Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

Believe It or Not: Acoustic-Prosodic Cues to Trust and Mistrust in Spoken Dialogue

Sarah Ita Levitan[1], *Julia Hirschberg*[2]

[1]Hunter College, City University of New York
[2]Columbia University

sarah.levitan@hunter.cuny.edu, julia@cs.columbia.edu

Abstract

Trust is a fundamental component of human-human and human-computer interaction. In this work we examine the acoustic-prosodic features of trust in a corpus of interview dialogues. While previous studies have explored the characteristics of speech that is trusted or mistrusted by others, we study a complementary problem: what are the characteristics of trusting vs. mistrusting speech? That is, are there specific acoustic-prosodic cues in an interviewer's speech that indicate whether the interviewer believes their interlocutor, or whether they are skeptical? We use a corpus of deceptive and truthful interview dialogues, where trust labels are explicitly provided by the interviewer for every question asked. We analyze acoustic-prosodic features extracted from interviewer turns and compare the features of trusting and mistrusting speech, finding several significant differences in features. Furthermore, we compare the features of trusting speech in our study of human-human dialogue, with previous findings from a study of trusting speech in human-computer dialogue. This work sheds light on the nature of trusting speech, and how it manifests itself when humans communicate with human vs. machine interlocutors.

Index Terms: human-computer interaction, computational paralinguistics, trust

1. Introduction

Trust is a critical component of all forms communication. It enables interlocutors to collaborate effectively and is a key factor which contributes toward successful interactions. In order to understand trust, researchers across many fields have sought to find a specific signal or set of signals of trust. In particular, researchers have explored nonverbal and verbal cues that make a person more likely to be trusted by others, i.e. cues to trustworthiness [1, 2].

However, very little work has been done to understand the nature of *trusting* speech and behavior. That is, what are the verbal or non-verbal characteristics that indicate that a speaker trusts their conversational partner? We address this problem in this work and focus on studying cues to trusting speech. Our goal is to identify specific acoustic-prosodic features that characterize the speech of someone who trusts their conversational partner. Recent work by [3] explored the nature of trusting speech in human-machine interaction; in this work we focus on human-human interaction but there would be many applications for our findings to human-machine interaction.

Specifically, we address the following research questions in this work:

1. What are the acoustic-prosodic characteristics of trusting and mistrusting speech?
2. Are there universal characteristics of trusting speech, or are there differences in production of trusting speech across gender and native language of the speaker?
3. How do cues to trusting speech in human-human interactions compare with cues previously identified in human-computer interaction?

There are a number of potential applications for this work. Identifying cues to trusting speech in humans can be particularly useful for identifying problems and then improving human-machine interactions. For example, we envision a system that can monitor whether a user trusts it or not, and can leverage that information to try to build the user's trust. It can then check to see whether those efforts were successful, i.e. whether cues to trust are present in the user. This approach would be particularly important for recommender systems that aid humans in shopping or selecting music to listen to or movies to view [4, 5]. It would also be particularly important in conversational systems or robots that provide exercise or other health-care advice or advice on how to improve one's business effectiveness or news stories or even conversational story-telling[6, 7, 8, 9, 10, 11].

In addition, identifying cues to trusting speech can be used to gain insights about successful vs. unsuccessful human-human dialogues, perhaps by inferring levels of trust between interlocutors from the patterns of their speech.

The remainder of this paper is organized as follows. Section 3 describes the data used for this work. In Section 4 we describe the methods used for feature extraction and analysis of trusting speech. We present an analysis of acoustic-prosodic indicators of trusting vs. mistrusting speech in Section 5, and include further analysis of differences across gender and native language. We compare the results of our study with a prior study of cues to trusting speech in human-computer interaction in Section 6. We conclude in Section 7 with a discussion of our findings and ideas for future work.

2. Related Work

There has been little work done to identify the prosodic characteristics of trusting speech. However, prior work has studied a complementary problem — the characteristics of trustworthy speech, or speech that is trusted by others. [12] conducted an experiment in which they manipulated the prosody of sythesized truthful and deceptive statements, and then asked subjects to judge the synthesized speech as true or false. They then identified prosodic features associated with perceived trustworthiness.

[13] and [2] studied cues to trustworthy speech in interview dialogues, where interviewees lied or told the truth to biographical questions, and interviewers provided judgments of decep-

10.21437/SpeechProsody.2022-124

tion for each question. They analyzed the interviewee speech and compared features of speech that was trusted by the interviewer with the features of speech that was mistrusted. They then extended this work in a crowdsourced study which analyzed many more perceptual ratings of trust. Based on these studies, they identified several linguistic cues to trust. For example, they found that speech with higher mean pitch and mean intensity were associated with greater trust levels. Speech that was produced with a faster speaking rate was also judged as more trustworthy. [13] also examined individual differences across gender and native language in how people produce trustworthy speech.

Others have studied trust in the context of human-computer interaction. The work that is most closely aligned with our work is that of [3], which studied the acoustic-prosodic characteristics of trusting speech in human subjects interacting with a virtual assistant. Another related area of research examined the relationship between entrainment and trust in human-computer interaction. [14] studied the effects of lexical entraininment by a spoken dialogue system on user perception and found that users judged the entraining system as more likeable and having more integrity. [15] implemented a prosodically entraining dialogue system and similarly found that users preferred and trusted the entraining system more.

Our work builds on this prior work and focuses on human-human interactions. Rather than study the characteristics of trustworthy speech, we focus on the characteristics of speech produced by a speaker who trusts their conversational partner. We compare our findings to [3] to identify differences in trusting speech between human-human and human-machine interaction. Inspired by prior work on individual differences, we analyze cues to trusting speech in subsets of speakers by gender and native language, to understand how these traits may affect the production of trusting speech.

3. Data

We study patterns of trusting speech using the Columbia X-Cultural Deception (CXD) Corpus [16], a collection of within-subject deceptive and non-deceptive speech from native speakers of Standard American English (SAE) and Mandarin Chinese (MC), all speaking in English. The corpus was collected using a fake resume paradigm, where pairs of previously unacquainted subjects played a lying game centered around a 24-item biographical questionnaire. Subjects were instructed to complete the questionnaire, creating a false response to a random half of the questions (which were chosen to balance deceptive and truthful responses to different questions), and a truthful response for the remaining half.

Subjects took turns playing the role of interviewer or interviewee. The interviewer asked the 24 questions while their partner answered them truthfully or deceptively as indicated on their questionnaire. The goal of the interviewer was to determine the veracity of the interviewee responses, and interviewers recorded their judgments for each of the questions. The interviewee objective was to effectively deceive their partner and they too recorded each utterance they produced as either truth or lie. To provide incentive for the two roles in the game, interviewers earned $1 for every correct judgment and lost $1 for every incorrect judgment, while interviewees earned $1 for every successful lie (i.e. that was believed by their partner) and lost $1 for every unsuccessful lie. Game sessions took place in a soundproof booth where participants were seated across from each other, separated by a curtain to remove potential visual cues. Subjects were recorded via head-mounted close-talking microphones.

The CXD corpus is ideal for this study of trust. It includes interviewer judgments of deception for every question, which provide implicit measures of interviewer trust. Interviewers were financially incentivized to perform well at deception detection, which increases the validity of their trust ratings. Lastly, the corpus is balanced by participant gender and native language, enabling a study of how these traits may affect cues to trusting and mistrusting speech.

To analyze the differences between trusting and mistrusting speech, we selected interviewer turns in the CXD corpus that immediately followed an interviewee's response to a question. These interviewer turns were selected because they capture the interviewers' initial reaction to an interviewee response. The interviewer turns are labeled as trusting (T) or mistrusting (MT) based on the judgment that the interviewer recorded for that question. If the interviewer judged the question response as deceptive, we say the turn is mistrusting, and if the interviewer judged the response as truthful, we label the corresponding interviewer speech as trusting. In total, 8,009 interviewer turns from 340 unique interviewers were used for the analysis in this paper.

Table 1 displays two sample turn exchanges from the CXD corpus in order to highlight which turns are analyzed in this work. We focus on turns that are interviewer "Reaction" turns, which are the first turn immediately following an interviewee response to a question. We hypothesize that acoustic-prosodic features extracted from those interviewer reaction turns will be different when the interviewer trusts vs. mistrusts the interviewee response.

4. Method

4.1. Features

To identify differences between trusting vs. mistrusting speech behaviors, we extracted a set of acoustic-prosodic features that have also been widely used in analyses of deceptive and trustworthy speech, and are also commonly studied in general speech research. In total, we extracted 12 features:

- (1) Duration
- (2-6) Pitch minimum, maximum, mean, median, and standard deviation
- (7-11) Intensity minimum, maximum, mean, median, and standard deviation
- (12) Speaking rate

Duration is the length of each turn, measured in seconds. Pitch describes the fundamental frequency of a voice, measured in Hz. Intensity describes the degree of energy in a sound wave, measured in dB. Speaking rate is estimated using the ratio of voiced to total frames. All features were automatically extracted using Praat [17], an open-source software for speech analysis. After feature extraction, all features were Z-normalized by speaker ($z = (x - \mu)/\sigma$; x = value, μ = speaker mean, σ = speaker standard deviation).

4.2. Analysis

To identify differences between features of trusting and mistrusting speech, we computed a series of paired t-tests between the features of trusting and mistrusting interviewer turns. All tests for significance correct for family-wise Type I error by

Table 1: *Sample transcribed interviewer and interviewee turn exchanges, labeled for trust (T) and mistrust (MT).*

Role	Transcribed turn	Description	Trust Label
Interviewer	Did you ever have a cat?	Question	
Interviewee	No I never had a cat.	Answer	
Interviewer	Okay me neither ha.	Reaction	T
Interviewer	Have you ever broken a bone?	Question	
Interviewee	I broke my wrist when I was riding my bike.	Answer	
Interviewer	Ah you fell off your bike I guess.	Reaction	MT

controlling the false discovery rate (FDR) at $\alpha = 0.05$. The k^{th} smallest p value is considered significant if it is less than $\frac{k*\alpha}{n}$. In all the tables in this paper, we use T to indicate that a feature was significantly increased in trusting speech, and MT to indicate a significant indicator of mistrusting speech. We consider a result to approach significance if its uncorrected p value is less than 0.05 and indicate this with () in the tables.

5. Acoustic-prosodic Cues to Trusting and Mistrusting Speech

In this section we present the results of our analysis of acoustic-prosodic characteristics of trusting and mistrusting interviewer speech. In our first analysis, we aim to answer the following question: **What are the acoustic-prosodic characteristics of trusting and mistrusting interviewer speech?**

We analyze 12 acoustic-prosodic features extracted from each interviewer turn immediately following an interviewee response. This provides a speech sample with the interviewer's initial reaction to an interviewee answer. Trusting (T) interviewer speech is speech from an interviewer that believed their partner's response (i.e. judged it as truthful), and mistrusting (MT) interviewer speech is speech from an interviewer that did not believe their partner's response (i.e. judged it as deceptive). The trusting and mistrusting labels are determined by the interviewer's judgment alone, and are independent of the veracity of the interviewee speech (i.e. whether they were in fact lying or telling the truth). Thus, trusting/mistrusting labels capture the perception of deception. Table 2 shows the results of this analysis.

Table 2: *T-test results comparing trusting vs. mistrusting interviewer turns.*

Feature	t	p	label
Duration	-3.88	0.00011	T
Min Pitch	0.76	0.45	
Max Pitch	-1.09	0.28	
Mean Pitch	1.4	0.16	
Median Pitch	1.52	0.13	
SD Pitch	1.26	0.21	
Min Intensity	2.04	0.041	(MT)
Max Intensity	-1.49	0.14	
Mean Intensity	1.47	0.14	
Median Intensity	2.8	0.0051	MT
SD Intensity	-1.86	0.062	
Speaking Rate	2.03	0.042	(MT)

Our results show two significant indicators of trusting vs. mistrusting speech: speech duration and speech median intensity. Across all speakers, interviewer turns were significantly longer when they trusted the preceding interviewee's speech, and median intensity values were higher in interviewer turns following speech from the interviewee that they did not trust. We were also able to identify two trends in this analysis: minimum intensity values and speaking rate tended to be increased in mistrusting interviewer speech. Overall, it appeared that interviewers in general spoke louder in their intensity range and used shorter turns and a faster speaking rate when they did not trust their conversational partners.

5.1. Individual Differences

The trends in Table 2 were observed across all interviewers in our corpus. In our next analysis, we aimed to answer the following question: **Are there differences in the characteristics of trusting or mistrusting speech across gender or native language of the speaker?**

To answer this question, we selected subsets of interviewers from the CXD corpus by gender (male vs. female) or native language (Standard American English vs. Mandarin Chinese). Next, we computed t-tests between trusting and mistrusting interviewer speech for specific subsets of speakers. We ran four sets of experiments, considering: 1) only male speakers, 2) only female speakers, 3) only native speakers of Standard American English (SAE), and 4) only native speakers of Mandarin Chinese (MC). Table 3 presents the results of these experiments, along with the results across all speakers for comparison.

Table 3: *T-test results comparing trusting vs. mistrusting interviewer turns, for speaker subsets: M=male, F=female, E=English, C=Chinese. A cell with a value of T indicates that the feature is significantly increased in trusting speech; MT indicates that the feature is significantly increased in mistrusting speech. () indicates that the uncorrected p-value is ≤ 0.05.*

Feature	All	M	F	E	C
Duration	T	T		(T)	T
Min Pitch					
Max Pitch					
Mean Pitch					MT
Median Pitch					(MT)
SD Pitch					
Min Intensity	(MT)	MT			
Max Intensity					
Mean Intensity					
Median Intensity	MT	(MT)			(MT)
SD Intensity		(T)			
Speaking Rate	(MT)			(MT)	

As shown in Table 3, some findings that we observed across all speakers were also stable across speaker groups. For exam-

ple, speaker turns were longer in duration for trusting speech across all speakers, and this was also true for male, English, and Chinese interviewer subsets. Median intensity was increased in mistrusting speech across all speakers also, as well as in male and Chinese subsets. Other findings that were previously observed across all speakers were found only to be associated with particular subgroups in this analysis. For example, while we previously observed a trend that speaking rate was increased in mistrusting speech across all speakers, this analysis revealed that in fact speaking rate was only increased in mistrusting speech for native speakers of Standard American English. It is intuitive that speaking rate is a salient feature of trust for only native speakers of SAE. Non-native speakers in the corpus spoke significantly slower than native speakers ($t(7992) = 24.139, p \approx 0$) and their speaking rate is likely affected by fluency rather than trust.

There were also some features that were significant only for particular groups of speakers and were not significant in our analysis of features across all speakers. For example, mean pitch was significantly increased in mistrusting speech only for native Chinese speakers. It is possible that, because Mandarin Chinese, is a tonal language, changes in pitch may play a different role for native Chinese speakers than for native Standard American English speakers in signalling trust or mistrust. Increased variation in intensity, measured by the standard deviation of intensity, was associated with trusting speech for male speakers only. Interestingly, there were no significant indicators or even trends of trusting or mistrusting speech when considering only female speakers; this suggests that female speakers are not a group that consistently exhibits trusting speech in similar ways.

6. Human vs. Machine Interlocutor

In this section, we compare our findings with a previous study of trusting and mistrusting speech which focused on human speech directed at a virtual assistant [3]. In the study, subjects interacted verbally with a virtual assistant in order to find answers to a series of factual questions. Subjects were told that the virtual assistant (VA) that they would interact with was previously rated by other users with either a very high (4.9) or very low (1.4) score, in order to bias the subjects to trust or distrust the VA. The VA performed consistently with those provided scores — making several mistakes in the low score condition, and making no mistakes in the high score condition. The experiments were conducted in Spanish. Acoustic-prosodic features were extracted and normalized by speaker, and compared across H and L conditions to discover the features associated with trusting speech.

The authors trained predictive models of trusting speech and analyzed the top performing features: speaking rate and pitch median. They found that subjects tended to speak faster and with higher pitch in the H condition, when subjects were biased to trust the VA.

We compare our findings with those of [3] in order understand how trusting speech between two human interlocutors compares with trusting speech toward a machine. For median pitch, we found no significant difference between trusting and mistrusting speech across all speakers, but observed a trend toward increased median pitch in mistrusting speech in native Chinese speakers only. This is in contrast to [3], which observed an increase in median pitch in trusting speech. For speaking rate, we observed a trend toward faster speaking rate in mistrusting speech across all speakers. When broken down by interviewer traits, we found that this trend was only observed in native speakers of Standard American English. This is in direct contrast with [3], who observed faster speaking rates in trusting speech.

It is difficult to conclude whether these differences in features of trusting speech are due to inherent differences between human-human and human-machine interaction or not. There are several other important differences between our study and [3]. Our study took place in English, while theirs was in Spanish. The labels of trust are also fundamentally different. The CXD corpus, which we used for this study, includes labels of trust for every utterance. These trust labels were provided by the interviewers, as judged each interviewee response as truthful or deceptive, indicating their level of trust or belief in the veracity of the interviewee response. In contrast, [3] determined the subjects' level of trust in the system performance (High vs. Low), from the information they had given the subjects about prior ratings of the system which they also matched with system performance in order to influence the subjects' trust or mistrust of the system. They did not use explicit trust ratings of the system provided by the subjects in this study. Finally, there may also be implementation differences in segmentation and feature extraction.

However, despite all of these differences, we conclude that the nature of trusting speech may vary substantially depending on a number of factors, including the culture or language of the speakers, whether the interlocutor is a machine or a human, and also depending on the task at hand. In this work, we identified some differences in the features of trusting speech across speaker gender and native language. We believe that further investigation is needed to identify additional differences that may be due to language (e.g. Spanish vs. English) or due to the kind of interlocutor (human vs. machine).

7. Conclusions

In this paper we have presented a study of trusting speech in the context of interview dialogues. We analyzed the speech of interviewers that rated their partner speech as truthful vs. deceptive, in order to identify the acoustic-prosodic characteristics of trusting speech. Our analysis identified features associated with trusting vs. mistrusting speech, including duration, intensity features, and speaking rate. We also studied how these cues vary across gender and native language of the interviewers, and highlighted cues that are specific to particular subsets of speakers. Finally, we compared our results with a prior study of trusting speech in human-computer interaction and identify key differences. This work provides insights about the nature of trusting speech and has potential applications for improving human-computer interactions.

There are limitations to our comparison of findings with a prior study because of the many differences in our experimental paradigms. In future work, we would like to conduct a more direct comparison of cues to trusting speech between human-human vs. human-machine interaction which requires a controlled study. We also plan to use the insights gained from this work to build predictive models of trusting speech by training machine learning models to automatically identify trusting vs. mistrusting speech. Finally, we would like to explore additional features that characterize trusting speech, including lexical features and potentially visual features.

8. Acknowledgements

This work was funded by AFOSR grant FA9550- 18-1-0039, "Spoken Indicators of Trust Across Cultures."

9. References

[1] J. J. Lee, B. Knox, J. Baumann, C. Breazeal, and D. DeSteno, "Computationally modeling interpersonal trust," *Frontiers in psychology*, vol. 4, p. 893, 2013.

[2] X. L. Chen, S. Ita Levitan, M. Levine, M. Mandic, and J. Hirschberg, "Acoustic-prosodic and lexical cues to deception and trust: deciphering how people detect lies," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 199–214, 2020.

[3] L. Gauder, L. Pepino, P. Riera, S. Brussino, J. Vidal, A. Gravano, and L. Ferrer, "A study on the manifestation of trust in speech," *arXiv preprint arXiv:2102.09370*, 2021.

[4] T. Bickmore and R. Picard, "Establishing and maintaining long-term human computer relationships," *ACM Transactions on Computer-Human Interaction*, vol. 12, no. 2, 2005.

[5] T. Bickmore, D. Schulman, and L. Yin, "Maintain engagement in long-term interventions with relational agents," *Applications of Artificial Intelligence*, vol. 1, no. 24(6), pp. 648–666, 2010.

[6] D. DeVault, R. Arstein, G. Benn, T. Dey, E. Fast, and A. e. a. Gainer, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014)*, Paris, 2014.

[7] J. Lucas, Gale M. and Boberg, D. Traum, J. Gratch, A. Gainer, E. Johnson, A. Leuski, and M. Nakano, "Getting to know each other: The role of social dialogue in recovery from errors in social robots," in *Proceedings of Human Robot Interaction (HRI)*, Chicago, 2018.

[8] G. I. Winata, O. Kampman, Y. Yan, A. Dey, and P. Fung, "Nora the empathetic psychologist," in *Interspeech 2017*, Stockholm, 2017.

[9] R. Zhao, T. Sinha, A. W. Black, and J. Cassell, "Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior," in *Intelligent Virtual Agents 2016, LNAI 10011*, Los Angeles, 2016, pp. 218–233.

[10] M. Ochs, C. Pelachaud, and D. Sadek, "An empathic virtual dialog agent to improve human-machine interaction," in *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems (AA-MAS 2008)*, Estoril, Portugal, 2008, pp. 89–96.

[11] M. Czerwinski, J. Hernandez, and D. McDuff, "Building an ai that feels," *IEEE Spectrum*, vol. 58, no. 50, pp. 32–38, 2021.

[12] R. H. Gálvez, S. Benus, A. Gravano, M. Trnka, F. Lacerda, S. Strombergsson, M. Wlodarczak, M. Heldner, J. Gustafson, D. House *et al.*, "Prosodic facilitation and interference while judging on the veracity of synthesized statements." in *INTERSPEECH*, 2017, pp. 2331–2335.

[13] S. I. Levitan, A. Maredia, and J. Hirschberg, "Acoustic-prosodic indicators of deception and trust in interview dialogues." in *Interspeech*, 2018, pp. 416–420.

[14] G. A. Linnemann and R. Jucks, "'can i trust the spoken dialogue system because it uses the same words as i do?'—influence of lexically aligned spoken dialogue systems on trustworthiness and user satisfaction," *Interacting with Computers*, vol. 30, no. 3, pp. 173–186, 2018.

[15] R. Levitan, S. Benus, R. H. Gálvez, A. Gravano, F. Savoretti, M. Trnka, A. Weise, and J. Hirschberg, "Implementing acoustic-prosodic entrainment in a conversational avatar." in *Interspeech*, vol. 16. San Francisco, CA, 2016, pp. 1166–1170.

[16] S. I. Levitan, G. An, M. Wang, G. Mendels, J. Hirschberg, M. Levine, and A. Rosenberg, "Cross-cultural production and detection of deception from speech," in *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, 2015, pp. 1–8.

[17] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (version 6.0. 11)[software]," 2016.