction effects between nonverbal im-
mediacy (e.g., high vocal variety) and verbal caring on perceived trustworthiness in an educational context. Designers for affective systems should consider the trade-off between predictive performance and computational cost. For example, when the primary goal of affective systems is to assess individuals accurately, and measurement of multimodal data incurs a reasonable marginal cost (e.g., infrastructure for speech-to-text systems has already been set up), they should incorporate multiple behavioral channels when possible. If setting up the infrastructure is too costly, system designers should prioritize analysis of historic actions of group members or the modality that is easy for them to measure.

6.1 Limitations and Future Work

There are several important limitations to acknowledge in this work. One potential limitation of this research is that survey measures of trusting intentions may cause variations in behavior that would not have otherwise occurred in the absence of surveys. Objective measures of trust, like funding success of Peer-to-Peer (P2P) loan requests [16], could mitigate this issue in some contexts.

Aside from the lack of technical tools for matching and tracking senders and receivers in group interactions, it is even more challenging to identify the different parties in large group communication. For example, when a speaker is talking in front of a group, the speaker might be addressing the whole group, or might be just talking to their neighbor, which makes identifying the message receivers difficult before a thorough understanding of the context. Other research in this area has measured personality traits with social network measures [62], which could be adapted to identify one-to-one communication versus that which is directed at the entire group.

Our sample contains data from a variety of cultures around the world and allows us to draw some tentative generalizations across them, but because of the relatively small number of games from each location, we are not able to conduct a more thorough analysis of individual differences. Future work in this area should examine the interplay between culture, behavior, and trust.

Further, our study evaluated trust for only one situation and setting. It is likely that cues to trust will vary depending on high-stake versus low-stake situations and varying modalities of communication such as sitting versus standing, ability to move further or closer to group members, or speech recognition, natural language processing, and video-based analysis.

7 CONCLUSION

Affective computing research has largely focused on recognizing basic emotions in individuals. As technology advances and affective responses further explored, it becomes apparent that researchers need to continue to push the boundaries of affective understanding from a multi-person perspective. Trust plays a vital role in the progress of group collaboration and how affective bonds between people develop. However, there is little affective computing research on behavioral indicators of perceived trustworthiness in a group setting. The research presented in this paper represents a step towards a better understanding of affective computing in multi-person environments.

8 ACKNOWLEDGMENT

This work was supported in part by a grant from Army Research Office Grant W911NF-16-1-0342. (PI: Judee Burgoon).

9 REFERENCES


[37] V. Smith, S. A. Siltanen, and L. A. Hosman, “The effects of powerful and powerless speech styles and


Lee A. Spitzley received his Bachelor of Science degree in 2010 and his Master of Business Administration degree in 2012, both from Central Michigan University. He received his Ph.D. in Management Information Systems from the University of Arizona in 2018. Lee is an Assistant Professor of Information Security and Digital Forensics at the University at Albany School of Business. His research interests include understanding how language and nonverbal behavior reflects underlying psychological states and traits, linking language measures to psychological constructs, and using these methods to solve real-world problems in the areas of financial fraud, job interviews, and group interactions.

Xinran Wang received the Bachelor’s degree of Management Sciences with a specialty in Accounting from Zhejiang University in 2016 and the degree of Master of Science in Accounting from Michigan State University in 2018. She is a doctoral student in the Department of Management Information Systems at the University of Arizona and a research associate for the Center for the Management of Information (CMI). She has primarily worked on analyzing multi-modal behavioral cues of deception and other relational constructs (e.g., dominance, trust, and nervousness) in cross-cultural group interaction and automated interviews. She is a student member of Association of Information Systems (AIS) and Association of Computing Machinery (ACM).

Xunyu Chen received his Bachelor’s degree in Statistics from Renmin University of China in 2015 and the degree of Master of Science in Statistics from the University of Illinois in 2018. He is working towards a Ph.D. in the Management Information Systems program at the University of Arizona. His research interests include deception detection, affective computing, and group communication.

Steven J. Pentland received his Bachelor’s degree in Public Administration from the University of Arizona in 2008. He later received his Ph.D. in Management Information Systems in 2018 also from the University of Arizona. Currently, he is an Assistant Professor of Information Technology & Supply Chain Management at Boise State University. He has contributed to a variety of projects supported by the National Science Foundation, Department of Homeland Security, and the Army Research Office. Dr. Pentland’s research interests include automated interviewing, affective computing, and human analytics. His current research focus is in the area of deception detection and personnel selection.

Jay F. Nunamaker Jr. is Regents and Soldwedel Professor of MIS, Computer Science and Communication, and director of the Center for the Management of Information and the National Center for Border Security and Immigra-

tion at the University of Arizona. He received his Ph.D. in Operations Research and Systems Engineering from Case Institute of Technology. Dr. Nunamaker has held a professional engineer’s license since 1965. He was inducted into the Design Science Hall of Fame and received the LEO Award for Lifetime Achievement from the Association for Information Systems. He was featured in the July 1997 issue of Forbes Magazine on technology as one of eight key innovators in information technology. His specialization is in the fields of system analysis and design, collaboration technology, and deception detection. The commercial product GroupSystems ThinkTank, based on his research, is often referred to as the gold standard for structured collaboration systems. He founded the MIS Department at the University of Arizona and served as department head for 18 years.